Long Range Audio and Audio-Visual Event Detection Using a Laser Doppler Vibrometer

Tao Wang[ab], Zhigang Zhu[ab], Ajay Divakaran[c]
[a]Dept. of Comp Sci, City College, City University of New York, NY, USA 10031
[b]Dept. of Comp Sci, Graduate Center, City University of New York, NY, USA 10031
[c]Sarnoff Corp., 201 Washington Rd., P.O.Box 5300, Princeton, NY, USA 08543

ABSTRACT

Association of audio events with video events presents a challenge to a typical camera-microphone approach in order to capture AV signals from a large distance. Setting up a long range microphone array and performing geo-calibration of both audio and video sensors is difficult. In this work, in addition to a geo-calibrated electro-optical camera, we propose to use a novel optical sensor - a Laser Doppler Vibrometer (LDV) for real-time audio sensing, which allows us to capture acoustic signals from a large distance, and to use the same geo-calibration for both the camera and the audio (via LDV). We have promising preliminary results on association of the audio recording of speech with the video of the human speaker.

Keywords: LDV, Laser-Doppler Vibrometry, long-range, audio-visual, event detection, integration.

1. INTRODUCTION

Event detection using both audio and video is receiving growing interests in surveillance applications. In general, most human activity detection systems depend mainly on visual information[1,2,3], whereas the audio modality is used as complementary information to discover and explain interesting or abnormal patterns in a scene[4,5,6]. In some scenarios, however, audio conveys more significant information than video, for example, a human talking behind an object, or two people with similar appearance facing back against the camera. In the past, microphones or microphone arrays have been employed in audio-visual surveillance, but they have the limitation of very short ranges. Furthermore, these types of sensors need to be fixed at pre-determined locations. If the targets move out of their sensing ranges, they will not be able to obtain any signals. A parabolic microphone can capture voice signals at a fairly large distance; however, when it points to the direction of the target, all the signals on the way are captured. A Laser Doppler Vibrometer (LDV), on the other hand, is a long-range, non-contact acoustic measurement device to detect the speed of the target's vibration based on Doppler frequency shift. Within about two hundred meters with sensitivity on the order of 1um/s[7], the LDV can be used to obtain the acoustic signals of a target (a human or other target) in a large distance caused by the sound of the target next to a reflecting surface on which the laser points to.

In the last few years, we have studied the long range voice detection using an LDV with the aid of a Pan-Tilt-Zoom (PTZ) camera[8,9]. The signal quality acquired from the LDV is mainly determined by the reflection and the vibration properties of the selected background surface nearby the target. The reflection problem can be dealt with using a retro-reflective tape on the selected surface or selecting a good reflecting surface using the image processing. Usually smooth and brighter segmented regions in the image can be selected as possible candidates, the one with sufficient strong signal returns when pointing the laser to the surface is selected. This part of work has been presented in our earlier paper[10]. The vibration properties on various surfaces are the main issue we will discuss in this paper. Because we may need to track a moving object, the nearby surfaces, which are selected to point the laser beam to, need to be changed as the target moves. The main contribution of this work is to build robust acoustic background models of various reflecting surfaces as one type of prior knowledge in order to detect acoustic events. The other prior knowledge is the calibration between the 1D laser point of the LDV and 2D images of the PTZ camera in order to find the target in a 3D position so that we can easily and quickly find the surfaces nearby the target and detect the audio event clearly. The PTZ camera serves to detect the human target and find laser pointing surfaces. The signals from the audio and video are obtained synchronously, and event detection is performed separately on each sensor modality. Then integration is made on the decision level by combining both audio and video information.

The rest of paper is organized as follows. The system framework and configuration are presented in section 2. In section 2, the calibration between the PTZ camera and the LDV is also discussed. In section 3, the audio background modeling for outlier detection is described. In section 4, the event detection method using both audio and video information is presented. In section 5, the experimental results for the audio modeling and detection, video detection and their integration are demonstrated. Conclusions are provided in section 6.

2. SYSTEM FRAMEWORK AND CONFIGURATION

2.1 System framework

The framework for long range audio-video event detection is shown in Fig. 1. The audio and video signals are obtained synchronously from the LDV and the PTZ, respectively. The initial background scene and possible surface candidates are analyzed via video processing. Audio background models are constructed based on the initial knowledge of the acoustic scene and are not updated until the scene is changed. The modeling technique will be explained in section 3. The detection in the video modality consists of two parts: surface detection and object detection. The surface detection module is used to assist the pointing of the laser beam of the LDV to a proper reflecting surface in order to acquire acoustic signals. The object detection module is used to make a decision of target presence from its visual information. The integration between the audio and video information is made at the decision level.

