Microphone Arrays as Generalized Cameras for Integrated Audio Visual Processing

Adam O'Donovan and Ramani Duraiswami
Perceptual Interfaces and Reality Laboratory
Department of Computer Science and UMIACS, University of Maryland, College Park
adamod@gmail.com, ramani@umiacs.umd.edu

Jan Neumann
Siemens Corporate Research, Princeton, NJ
jan.neumann@siemens.com

Abstract

Combinations of microphones and cameras allow the joint audio visual sensing of a scene. Such arrangements of sensors are common in biological organisms and in applications such as meeting recording and surveillance where both modalities are necessary to provide scene understanding. Microphone arrays provide geometrical information on the source location, and allow the sound sources in the scene to be separated and the noise suppressed, while cameras allow the scene geometry and the location and motion of people and other objects to be estimated. In most previous work the fusion of the audio-visual information occurs at a relatively late stage. In contrast, we take the viewpoint that both cameras and microphone arrays are geometry sensors, and treat the microphone arrays as generalized cameras. We employ computer-vision inspired algorithms to treat the combined system of arrays and cameras. In particular, we consider the geometry introduced by a general microphone array and spherical microphone arrays. The latter show a geometry that is very close to central projection cameras, and we show how standard vision based calibration algorithms can be profitably applied to them. Experiments are presented that demonstrate the usefulness of the considered approach.

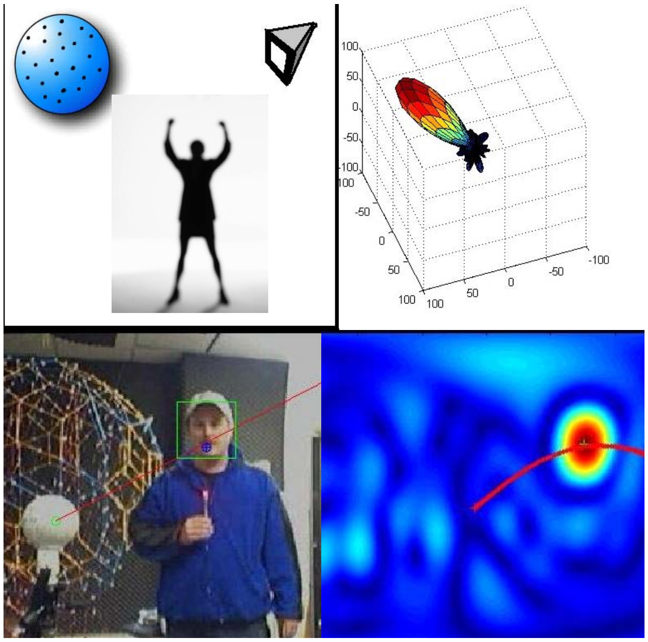

Figure 1. Cameras and spherical microphone arrays are used to record a speaking person. The spherical array (small white sphere) acts as a central projection camera for the sound field, and with a calibrated camera can be used to perform source localization. Epipolar lines are shown in red in the camera image and the sound-field image. In the latter, the epipolar line is seen as a curve, as space is distorted by the Cartesian plot in angular variables. The system can be used to perform beamforming of speech in presence of noise, or to register sounds to image objects.

1. Introduction

Most animals use eyes and ears together to make sense of the place of objects and other animals in the world. There are many applications in machine vision, e.g., surveillance, gunshot detection, meeting recording and analysis, or in the identification of noise sources where computer vision and computer audition can be used profitably together, and there are recent works in all these areas, that combine audio analysis and image-processing/computer vision to various degrees to solve problems of interest. The analysis of the two modalities has in the past been developed with different

1-4244-1180-7/07/$25.00 ©2007 IEEE

sets of tools, and usually the combinations that are made are done at a later stage in the analysis, fusing the outputs of the audio and visual analysis. While this strategy may be appropriate for many problems, there are a few problems where an integrated analysis may provide benefit. This is especially the case where the geometrical location of a sound source is being inferred, and both modalities provide only partial information.

With the identification of the pinhole camera model and the use of the epipolar restriction in stereo, geometry has played a central role in computer vision. Over the past decade or so, the study of geometrical estimation of the world using combinations of cameras of different kinds has seen tremendous progress. Because of the centrality of geometrical estimation in computer vision, it would be fair to say that the geometric estimation algorithms available to practitioners of computer vision are more advanced than in audio. Many algorithms for processing audio data, such as source separation and noise suppression, are often formulated to work in a "blind" setting, without knowledge of the source location or source characteristics. However, despite this source localization and tracking form an important part of audio processing, and knowledge of the source location can lead to better performance.

