CALIBRATION OF AUDIO-VIDEO SENSORS FOR MULTI-MODAL EVENT INDEXING

Thorsten Kühnapfel [1], *Tele Tan* [1], *Svetha Venkatesh* [1] *and Eric Lehmann* [2]

[1] Department of Computing, Curtin University of Technology, Western Australia
[2] Western Australian Telecommunications Research Institute, Western Australia

ABSTRACT

This paper addresses the coordinated use of video and audio cues to capture and index surveillance events with multi-modal labels. The focus of this paper is the development of a joint-sensor calibration technique that uses audio-visual observations to improve the calibration process. One significant feature of this approach is the ability to continuously check and update the calibration status of the sensor suite, making it resilient to independent drift in the individual sensors. We present scenarios in which this system is used to enhance surveillance.

Index Terms— Calibration, Multimedia system

1. INTRODUCTION

In recent years, surveillance systems have become increasingly sophisticated by combining different types of sensors to enhance performance. Common choices of sensors are cameras and microphones, which are used predominantly for tracking and identifying the speaker [1, 2, 3, 4]. This paper explores the introduction of an audio array in a video surveillance domain, to explore the use of extracting and indexing "coordinated" audio and video events. For example, consider a video and audio stream in an area where there are 2 groups of people - one speaking German and the other English. Our system detects the video event, say "gathering of more than one person", and than extracts "beam-steered" audio from the direction of the video event. Thus, in the above example, the two video events "gathering 1" and "gathering 2" will be indexed by their corresponding beam-steered audio, thereby allowing "gathering 1" to be indexed by the enhanced German audio content, and "gathering 2" to be indexed by the enhanced English audio. Note that the video event detection is not the focus of this paper. We focus on the calibration of the coordinated video-audio array, and demonstrate its use for video surveillance. Such calibration can be done in a static setting, but drift often perturbs such systems over time. Instead, we propose a dynamic calibration system triggered by a single, talking person walking across the visual field of view.

Thanks to Anders Johansson[2] for his support on sound processing and audio hardware advices.

In previous work, audio-video information is fused for tracking purposes, mainly in video conferencing systems. These systems use video segmentation techniques, such as background subtraction, motion detection, shape detection or skin colour detection to locate a speaker. As tracking is sensitive to light changes, people in the background or occlusion, there is a need for these systems to incorporate sound source localisation algorithms to enhance tracking. For this purpose, the proposed approaches in [1, 2, 3, 4] all assume known geometrical alignment of the sensors. In [5] the number of audio and video sources are assumed to be known at all times. Likewise, in [4, 2] a linear mapping between the audio and video data is assumed.

We focus on the issue of dynamic audio-visual sensor calibration based on audio-video observations. This involves estimating a calibration function that maps the two dimensional video coordinates of a stationary camera to the one dimensional audio angle of arrival in a linear microphone array. Importantly, the calibration function is automatically computed from a sequence of audio-visual observations and does not assume linear mapping or known relative audio-video sensor orientation. To enhance the accuracy of the calibration function, the estimated sound source location is smoothed by a novel approach. After the calibration function is estimated, the system is able to steer an audio beam towards a "target" based on the image location. Targets can be a wide range of video events such as single people, multiple people and so on. We demonstrate our system in a surveillance situation by enhancing speech for detected visual events. The novelty of such a system is the self calibration of audio-video sensor pairs that combine audio and video information, and its application for surveillance in which video events are indexed by enhanced audio.

The paper is organised as follows: Section 2 provides an overview of related work. Section 3 presents the methodology and the results of our experiments are shown in section 4. Section 5 concludes our work.

2. RELATED WORK

The authors in [4] propose a sensor fusion model based on Bayesian networks, where a probabilistic graphical model is chosen to approximate a linear mapping between the video

and audio with known sensor geometry.

Another way to fuse the data is via the Sequential Monte Carlo method (SMC) [1, 5]. Here the location of the foreground objects in the video stream and the location of sound sources are converted into a probability density function (pdf). In [1] the pdf is computed based on the assumption that the optical center of a static camera is aligned with the middle of the audio array. In [5] a prerecorded sequence with a known number of objects is used to estimate the correlation of the video and audio signals.

The speaker tracking system proposed in [3] uses a simple summing voter to combine all sound sources, skin colour and motion locations. In order to map the video based information with the audio information, the authors discretise the audio information into 30° sectors and the video in sectors according to their alignment setup. The audio information is also normalised to a reference power profile, generated in an anechoic chamber.