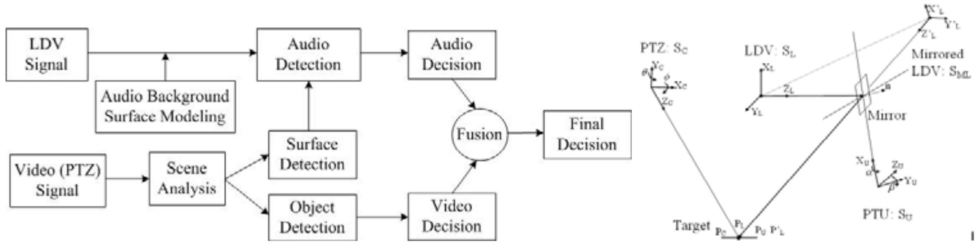

Figure 1. System framework (left) and configuration (right) for long range audio-video event detection using the LDV

2.2 System configuration

In order to detect audio signals at a large distance, an LDV is employed as the main component in the system. The LDV works according to the principle of laser interferometry. It uses a HeNe red laser with the wavelength of 633.8 *nm* and 1 *mm/s* velocity range for the voice vibration. Whenever the object has moved by half the wavelength, which is 0.3169 μm (or 12.46 micro inches), the intensity has gone through a complete dark-bright-dark light cycle. A detector converts this signal to a voltage fluctuation. The Doppler frequency f_D of this sinusoidal cycle is proportional to the velocity v of the object according to $f_D = 2v/f$. In our previous work[8,9], we have found that the vibration of most objects in man-made environments caused by voice waves can be readily detected by the LDV at distances of up to 200 meters.

The PTZ camera with 26x optical zoom is able to detect humans or other objects at a great distance. It is mounted on the top of the LDV to assist the laser pointing to surfaces in the background scene to pick up acoustic signals of the targets. Signals are acquired synchronously from the LDV and the PTZ. In order to automatically and quickly turning the laser point at various locations, a mirror mounted on a pan-tilt-unit (PTU) is used to reflect the laser beam to the selected surface. It also provides the pan and tilt angles that enable a laser point observed in the image. Therefore, each laser point is associated with a 4D vector, the x-, y- pixel locations in the PTZ image and the pan- and tilt- angles of the PTU which will be used to obtain the 3D location of the laser point on the selected surface via triangulation, and the 3D information will be used to focus the laser in real-time.

2.3 Calibration

In a multimodal sensor fusion system, we may need to perform geometric calibration among various sensor modalities. In our case, it is the calibration of a 2D camera and a 1D laser sensor (Fig 1). The calibration of laser range finders and cameras can be found in literature[11,12] and most of them use markers of features like edges and corners that are visible/measurable from both sensors. In our system, we use a controllable mirror instead of a scanning laser, which makes the calibration a little more complicated. The intrinsic parameters of the PTZ camera are calibrated using the method in[13]. In order to obtain the extrinsic parameters, the relation between the LDV coordinate system S_L and the camera coordinate system S_C is first defined as:

$$P_L = R_C P_C + T_C \tag{1}$$

where P_L and P_C are the 3D point represented in S_L and S_C, respectively. The R_C and T_C are the rotation matrix and translation vector between S_L and S_C. Next, the relation between the LDV coordinate system S_L and the mirrored LDV coordinate system S_{ML} is defined as:

$$P_L = R_U R_{LR} R_U^T (P_{ML} - T_U) + T_U \tag{2}$$

where P_L and P_{ML} are the 3D point represented in the S_L and the S_{ML}, respectively. The R_U and T_U are the rotation matrix and translation vector between S_L and the PTU coordinate system. The R_{LR} is the rotation matrix that converts a right hand coordinate system to a left hand coordinate system. Then the extrinsic parameters are estimated by combining Eq. (1) and Eq. (2), as

$$R_C P_C = R_U R_{LR} R_U^T (P_{ML} - T_U) + (T_U - T_C) \tag{3}$$

The image data and laser data are obtained in different positions of a checkerboard. Because the variables P_{ML} - T_U and T_U -T_C are not independent, the distance between the fore lens of the LDV and the laser point on the mirror is estimated initially, and will be refined later. Giving n 3D points, 3n linear equations that include n+14 unknowns are constructed. Therefore, at least 7 points are needed. After the calibration of the multimodal sensing system, the camera and LDV can be precisely coordinate to find the appropriate surfaces and to focus the laser beam, based on the distance measurements and 3D geometry.