1.1. Motivation and present contribution

Arrays of microphones can be geometrically arranged and the sound captured can be used to extract information about the geometrical location of a source. Our interest in this subject was raised by the idea of using a relatively new sensor and an associated beamforming algorithm reported in [21, 22], for audiovisual meeting recordings (see Figure 1). This array has since been the subject of some research in the audio community. While considering the use of the array to detect and to beamform (isolate) an auditory source in the meeting system, we observed that this microphone array is a central projection device for far-field sound sources, and can be easily treated as a "camera" when used with more conventional video cameras. Moreover, certain calibration problems associated with the device can be solved using standard approaches in computer vision.

While this paper is primarily concerned with the spherical microphone arrays, we were naturally led to how other microphone arrays could be included in the framework as generalized cameras, similar to the recent work in vision on generalized cameras, that are imaging devices that do not restrict themselves to the geometric or photometric constraints imposed by the pinhole camera model [12], including the calibration of such generalized bundles of rays [24]. In the most general case, any camera is simply a directional sensor of varying accuracy.

Microphone arrays that are able to constrain the location of a source can be interpreted as directional sensors. Due to this conceptual similarity between cameras and microphone arrays, it is possible to utilize the vast body of knowledge about how to calibrate cameras (i.e. directional sensors) based on image correspondences (i.e. directional correspondences). Specifically, we utilize the fact that spherical arrays of microphones can be approximated as directional sensors which follow a central projection geometry. Nevertheless, the constraints imposed by the central projection geometry allow us to apply proven algorithms developed in the computer vision community as described in [20] or [13] to calibrate arbitrary combinations of conventional cameras and spherical microphone arrays.

Paper Organization: Below we briefly review some relevant work. Next, in section 3, we provide some background material on audio processing, to make the paper self contained, and to establish notation. Sections 4 describes the algorithms developed for working with the spherical array and cameras, and results are described in Section 5. Section 6 concludes the paper and discusses possible work with other types of microphone arrays.

2. Prior Work

Microphone arrays have long been used in many fields (e.g., to detect underwater noise sources), to record music, and more recently for recording speech and other sound. The latter is of our concern here, and there is a vast literature on the area. An introduction to the field may be obtained via a pair of books that are collections of invited papers that cover different aspects of the field [3, 8]. Solid spherical microphone arrays were first developed (both theoretically and experimentally) by Meyer and Elko [21, 22] and extended by Li et al. [15, 16].

There are several papers that consider combined audio visual processing, and we will only mention a few here. Pointing a pan-tilt-zoom camera at a sound source has been achieved by several authors [4], while a few employ the knowledge of the location of the sound source obtained from vision to improve the audio processing [4, 32] Several authors have performed joint audio-visual tracking using various approaches (particle filtering [31], learning a probabilistic graphical model using low level audio and visual features [2], finding the pixels that create sound via an efficient formulation of canonical correlation analysis [14], built a large efficient industrial system [6]). Modern image processing and computer vision techniques were used to define new features for sound recognition in [9]. These references only form a small part of a large body of work in this area.

The closest to our work is that of Negahdaripour [23], where the joint geometry of an underwater sonar camera system is developed. There is a difference however in the methods used in that paper, which relies on active probing of the scene using acoustic pulses, and then images it rather like LADAR, using a time of flight map for the reflected signals. Due to the large error in the 3rd coordinate of their estimates the authors choose to treat the sensor as a 2D sensor, with the two retained image dimensions as range and one angular coordinate. In contrast, we discuss microphone arrays whose "image" geometry is similar to that in regular central projection cameras, and do not actively probe the scene but rely on sounds created in the environment. We envisage the sensor we describe would be useful in indoor people and industrial noise monitoring situations, while [23] would be useful in underwater imaging.

3. Background

3.1. Source Localization and Beamforming

Assume that the acoustic source that produces an acoustic signal $y(t)$ is located at point $\mathbf{p}$ and K microphones are located at points $\mathbf{q}_1, ..., \mathbf{q}_K$. The signal $s_m(t)$ received at the m^{th} microphone contains delayed versions of the source signal, its convolution with the channel impulse response, and noise (or other sources) and is given by

$$s_m(t) = r_m^{-1} y(t - \tau_m) + y(t) \star h_m^*(q_m, p, t) + z_m(t). \quad (1)$$

where the first term on the right is the direct arriving signal, $r_m = ||\mathbf{p} - \mathbf{q}_m||$ is the distance from the source to the mth microphone, c is the sound speed, $\tau_m = r_m/c$ is the delay in the signal reaching the microphone, $h_m^*(\mathbf{q}_m, \mathbf{p}, t)$ is the filter that models the reverberant reflections (called the room impulse response, RIR) for the given locations of the source and the m^{th} microphone, star denotes convolution, and $z_m(t)$ is the combination of the channel noise, environmental noise, or other sources; it is assumed to be independent at all microphones and uncorrelated with $y(t)$ [7].