In comparison to all these methods, the system proposed in this paper does not require any prior cross sensor measurements, known relative audio-video sensor geometry or any linearity assumptions in mapping audio-video pairs.

3. METHODOLOGY

This section details the tasks in the proposed surveillance system. First, the sound source localisation algorithm and the smoothing is discussed. Then, the video segmentation and face detection algorithms are presented, followed by the calibration approach. Finally, the surveillance system is described to demonstrate the indexing of video events with enhanced audio.

3.1. Angle of Arrival Algorithm

For a sound source in the far-field region, the time delay of arrival (TDOA), τ, between two microphones can be computed by the general cross-correlation with PHAT weighting (GCC-PHAT) [6]. The generalization of the GCC-PHAT is the SRP-PHAT [7] algorithm and it is used to compute a distribution β, for a one dimensional, uniform alignment, microphone array as:

$$\beta(\tau) = \nu 2\pi \sum_{l=1}^{N} \sum_{k=1}^{N} \int_{-\infty}^{+\infty} \frac{\Gamma_{lk}(\omega)}{|\Gamma_{lk}(\omega)|} e^{j\omega\tau(l-k)} d\omega \quad (1)$$

where τ is the time delay, l and k are microphones indices ranging from 1 to N, and Γ_{lk} denotes the cross power density spectrum. In this discretised distribution, τ ranges from $-\tau_{max}$ to $+\tau_{max}$, which corresponds to a scan for sound sources from $-\frac{\pi}{2}$ to $\frac{\pi}{2}$, relative to the center of the array. The distribution β is normalised using ν. An estimate of a sound source's TDOA, $\hat{\tau}_s$, is obtained as the lag τ that maximises $\beta(\tau)$, and can then be converted to an angle of arrival (AOA), α, as follows [7]:

$$\alpha(\tau) = \arccos\left(\frac{c}{dF_s}\tau\right) \quad (2)$$

where F_s is the sampling frequency, d the distance between the microphones and c the speed of sound. Thus, $\beta(\tau)$ can be translated to $\beta(\alpha)$. The max of $\beta(\alpha)$, $\widehat{\alpha}$, is called AOA. Since the distribution $\beta(\alpha)$ can be estimated at every sampling instance t, we refer to this distribution as $\beta_t(\alpha)$.

3.1.1. *Smoothing of AOA result*

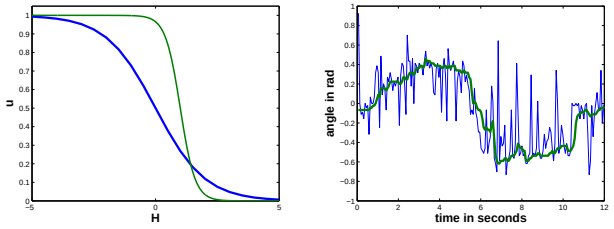
The estimated AOA has quite a large variance around the true direction of the source. To reduce this variance in estimation, we propose a smoothing approach based on the entropy, H, of the discrete AOA distribution $\beta_t(\alpha)$. For smoothing, we use exponential forgetting to combine β_t and β_{t-1} as

$$\overline{\beta}_t(\alpha) = \left(1 - \frac{u}{\epsilon}\right)\beta_{t-1}(\alpha) + \frac{u}{\epsilon}\beta_t(\alpha) \quad (3)$$

where ϵ is a appropriate normalising factor, and u as update factor computed as:

$$u = \frac{1}{1 + e^{\frac{1}{\sigma}[H-\mu]}} \quad (4)$$

where μ is the mean and σ the variance. To understand u, assume without loss of generality that $\mu = 0$, $\sigma = 1$. Then, when the distribution β is not uniform (i.e. very peaked), $H = 0$, and $u = 1$. This corresponds to the update factor being 1 when there is little uncertainty in β. Conversely, when the distribution β is uniform, $H = 1$, and $u = 0$. μ and σ control the rate of "fall" as shown in figure 1(a). Figure 1(b) presents the sound source location, in relation to the center of the array, based on the unsmoothed and smoothed $\overline{\beta_t}$.



(a) (b)

Fig. 1. **Smoothing:** (a) shows two functions of u, with $\mu = 0$ and $\sigma = 1$ (dark) and with for $\mu = 1$ and $\sigma = 0.3$ (light). (b) shows the result of the sound source localisation with a frequency range from 800 Hz to 3000 Hz, for the smoothed (dark) and unsmoothed (light) AOA.

