

# ROBUST SPEAKER TRACKING BY FUSION OF COMPLEMENTARY FEATURES FROM AUDIO AND VISION MODALITIES.

*Mandar Rahurkar, Amit Sethi and Thomas S. Huang*

ECE Department  
University of Illinois-Urbana Champaign  
email: {rahurkar,asethi,huang}@ifp.uiuc.edu

## ABSTRACT

*In this work we address the challenging problem of target tracking and source separation for surveillance using both the audio and video modalities generated by the object. Our algorithm uses time delay of arrival from audio modality to estimate the position of the target to initialize and re-initialize visual tracking whenever it fails. Position of the object ascertained by visual tracking is used to estimate the delay which is used to separate the sound coming from the target source from background noise. The emphasis of this paper is to demonstrate robust tracking performance using a novel scheme for integration of audio and visual modalities. Moreover our algorithm is fully automatic which does not require any initialization. It is empirically shown that having two complementary modalities helps us track the object more robustly, which is intuitively expected. We demonstrate this robustness on some challenging video sequences where the target is occluded, and also when some of the frames in the sequence are completely corrupted by noise.*

## 1. INTRODUCTION

In surveillance, target tracking and speech signal enhancement/separation are two of the important tasks. These two problems can be solved jointly in a synergetic manner. The spatial motion of a moving target can be followed using video data, captured by a camera. If the object emits sound (e.g., person speaking) audio data can be used to estimate the time delay of arrival of sound between two (or more) microphones and thus used for tracking. Tracking using audio is robust to occlusions and variations in lighting whereas tracking using video alone gives us both x and y co-ordinates. This point is demonstrated in Fig. 2, where the visual modality loses track of the region of interest (ROI) due to occlusion. Thus, intuitively it is obvious that these modalities should complement each other and when used together should provide a more robust system with collective capabilities that is more than sum of its parts. The audio and

video signals are correlated at various levels; lip movement of the speaker is correlated with the amplitude of part of the signal and can also help us narrow down the ROI to sound generating source. Also the time delay of arrival (TDOA) between the two microphones is correlated with the position of the speaker in the image. We also exploit TDOA for the audio based estimate of the person's position using two microphones. When visual tracking fails due to occlusion, instability of the tracker, or corruption of frames by random noise, audio modality can be used to re-initialize the visual tracker. On the other hand, when visual tracking is robust, the estimate of the position of the object can be used to get an estimate of the time delay of the component of the sound coming from the target at the microphone pair, thus helping in source separation and noise cancellation.

We consider a surveillance system with audio and video subsystems. The system blocks are shown in Figure 1. These subsystems can be used together either by viewing it as a feature integration problem or these subsystems can interact amongst themselves to give a better solution than what either one of them can generate individually. We demonstrate that using audio alongside video our system is robust to occlusion as well as in the case when some frames are totally corrupted by noise. Audio is used for automatic initialization of visual tracking. Result of the video tracking is used to estimate the time delay for the audio signal generated by the target in a robust manner. This delay is further used to separate the target audio signal from the background noise. To the best of our knowledge neither has been attempted before in a multi-modal manner.

The paper is organized as follows. In Section 2 we describe our methods for video tracking, audio-based position estimation, and robust multi-modal tracking. We also present our algorithm for source separation and noise cancellation of the target sound component. We present some experiments and results to demonstrate the robustness of our multi-modal target tracking and noise cancellation in Section 3. Finally we conclude with discussion on future directions in Section 4.

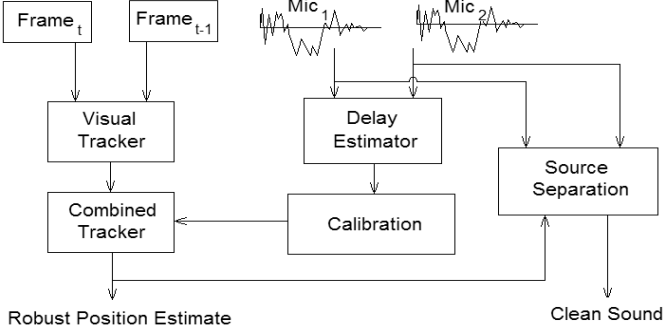


Fig. 1. System Diagram.