In general τ_m will not be measurable as the source position is unknown. Knowing the locations of two microphones, m and n respectively, We denote the time difference of arrival (TDOA) of a signal between receivers m and n as $\tau_{mn} = \tau_n - \tau_m$. TDOAs are usually obtained using a generalized cross-correlation (GCC) between signal frames (short pieces of the signal of length N) s_m and s_n acquired at the m^{th} and n^{th} sensors respectively [10]. Let us denote by $r_{mn}(\tau)$ the GCC of $s_n(t)$ and $s_m(t)$ and its Fourier transform by $R_{mn}(\omega)$. Then,

$$R_{mn}(\omega) = W_{mn}(\omega) S_m(\omega) S_n^*(\omega), \quad (2)$$

where $W_{mn}(\omega)$ is a weighting function. Ideally, $r_{mn}(\tau)$ (computed as the inverse Fourier transform of $R_{mn}(\omega)$) will have a peak at the true TDOA between sensors m and n (τ_{mn}). In practice, many factors such as noise, finite sampling rate, interfering sources and reverberation might affect the position and the magnitude of the peaks of the cross correlation, and the choice of the weighting function can improve the robustness of the estimator. The phase transform (PHAT) weighting function was introduced in [10]:

$$W_{mn}(\omega) = |S_m(\omega) S_n^*(\omega)|^{-1}. \quad (3)$$

The PHAT weighting places equal importance on each frequency by dividing the spectrum by its magnitude. It was later shown [7] that it is more robust and reliable in realistic reverberant acoustic conditions than other weighting functions designed to be statistically optimal under specific non-reverberant noise conditions.

Source localization using time delays: The availability of a single time delay between a pair of receivers, places the source on a hyperboloid of revolution of two sheets, with its foci at the two microphones (see Figure 5). In human hearing, time delays between the two ears places the source on this hyperboloid (also mislabeled the "cone of confusion"), and humans have to use other cues to resolve ambiguities. In general purpose arrays, we can add further microphones, and intersect the hyperboloids formed by delay measurements with each pair. Measurements at three collinear microphones restrict the source to lie on a circle whose center lies on the axis formed by the microphones, while knowing the time delays between 4 non-collinear microphones in principle can provide the exact source location. However TDOAs are very noisy, and the non-linear intersection algorithms may give poor results with the noisy input data, and various methods to improve the algorithms are still being developed by researchers.

Beamforming: The goal of beamforming is to *"steer"* a *"beam"* towards the source of interest and to pick its contents up in preference to any other competing sources or noise. The simplest *"delay and sum"* beamformer takes a set of TDOAs (which determine where the beamformer is steered) and computes the output $s_B(t)$ as

$$s_B(t) = \frac{1}{K} \sum_{m=1}^{K} s_m(t + \tau_{ml}), \quad (4)$$

where l is a reference microphone which can be chosen to be the closest microphone to the sound source so that all τ_{ml} are negative and the beamformer is causal. To steer the beamformer, one selects TDOAs corresponding to a known source location. Noise from other directions will add incoherently, and decrease by a factor of K^{-1} relative to the the source signal which adds up coherently, and the beamformed signal is clear. More general beamformers use all the information in the K microphone signal at a frame of length N, may work with a Fourier representation, and may explicitly null out signals from particular locations (usually directions) while enhancing signals from other locations (directions). The weights are then usually computed in a constrained optimization framework.

Beampattern: The pattern formed when the, usually frequency-dependent, weights of a beamformer are plotted as an intensity map versus location are called the beampattern of the beamformer. Since usually beamformers are built for different directions (as opposed to location), for source that are in the "far-field," the beampattern is a function of two angular variables. Allowing the beampattern to vary with frequency gives greater flexibility, at an increased optimization cost and an increased complexity of implementation.

Localization via Steered Beamforming: One way to perform source localization is to avoid nonlinear inversion, and scan space using a beamformer. For example, if using the delay and sum beamformer the set of time delays $\hat{\tau}_{mn}$ corresponds to different points in the world being checked for the position of a desired acoustic source, and a map of the beamformer power versus position may be plotted. Peaks of this function will indicate the location of the sound source. There are various algorithms to speed up the search as discussed in [32].