3.2. Video segmentation

Objects in the video stream are detected by background subtraction [8]. For calibration purposes, we use the AdaBoost face detection algorithm [9] to locate the face regions.

3.3. Calibration

The calibration function for the proposed system maps the video coordinates to the one dimensional audio information of a linear microphone array. To accurately estimate such a calibration function, 3 restrictions apply: Only one foreground object must be detected in the image, the object has to emit sound and the sound has to be directed towards the microphones. This proposed system uses a single person as the calibration object and speech as the sound source. Therefore, the system must have the ability to detect speech and faces. We use the support vector machine (SVM) [10], trained with Mel Frequency Cepstral Coefficients (MFCC) [11] of speech and background noise samples. The face detection algorithm computes the probability of a frontal face. For the mapping, the image is divided into I vertical bins. This discretisation allows the system to categorise the estimated AOA's to a bin i, based on the mean position of the person. Given n observations $[AOA_k, p\,(face_k)]$, where AOA_k and $p\,(face_k)$ are the angle of arrival and the probability that the person is facing the microphones in the k^{th} observation, for each of the bins, a mean AOA is computed as

$$\overline{AOA_i} = \frac{\sum_{k=1}^{n} AOA_k^i p\,(face_k)}{\sum_{k=1}^{n} p\,(face_k)} \tag{5}$$

where $i \in 1 : I$. A fitted polynomial function with degree 3 is shown in figure 2. The calibration function is quite linear, which is expected from the sensor geometry. The Levenberg-Marquardt algorithm is used for fitting the calibration function, which interpolates a correlated AOA for all vertical image locations.

3.4. Video surveillance

The aim of the dynamic calibration is to enhance the recorded sounds of a corresponding image events. When the spatial distance between individuals is within a certain threshold, the visual event "crowd" is detected, suggesting a possible conversation. Due to noisy backgrounds or other groups of people, the recorded conversation from a single microphone is inadequate. Therefore, the system can steer an audio beam towards the groups, based on the calibration, to enhance the audio signal in the given direction. To steer an audio beam, a delay and sum beamformer is implemented in the frequency domain [12] as follows:

$$B\,(k, \alpha) = \sum_{n=1}^{N} W_n(k) X_n\,(k)\, e^{-j2\pi f_k \frac{d}{c}(n-1)\sin\alpha} \tag{6}$$

$$f_k = k\frac{F_s}{M} \tag{7}$$

where k is the frequency index and α is the steering angle. The side lobe level is controlled with $W_n(k)$, $X_n(k)$ is the complex value of the FFT transformed audio signal, where n is the index of the microphone and M the size of the FFT.

4. EXPERIMENTS

The performance of the proposed system was evaluated by a set of 3 experiments. For all experiments, the sound was emitted in a 2 to 3 meters range from the microphones and the camera was mounted behind the microphones, so that it faces in approximately the same direction as the microphones. The array consists of 7 omnidirectional, condenser microphones with a spacing of 4 cm. The AOA analysis was performed in a frequency range from 800 Hz to 3000 Hz. Audio enhancement was done over the full frequency range from 0 Hz to 4000Hz. The calibration function was estimated from a single, talking person walking in front of the sensor system. Audio was captured at 8000 Hz, 16 bits per sample. The image resolution was 320x240 pixels, captured at 15 frames per second.

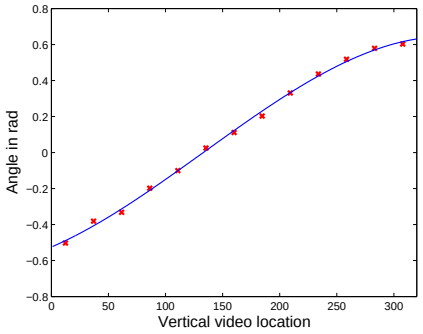


Fig. 2. **Calibration:** The $\times$ are the $\overline{AOA}$ for each video bin and the line is the calibration function.

4.1. Calibration

The accuracy of the calibration was tested by using two pure sine waves, with different frequencies. Two audio speakers were placed on the far right and left sides of the camera field of view, at a distance of 3 meters. Then the audio beam was first steered towards the speaker with the 3500 Hz sine wave, and then towards the one emitting 2500 Hz, based on the image location. The difference in the enhanced to original signal strange at 2500 Hz was 16.3 dB and 29.5 dB at 3500 Hz.