## 2. ALGORITHM

Though similar problem has been addressed before [1, 2, 3] for audiovisual object tracking. However, none of these works deal with occlusion of the target or problem of noisy frames. The problem of audio source separation using visual tracking has also not been addressed. We start by developing the audio and video subsystems of the surveillance system independent of each other. Then we combine the two subsystems for dealing with problems of visual tracking: initialization, occlusion, and frames corrupted by noise. We also solve the problem of source separation and noise cancelation in audio using the result of visual tracking.

We reiterate that focus of the paper is on combining these modalities in a synergetic manner and demonstrating that we can track a target more robustly than by using either modality in itself.

### 2.1. Time Delay of Arrival Estimation using Audio Signals

The delay  $\tau$  is estimated as follows. We consider windowed audio frames of  $N$  samples with 75% overlap. It should be noted that several audio frames make up one video frame in terms of time. We used a coherence measure of cross-correlation to determine the delay between the two microphones expressed as:

$$R_{ij}(\tau) = \sum_{n=0}^{N-1} x_i[n]x_j[n - \tau] \quad (1)$$

where  $x_i[n]$  is the discrete sample signal received by microphone  $i$  and  $\tau$  is the TDOA between the two received signals. In our case we had two microphones. The cross-correlation is maximal when  $R_{ij}(\tau)$  is maximal when  $\tau$  is equal to the offset between the two signals. The complexity in computing  $R_{ij}(\tau)$  using Equation 1 is  $O(N^2)$ . This can be approximated by computing the inverse Fourier transform of the cross-spectrum as given by:

$$R_{ij}(\tau) \approx \sum_{n=0}^{N-1} X_i(K)X_j(K)^* e^{j2\pi k\tau/N} \quad (2)$$

We have developed this algorithm closely along the lines of Knapp et al. [4].

In actuality, the calibration process fixes the mapping between  $\tau$  (which is estimated for each visual frame) and  $x$ .

The mean of the  $x$ 's collected for all the audio frames corresponding to a video frame gives the audio estimate of the position of the object in  $x$ -direction at time  $t$ ;  $x_t^{audio}$ . The inverse of the standard deviation of the  $x$ 's collected for all the audio frames corresponding to a video frame gives the confidence in audio-based estimate of the position  $AudConf$ , which is used in combining the two modalities to improve object tracking as described later.

### 2.2. Visual Tracker

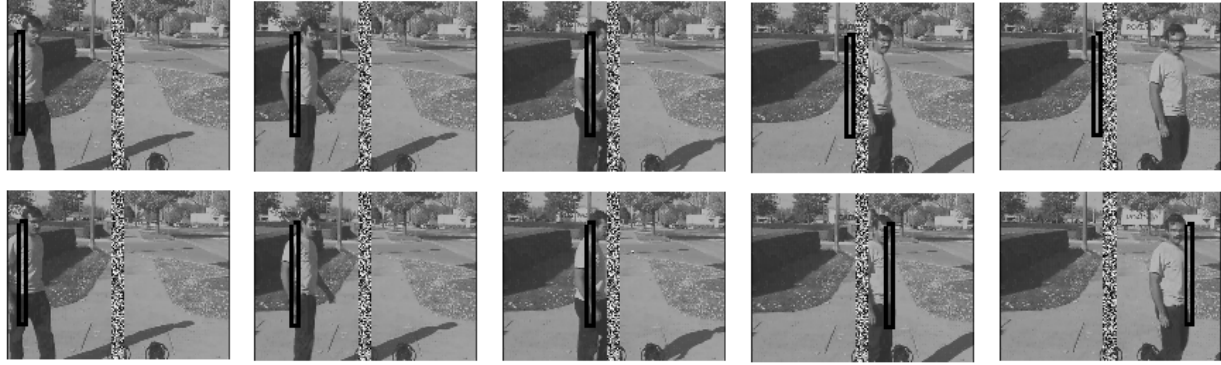
Our visual tracker algorithm is based on the mean shift tracker [5]. The matching algorithm finds an exact solution by searching with one pixel-shifts in a region around the expected position of the window in the current frame, and picking the window that gives the maximum Bhattacharya [6] coefficient with the window from the previous frame. This is described in Equation 3.