3.2. Spherical Microphone Arrays

We will be concerned with solid spherical microphone arrays (as in Figure 1) on whose surface several microphones are embedded. In [21] an elegant prescription that provided beamformer weights that would achieve as a beampattern any spherical harmonic function $Y_n^m(\theta_k, \varphi_k)$ of a particular order n and degree m in a direction (θ_k, φ_k) was presented. Here

$$Y_n^m(\theta, \varphi) = (-1)^m \sqrt{\frac{2n+1}{4\pi} \frac{(n-|m|)!}{(n+|m|)!}} P_n^{|m|}(\cos\theta) e^{im\varphi}, \tag{5}$$

where $n = 0, 1, 2, ...$ and $m = -n, ..., n$, and $P_n^{|m|}$ is the associate Legendre function. The maximum order that was achievable by a given array was governed by the number of microphones, S, on the surface of the array, and the availability of spherical quadrature formulae for the points corresponding to the microphone coordinates (θ_i, φ_i) , $i = 1, \ldots, S$. Li et al. [15] extend the analysis to arbitrarily placed microphones on the sphere.

Since the spherical harmonics form a basis on the surface of the sphere, building the spherical harmonic expansion of a desired beampattern, allowed easy computation of the weights necessary to achieve it. In particular if one desires a beampattern that is a delta function, truncated to the maximum achievable spherical harmonic order p, in a particular direction (θ_0, φ_0), then we can use

$$\delta^{(p)}(\theta - \theta_0, \varphi - \varphi_0) = 2\pi \sum_{n=0}^{p-1} \sum_{m=-n}^{n} Y_n^{m*}(\theta_0, \varphi_0) Y_n^m(\theta, \varphi), \tag{6}$$

to compute the weights for any desired look direction. This beampattern is often called the "ideal beampattern," since it enables picking out a particular source. The beampattern achieved at order 6 is shown in Figure 1. A spherical array can be used to localize sound sources by steering it in several directions and looking at peaks in the resulting intensity image formed by the array response in different directions.

The ability of an array to isolate a sound source from a given look direction is often quantified by the directivity index and is given in dB:

$$DI(\boldsymbol{\theta}_0, \boldsymbol{\theta}_s, ka) = 10 \log_{10} \left(\frac{4\pi \left| H(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0) \right|^2}{\int_{\Omega_s} \left| H(\boldsymbol{\theta}, \boldsymbol{\theta}_0) \right|^2 d\Omega_s} \right), \tag{7}$$

where $H(\boldsymbol{\theta}, \boldsymbol{\theta}_0)$ is the actual beampattern looking at $\boldsymbol{\theta}_0 = (\theta_0, \varphi_0)$ and $H(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0)$ is the value in that direction. The DI is the ratio of the gain for the look direction $\boldsymbol{\theta}_0$ to the average gain over all directions. If a spherical microphone array can precisely achieve the regular beampattern of order N as in [17], its theoretical DI is $20\log_{10}(N+1)$. In practice, the DI index will be slightly lower than the theoretical optimal due to errors in microphone location and signal noise.

Spherical microphone arrays can be considered as central projection cameras. Using the ideal beam pattern of a particular order, and beamforming towards a fixed grid of directions, one can build an intensity map of a sound field in particular directions. Peaks will be observed in those directions where sound sources are present (or the sound field has a peak due to reflection and constructive interference). Since the weights can be pre-computed and a relatively short fixed filters, the process of sound field imaging can proceed quite quickly. When sounds are created by objects that are also visualized using a central projection camera, or are recorded via a second spherical microphone array, an epipolar geometry holds between the camera and the array, or the two arrays. Below we describe experiments which confirm this hypothesis.

4. Experiments with Spherical Arrays and Cameras

We had available a previously built 60-microphone spherical microphone array of radius 10cm in our laboratory [17]. The array interfaces to a computer via a 64 channel signal acquisition interface using PCI-bus data acquisition cards that are mounted in the analysis computer and connected to the array, and the associated signal processing apparatus. This array can capture sound to disk and to memory via a Matlab data acquisition interface that can acquire each channel at 40 kHz, so that a Nyquist frequency of 20 kHz is achieved. The same Matlab was equipped with an image-processing toolbox, and we acquired camera images via a USB 2.0 interface on the computer. A 320×240 pixel, 30 frames/s web camera was used. While, the algorithms should be capable of real-time operation, if they were to be programmed in a compiled language and linked via the Matlab mex interface, in the present work this was not done, and previously captured audio and video data were processed subsequently.

Camera and Array Calibration: The camera was calibrated using standard camera calibration algorithms in OpenCV, while the array microphone intensities were calibrated as described in the spherical array literature. We then proceeded with the task of relative calibration of the array and the camera. To calibrate this system, we built a wand that had an LED and a small speaker (both about 3mm × 3mm) attached to its tip (see Figure 2).