4.2. Indexing audio-video events

Two stationary groups of people converse in English and German, each group consists of two people, located on the right and left sides of the camera view, at about 3 meters. Two beamformed audio streams were produced, each from a beam steered towards the center of each group. To evaluate the quality of the beamformed audio streams, they were first normalised to the same signal strength as that of a single microphone. Then a test group listened to the resulting beamformed audio and gave a mean opinion score (MOS), expressed as a

single number in the range 1 (Bad) to 5 (Excellent). This score was based on the understanding of the conversation. The beamformed audio streams achieved an average score of 3.62 and the original signal an average score of 1.6, as shown in figure 3.

(a)

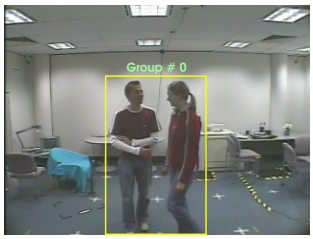
Video Event	MOS	
	Steered	Orig
Group 0	4	1.7
Group 1	3.75	1.5

(b)

Fig. 3. **Indexing audio-video events:** (a) shows a screen shot of the video event detection. The table (b) shows the mean opinion score between the beam steered audio and the original audio stream.

4.3. Audio events for moving targets

The last experiment consists of two people walking in front of the camera and having a conversation. To increase the background noise, a radio is playing in the far right end of the room. This experiment proves the ability to steer an audio beam towards a moving target. The results were evaluated again by using the MOS for the original and the normalised beamformed audio stream. As shown in figure 4 the average MOS for the beam steered audio stream was 3.9 compared with the original scored of 2.



(a)

Video Event	MOS	
	Steered	Orig
Group 0	3.9	2

(b)

Fig. 4. **Audio events for moving targets:** (a) shows a screen shot of the video event detection. Table (b) shows the mean opinion score between the beam steered audio and the original audio stream.

A video file that demonstrates the calibration and the last two experiments can be downloaded from:
www.computing.edu.au/∼thorsten/AudioVideoIndexing.avi.

5. CONCLUSIONS AND FUTURE WORK

We have demonstrated that the system can steer an audio beam towards a detected video event based on the video and audio observation. Experiments have demonstrated the ability of the system for surveillance.

In future, we will extend the system with a pan tilt and zoom camera to give the system the potential to detect a person based on an audio event, even so he might not be in the view of the camera. Also the system will be able to re-adjust itself should drifts amongst the video-audio sensors occur.

6. REFERENCES

[1] Y. Chen and Y. Rui, "Real-time speaker tracking using particle filter sensor fusion," *Proc. of the IEEE*, vol. 92, no. 03, pp. 485–494, 2004.

[2] J. Vermaak, M. Gangnet, A. Blake, and P. Perez, "Sequential monte carlo fusion of sound and vision for speaker tracking," *ICCV*, vol. 01, pp. 741, 2001.

[3] D. Lo, R.A. Goubran, R.M. Dansereau, G. Thompson, and D. Schulz, "Robust joint audio-video localization in video conferencing using reliability information," *IEEE Trans. on Instrumentation and Measurement*, vol. 53, no. 4, pp. 1132–1139, 2004.

[4] M. J. Beal, H. Attias, and N. Jojic, "Audio-video sensor fusion with probabilistic graphical models," *ECCV (1)*, pp. 736–752, 2002.

[5] H. Asano et al., "An application of a particle filter to bayesian multiple sound source tracking with audio and video information fusion," *Proc. Int. Conf. on Information Fusion*, pp. 805–812, 2004.

[6] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust. Speech Signal Process*, vol. 24, no. 4, pp. 320–327, 1976.

[7] A. Johansson, N. Grbic, and S. Nordholm, "Speaker localisation using the far-field SRP-PHAT in conference telephony," in *International Symposium on Intelligent Signal Processing and Communication Systems*, 2002.

[8] C. Stauffer, W. Eric, and L. Grimson, "Learning patterns of activity using real-time tracking," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 22, pp. 747–757, 2000.

[9] P. Viola and M. Jones, "Robust real-time object detection," *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2001.

[10] S. Gunn, "Support vector machines for classification and regression," Tech. Rep., Department of Electronics and Computer Science, University of Southampton, 1998.

[11] S. Davis and P. Mermelstein, "Comparison of parametric representation for monosyllable word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, pp. 357–366, 1980.

[12] B. Maranda, "Efficient digital beamforming in the frequency domain," *Journal of Acoustic Society of America*, vol. 86, pp. 1813–1819, 1989.