$$x_t^{video} = \arg \max_{x \in \mathcal{N}(x_{t-1})} \sum_{k=1}^n \sqrt{h_t^k(x)h_{t-1}^k(x_{t-1})} \quad (3)$$

where  $x_t^{video}$  is the position of the window in the frame  $t$  (current frame),  $h_t^k(x)$  is the histogram for  $k^{th}$  feature in the  $t^{th}$  frame at the window around position  $x$ , and  $\mathcal{N}(x_{t-1})$  is the neighborhood around  $x_{t-1}$  which is defined as a rectangular region around  $x_{t-1}$  plus a fraction of the previous motion vector if the previous motion vector is trustworthy, as determined by the maximum Bhattacharya coefficient. This gives a simple tracking strategy that can reduce the search space for the new window by predicting the position of the window in the next frame by using a simple strategy.

### 2.3. Multi-Modal Object Tracking

For initializing the visual tracker in the first frame, the position of the target as determined by the audio subsystem is used. After this the visual tracker performs two-frame tracking as described above. For cases where the visual tracking fails, a criteria to determine the failure was determined. In such a scenario, the estimate of the position of the target determined by audio was used to re-initialize the tracker. The estimate of the target position was fed back to the audio subsystem to estimate the delay associated with the sound component from the target to perform noise cancellation.



**Fig. 2.** (Top) Tracking performance using video alone in presence of occlusion(white strip). Note in the right most frame tracker is unable to follow the subject. (Below) Tracking performance using A/V. Now target is being followed after occlusion.

Failure of visual tracker was determined as follows. If the tracker loses the object due to drift, or occlusion, it settles on the background. When this happens the tracking window stops moving and settles on a constant window that does not change. This, will also result in the maximum Bhattacharya coefficient becoming close to one. Thus, when these two conditions happen simultaneously for consecutive frames, it is an indication of tracker failure. When frames become totally corrupted, or the tracker suddenly loses track of the object, it will result in a maximum Bhattacharya coefficient close to zero. This was the other criteria to indicate failure of visual tracking. The third way to determine failure was the indication of a highly confident estimate of target position from the audio subsystem that was far away from the visual tracking-based estimate. This is summarized in Equation 4.

$$VisFail = \begin{cases} \text{TRUE} & \text{if } bcMax \geq \theta_1 \\ & \text{AND } |x_t^{video} - x_{t-1}| = 0; \\ \text{TRUE} & \text{else if } bcMax \leq \theta_2; \\ \text{TRUE} & \text{else if } AudConf \geq \theta_3 \\ \text{FALSE} & \text{else.} \end{cases} \quad (4)$$

where  $VisFail$  is the boolean flag that tells whether visual tracking has failed or not,  $bcMax$  is the maximum Bhattacharya coefficient of matching the window from previous frame with windows in the current frame,  $x_t$  is the position of the window in  $t^{th}$  frame,  $AudConf$  is the confidence of the audio subsystem in its prediction of the position of the target, and  $\theta_1$ ,  $\theta_2$ , and  $\theta_3$  are empirically determined thresholds. The position  $x_t$  is set to  $x_t^{audio}$  (which is the estimate of the position of the target as determined by the audio subsystem) when  $VisFail$  is TRUE and to  $x_t^{video}$  if  $VisFail$  is FALSE, to give a robust estimate of  $x_t$ .