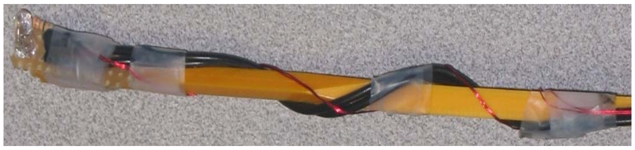

Figure 2. A calibration wand with a co-located sound and light source was constructed by attaching a microspeaker and a bright LED to one end of a long pencil.

When a button is pressed, the LED lights up and a sound chirp is simultaneously emitted from the speaker. Light and sound are then simultaneously recorded by the camera and microphone array respectively. We can determine the direction of the sound by forming a beam pattern as described above (also see Figure 1) which turns the microphone array into a directional sensor. In the figure below, we show an example sample acquisition

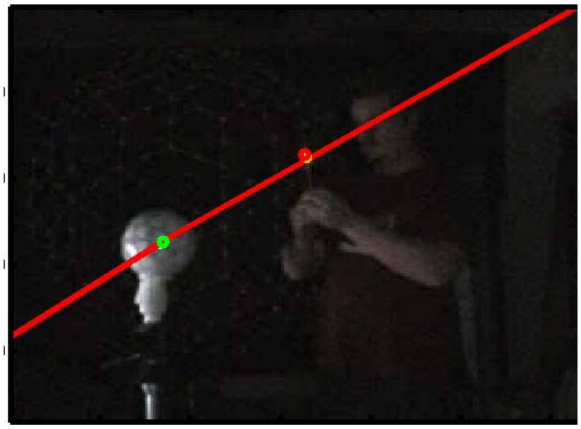

Figure 3. Image of Calibration procedure. Notice the epipolar line of the microphone correspondence in the camera image

As one can see the calibration recovered the epipolar geometry between the camera and the array very accurately. The same procedure can also be used to calibrate several (hemi-)spherical microphone arrays since both are equivalent to internally calibrated cameras, and thus also have to conform to the epipolar geometry. Figure 3 shows how the image ray projects into the spherical array and intersects the peak of the beam pattern.

4.1. One camera and one spherical array

In this case, the camera image and "sound image" are related by the epipolar geometry induced by the orientation and location of the camera and the microphone array respectively. We will assume that the camera is located at the origin of the fiducial coordinate system. For each sound we thus have the direction $\mathbf{r}_{mic}(\theta, \varphi)$, which we need to correspond to the projection of the 3D location of the sound source into the camera image $\mathbf{p}_{cam}$.

If we have precalibrated the camera, then we can transform $\mathbf{p}_{cam}$ into normalized image coordinates $\mathbf{r}_{cam} = \mathbf{K}^{-1}\mathbf{p}_{cam}$ where $\mathbf{K}$ is the internal calibration matrix of the camera (we disregard the radial distortion parameters). If the camera coordinate system and the microphone coordinate system are related by a rotation matrix $\mathbf{R}$ and a translation vector $\mathbf{t}$, then each correspondence is related by the essential matrix $\mathbf{E}$:

$$0 = \mathbf{r}^t_{mic}\mathbf{E}\mathbf{r}_{cam} = \mathbf{r}^t_{mic}[\mathbf{t}]_x\mathbf{R}\mathbf{r}_{cam} \qquad (8)$$

To compute the essential matrix $\mathbf{E}$ and extract $\mathbf{t}$ and $\mathbf{R}$, we follow [19]. We decide among the resulting four solutions by choosing the solution that maximizes the number of positive depths for the microphone array and the camera.

If the camera is not calibrated, then the direction in the microphone and the pixel in the image would be related by the fundamental matrix $\mathbf{F}$

$$0 = \mathbf{r}^t_{mic}\mathbf{F}\mathbf{p}_{cam} = \mathbf{r}^t_{mic}[\mathbf{t}]_x\mathbf{R}\mathbf{K}^{-1}\mathbf{p}_{cam} \qquad (9)$$

We can solve for $\mathbf{F}$ using a multitude of algorithms as described in [13], we chose to use a linear algorithm for which we need at least 8 correspondences, followed by nonlinear minimization. At this point we could also utilize robust and sensor specific error functionals that would take into account the different localization accuracies and noise characteristics of the image and microphone array "image" formation process. This is left for future work though.

The epipolar geometry induced by the essential or fundamental matrices, allows us interchangeably to transfer a point from an image to a 1-D curve in the microphone array directional space defined by the implicit equation $\mathbf{r}_{mic}(\theta, \varphi) \cdot (\mathbf{F}\mathbf{p}_{cam}) = f(\theta, \varphi) = 0$, or we can transfer a directional measurement from the microphone array to an epipolar line defined by the equation $\mathbf{p}_{cam} \cdot (\mathbf{F}^t\mathbf{r}_{mic}) = 0$.

4.1.1 N cameras and 1 spherical array

Multicamera systems with overlapping fields of view, attached to microphone arrays are now becoming popular to record meetings, and a product called RoundTable has been announced by Microsoft. The location of speakers in an integrated mosaic image is a problem of interest in such systems. An extension of the procedure described in section **??** can also be used to calibrate several (hemi-)spherical microphone arrays since both are equivalent to internally calibrated cameras, and thus also have to conform to the epipolar geometry.

For multiple cameras, we only need to know the calibration information from 2 cameras, to use a method similar to [1] to calibrate the remaining cameras. Since the microphone is already intrinsically calibrated, we only need to determine the internal calibration parameters for a single camera, compute the calibration between the spherical array and the calibrated camera, reconstruct the correspondences in space, and then use the 3D points to calibrate the system of cameras as described in[1].The results could then be further improved using bundle-adjustment [26] which again could take into account the different noise characteristics of sound and image-based directional imaging.

4.1.2 N cameras and N spherical arrays

Similarly, one could also use ≥ 2 (hemi-)-spherical microphone arrays [16], and an arbitrary number of uncalibrated cameras. First, we can calibrate the two microphone arrays using the epipolar constraint as described earlier. Then we can reconstruct the calibration points in space using the computed calibration. Due to the omnidirectional nature of the microphone array, we can be sure that all the calibration points are "visible" to both microphone arrays and thus can be reconstructed. We can now use the reconstructed structure to compute the projection matrices for each of the cameras. We can now use all the cameras and the microphone arrays together with the reconstructed points to initialize a bundle-adjustment procedure.

4.2. Example application: Speaker tracking and Noise Suppression

We now used the epipolar geometry between a spherical microphone array and a camera in a meeting room scenario. The microphone array was used to detect the direction of sound sources in the scene, in this case the speaker in the room, and then the epipolar geometry, to project the epipolar line into the camera image. We can now employ a simple face detector along the vicinity of the epipolar line to located the exact position of the speaker in the image. In our system we use a face detector based on Haar wavelets as implemented in OpenCV [18]. This allows us then to accurately zoom into the image and display a detailed view of the speaker. Since the search space is greatly reduced, the localization can be done extremely fast, and also switching from one speaker to the next can be done instantly. In Figure 4 we show the sound image where the peak indicates the mouth region, this peak is located and using the epipolar geometry projected into the image resulting in a epipolar line. We now search along this line for the most likely face position, triangulate the position in space and then set our zoom level accordingly.

The knowledge of the face location can help improve the recorded audio as well. In supplemental material (available on the first author's web site) we present an example in which an extremely loud music interference was played from a location to the left of the subject, and below him, after the face was initially detected as above. Once the face rectangle was extracted, a template match was used to detect the mouth region. The epipolar line from the image passing through this region was then constructed on the soundfield image. The lower panel of Figure 4 shows the sound field image generated, where the distracter can be seen to be extremely bright compared to the source. The location corresponding to the mouth was passed to the beamforming algorithms, and the sound from this location was extracted. The sound achieved by this process is also attached. A further refinement of the algorithm could be to throw an explicit null at the location of the other source [27], though we have not done this yet.

5. Conclusions and Future Work

We have presented a novel approach that considers the geometrical restrictions introduced by microphone array measurements, and those introduced by cameras in a joint framework, which allows localization and calibration problems to be more efficiently solved. The theoretical sections considered the general situation briefly, and then the case of the spherical array in more detail. The ideas were validated experimentally.

We believe that the approach considered here, of imaging the sound field using a spherical array(s) and the actual scene using camera(s) will have many applications, and several vision algorithms can be brought to bear. For example, when multiple cameras will be used with multiple spherical arrays, we can build a joint mosaic of the image and the soundfield image. Such an analysis can easily indicate locations where sounds are being created, their intensity and frequencies. This may have applications in industrial monitoring and surveillance.