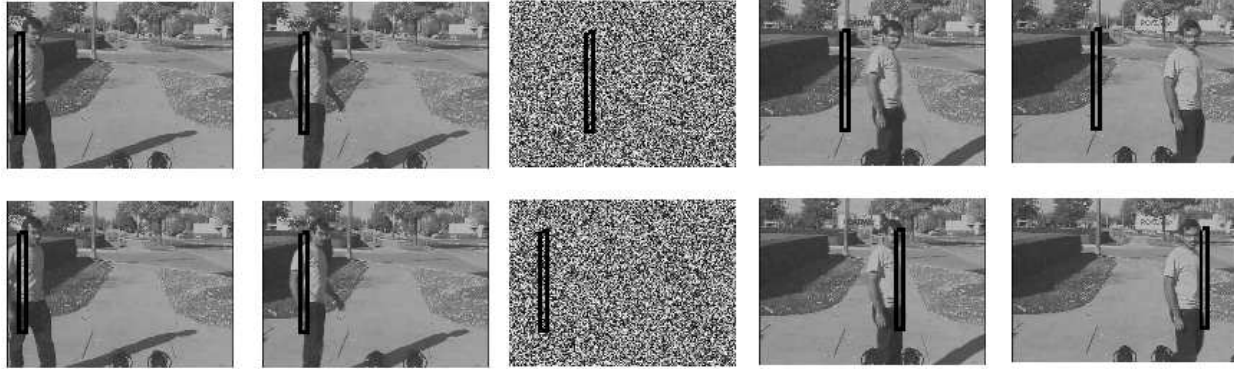
### 3. RESULTS

#### 3.1. Audio Visual Corpus

We tested the algorithm on several test cases and having a multi-modal object tracking improved the performance compared to having only either one of the modalities. Some of the cases due to which a visual tracker failed was due to drift of the tracker on to a matching background, change in its appearance, occlusion of the target, and corruption of frames with random noise. We also tested the algorithm for noise cancellation for source separation of the sound coming from the source of interest. Target motion was mostly horizontal and translational without any significant movement in the y-direction. However, there was significant change in the appearance of the target due to rotation and articulation. The algorithm was consistently able to track the speaking target in the presence of background noise, occlusion or even when frames were corrupted with noise. The output of the tracker was used for noise cancellation and perceptual improvement in the audio quality was noticed. Thus the target speech is separated from the background noise, again, without any speaker speech modelling or initialization.

The video capture rate was 15 frames per second while audio was digitized at 44100 Hz. Thus, we have 2940 audio samples for each video frame. The horizontal direction represents the time along the sequence. Since the delay locations  $\tau$  had to be mapped to image locations, 10 manually annotated frames were used only once and thereafter only raw data was given to the algorithm. No model parameters were set by hand and no initialization was required at any stage.

We present the results on two sequences which had occlusion and or had dropped frames. Audio waveforms were consistently corrupted with background noise. Occlusions were simulated by inserting a bar of randomly generated pixels as shown in the Figure 2.1. When visual tracker



**Fig. 3.** (Top) Tracking performance using video alone when frames are dropped due to corruption by random noise. Note in the right most frame tracker is unable to follow the subject. (Below) Tracking performance using A/V. Now target is being followed after frame drops due to noise.

reaches the portion where occlusion occurs it loses the target and is unable to locate it again as indicated by the the constant estimate of the x-coordinate. When we added the audio stream which is not affected by the occlusion the tracking performance improves as seen, where tracker now uses the estimates given by the audio stream and is thus able to locate the target.

Frame corruptions were simulated by replacing the entire intermittent frames with random noise pixels. The improvement in the performance by having a additional audio modality is demonstrated in Figure 3.1. Thus in both these case visual tracker lost the target, but was able to follow the target with the estimates from audio, consistently.

#### 4. CONCLUSION AND FUTURE WORK

In this paper we have presented a novel algorithm for tracking and surveillance which uses Audio and Video modalities which complement each other. The modalities are used in systematic fashion to improve the performance. Unlike other methods our algorithm does not require any initialization which is performed automatically. The output of this algorithm was also used for noise cancellation using beamforming and spectral subtraction. This algorithm finds several applications mainly in surveillance where it is important to track the movements of target robustly as well as listen to what is being said in presence of background noise in real time. Our future work will focus on extending this model for tracking multiple targets while simultaneously performing source separation when we have more sources than the number of microphones.

#### 5. REFERENCES

[1] Matthew J. Beal, Nebojsa Jojic, and Hagai Attias, "Graphical model for audiovisual object tracking.," *IEEE Trans on PAMI*,

vol. 25.

[2] J. Vermaak, M. Gangnet, A. Blake, and P. Perez, "Sequential monte carlo fusion of sound and vision for speaker tracking.," *Proc. Int'l Conf. Computer Vision*, June 2000.

[3] Y. Rui and Y. Chen, "Better proposal distributions: Object tracking using unscented particle filter.," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2001.

[4] Charles H. Knapp and G. Clifford Carter, "The generalized correlation method for estimation of time delay.," *IEEE Trans on ASSP*, vol. 24.

[5] D. Comaniciu, V. Ramesh, and P. Meer, "Real-time tracking of non-rigid objects using mean shift.," *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 142–149, June 2000.

[6] T. Kailath, "The divergence and bhattacharya distance measures in signal selection.," *IEEE Trans Communication Technology*, vol. 15, pp. 52–60, 1967.