We can also extend the work to consider the regular microphone arrays. Such microphones are nowadays found on handheld devices such as mobile phones, and these often include cameras in them as well. In mobile phones these microphones are used to perform echo and noise cancellation. Other locations where such arrays may be found include at the corners of screens, and in the base of videoconferencing systems. Using time delays, one can restrict the source to lie on a hyperboloid of revolution, or when

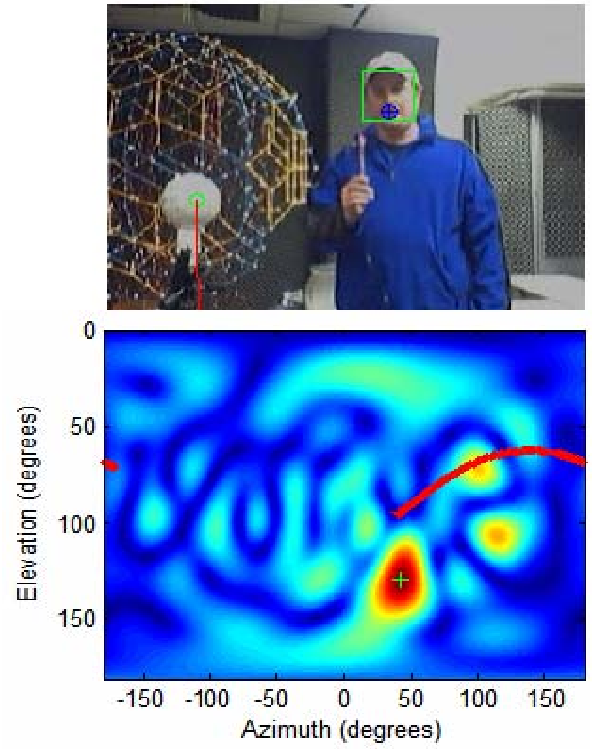

Figure 4. An example of the use of the system in speaker tracking with noise suppression. The bright red spot on the sound image (marked with a +) corresponds to the dominant source. The less dominant source however lies on the epipolar line in the sound image induced by the location of the mouth in the camera image, and this source is beamformed.

several microphones are present, at their intersection. If the processing of the camera image is performed in a joint framework, then the location of the source can be quickly performed, as is indicated in Fig. 5 below.

It would also be useful to consider some specialized systems where the camera and microphones are placed in a particular geometry. For example the human head can be considered to contain two cameras with two microphones on a rigid sphere. A joint analysis of the ability of this system to localize sound creating objects located at different points in space using both audio and visual processing means could be of broad interest.

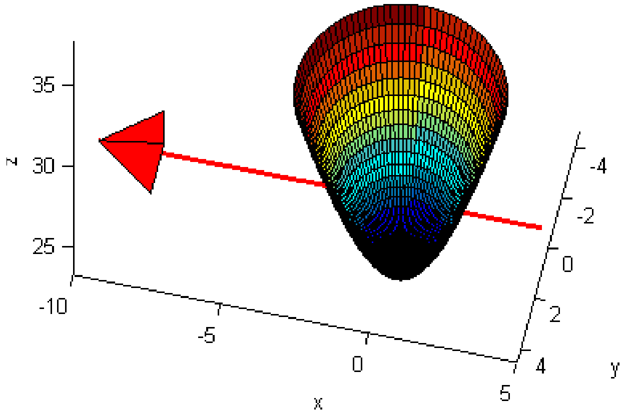

Figure 5. A ray from a camera to a possible sound generating object, and its intersection with the hyperboloid of revolution induced by a time delay of arrival between a pair of microphones. The source lies at either of the two intersections of the hyperboloid and the ray.

References

[1] J. P. Barreto and K. Daniilidis. "Wide area multiple camera calibration and estimation of radial distortion." In *OMNIVIS 2004 - Workshop on Omnidirectional Vision and Camera Networks*, Prague, 2004.

[2] M. Beal, N Jojic, H. Attias. "A Graphical Model for Audiovisual Object Tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **25**:828-836, 2003.

[3] M. S. Brandstein and D. B. Ward (editors) . *Microphone Arrays: Signal Processing Techniques and Applications*, Springer-Verlag, Berlin, Germany, 2001.

[4] U. Bub, M. Hunke, A. Waibel. "Knowing Who to Listen to in Speech Recognition: Visually Guided Beamforming." In *Proceedings IEEE ICASSP-95*, **1**:848-851, 1995.

[5] Y. T. Chan and K. C. Ho. "A Simple and Efficient Estimator for Hyperbolic Location," *IEEE Transactions on Signal Processing,* **42:**1905-1915, 1994.

[6] C. Choi, D. Kong, S. Lee, K. Park, S. Hong, H. Lee, S. Bang; Y. Lee, S. Kim. "Real-time audio-visual localization of user using microphone array and vision camera," *Proceedings IEEE/RSJ IROS 2005*, 1935-1940, 2005.

[7] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein. "Robust localization in reverberant rooms", in *Microphone Arrays: Signal Processing Techniques and Applications*, M. S. Brandstein and D. B. Ward ed., Springer-Verlag, Berlin, Germany, 157-180, 2001.

[8] Y.A. Huang and J. Benesty, ed. *Audio Signal Processing For Next Generation Multimedia Communication Systems*, Kluwer Academic Publishers 2004.

[9] Y. Ke, D. Hoiem, R. Sukthankar. "Computer Vision for Music Identification," *Proceedings IEEE CVPR*, **1**:597-604, 2005.

[10] C. H. Knapp and G. C. Carter. "The generalized correlation method for estimation of time delay", *IEEE Transactions on Acoustics, Speech and Signal Processing,* **24**:320-327, 1976.

[11] M. S. Brandstein and H. F. Silverman. "A robust method for speech signal time-delay estimation in reverberant rooms", *Proceedings IEEE ICASSP*, **1**:375-378, 1997.

[12] M. D. Grossberg and S. K. Nayar. "A general imaging model and a method for finding its parameters," In *Proc. ICCV 2001*, **2**:108–115, 2001.

[13] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, Cambridge, UK, 2000.

[14] E. Kidron and Y.Y. Schechner, M. Elad. "Pixels that Sound," *Proceedings IEEE CVPR*, **1**:88-96, 2005.

[15] Z. Li, R. Duraiswami, E. Grassi, and L.S. Davis. "Flexible layout and optimal cancellation of the orthonormality error for spherical microphone arrays," *Proceedings IEEE ICASSP*, **4**:41–44, 2004.

[16] Z. Li and Ramani Duraiswami. "Hemispherical microphone arrays for sound capture and beamforming," *Proceedings IEEE WASPAA*, 106–109, 2005.

[17] Z. Li and Ramani Duraiswami. "Flexible and Optimal Design of Spherical Microphone Arrays for Beamforming," *IEEE Transactions on Audio, Speech and Language Processing*, **15**:702-714, 2007

[18] R. Lienhart, L. Liang, and A. Kuranov. "A detector tree of boosted classifiers for real-time object detection and tracking," *Proceedings IEEE ICME*, **2**:277–280, 2003.

[19] Y. Ma, J. Kosecka, and S. S. Sastry. "Motion recovery from image sequences: Discrete viewpoint vs. differential viewpoint," *Proceedings ECCV*, **2**:337-353, 1998.

[20] Y. Ma, S. Soatto, J. Kosecka, and S. Sastry. *An Invitation to 3D Vision. From Images to Geometric Models*. Springer, 2003.

[21] J. Meyer and G. Elko. "A highly scalable spherical microphone array based on anorthonormal decomposition of the soundfield," *Proceedings IEEE ICASSP*, **2**:1781-1784, 2002.

[22] J. Meyer and G. Elko. "Spherical Microphone Arrays for 3D sound Recording," A*udio Signal Processing For Next Generation Multimedia Communication Systems* Ed. Y.A. Huang and J. Benesty, 67- 89, Kluwer Academic Publishers 2004.

[23] Shahriar Negahdaripour, "Epipolar Geometry of Opti-Acoustic Stereo Imaging," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, in press (available online), 2007.

[24] S. Ramalingam, P. Sturm, and S. Lodha. "Towards complete generic camera calibration." *Proceedings IEEE CVPR*, **1**:1093–1098, 2005.

[25] F. D. la Torre Frade, C. Vallespi-Gonzalez, P. Rybski, M. Veloso, and T. Kanade. "Learning to track multiple people in omnidirectional video," *Proceedings ICRA 2005*, 4150 - 4155, 2005.

[26] B. Triggs, P. F. McLauchlan, R. I. Hartley, and A. W. Fitzgibbon. "Bundle adjustment — a modern synthesis." In B. Triggs, A. Zisserman, and R. Szeliski, editors, *Vision Algorithms: Theory and Practice, LNCS:1883*. Springer-Verlag, 298–373, 1999.

[27] B.D. Van Veen and K.M. Buckley. "Beamforming: a versatile approach to spatial filtering," IEEE Signal Processing Magazine, **5**:4-24, 1988.

[28] J. Vermaak and A. Blake. "Nonlinear filtering for speaker tracking in noisy and reverberant environments", *Proc. IEEE ICASSP*, **5**:3021-3024, 2001.

[29] H. Wang and P. Chu. "Voice source localization for automatic camera pointing system in videoconferencing", *Proc. IEEE ICASSP* , **1**:187-190, 1997.

[30] M. Wax and T. Kailath (1983). "Optimum localization of multiple sources by passive arrays", *IEEE Transactions on Acoustics, Speech and Signal Processing*, **31**, 1210-1218.

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

[32] D.N. Zotkin and R. Duraiswami. "Accelerated speech source localization via a hierarchical search of steered response power," *IEEE Transactions on Speech and Audio Processing,* **12**:499–508, 2004.