3. AUDIO BACKGROUND MODELING

The main difficulty in audio event detection is linked to the environmental noise that is often non-stationary and may be loud comparing to the audio events to be detected. As we have discussed, the LDV is capable of detecting the acoustic signals from different kinds of vibration surfaces, including window frames, concrete walls, building pillars, telephone poles, traffic posts, etc.; however, a reliable background modeling technique should be employed in order to separate the outliers from the background "sound" that includes both the real background sound and the signals created by the electronic-optical effects of the LDV. A Gaussian Mixture Model (GMM) Φ is commonly used to model the feature distribution of signals using a weighted summation of a Gaussian distribution N. In this paper, the audio feature vector is generated using Mel-frequency cepstral coefficients (MFCCs) which are the perceptually representation of the frequency band response of the human auditory system. The likelihood of a feature vector x is defined as $\Phi(\mathrm{x}) = \sum_{k=1}^{K} \alpha_k N(x, \mu_k, \Sigma_\mathrm{k})$, where μ_k and Σ_k are mean and covariance matrix of kth Gaussian among K Gaussians, and α_k is a normalizing factor in range between 0 and 1. Due to the variations of the vibration properties from surface to surface, the GMM on each selected surface are constructed differently. They may have different numbers of components K. For example, in Fig. 2 & Fig. 3, the audio samples obtained in type A surface may contains 1 Gaussian, where the audio samples in type B surface may contain 2 Gaussians. However, the values of MFCCs in type A may be very close to the values of MFCCs in one of type B's components. In Fig. 3, the MFCC shape (in red) of type A component is very similar to the top shape (in green) of type B component. So, if we accidently use the similar background model for type B surface to detect the audio event acquired using type A surface, we may produce unreliable results. Therefore we need to select the right model for the right surface, and evaluate the correctness. The evaluation part is shown in the experiment section. Also note that the GMM can model the feature distribution, however, it cannot present temporal dependencies of each component. In order to model the internal dynamic between the components distribution in temporal domain, we use a score-based aggregation technique for the GMM with more than one component.

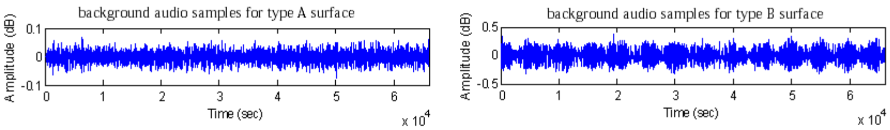

Figure 2. Background audio samples obtained from the surface type A (left) and surface type B (right)

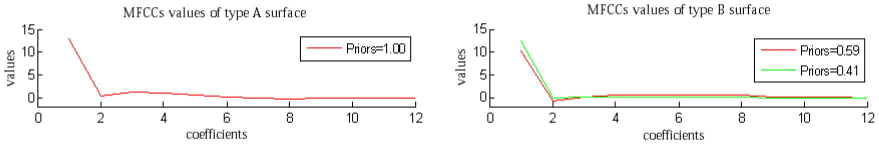

Figure 3. MFCC values of type A (left) which contains one component, of type B (right) which contains two components.

The intuitive use of the aggregation is to keep high true positive rate (TPR) of an audio sample using the right background model but decrease the TRP of an audio sample using a wrong background model. The technique works as follows. A sequence of overlapped windows is selected with each has size W and contains m consecutive features $f_1f_2 \ldots f_m$ in time series. Each feature f_j yields a score s_j that can be either the foreground or one of the background components. The normalized histogram $H_{W(i)}$ of all scores in the current window is constructed. A reference histogram H_p is created initially that presents the priors of the right background model's components. If the distance between H_p and $H_{W(i)}$ is not satisfied for a threshold, then convert all scores $s_1s_2 \ldots s_m$ to the foreground since they are not using the right background model; otherwise keep all scores and evaluate on the next window $H_{W(i+1)}$. The temporal pattern also needs to be updated once a new set of background data is acquired.

4. EVENT DETECTION USING BOTH AUDIO AND VIDEO

With the background models of various surfaces, we are able to better separate the foreground as "outlier" from the background to detect acoustic "events" if the correct background models are used. The use of video detection can verify the detection in the audio, whenever the target can be seen. In some other times, the target may be hidden behind objects such as walls, doors; in these cases the LDV audio is the only source of information. Furthermore, visual sensor is used to select reflecting surface regions to point the laser beam in order to obtain the audio signals of a nearby target (human or other subject). In object detection, moving objects (usually humans) are segmented from the scene background by using an adaptive background subtraction algorithm[14]. A reference background model with the first several frames without moving targets is initialized. In each scene, several reflection surface regions and their pointing locations for the LDV are trained using the initial background model. The 3D positions of these locations are known from our calibrated system. Only one of the surface locations next to the detected target is selected each time to point the LDV laser for collecting the audio signals. If the target moves far away from the location, another surface location close to the target is selected from the background image model (Fig. 4). The need of reselecting a surface location can be easily determined if no foreground audio event is detected but there is still a moving object detected in the video.

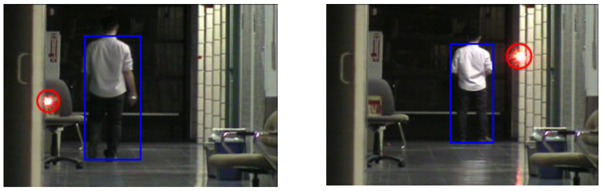

Figure 4. The image on the left shows a reflecting surface (in red) next to the selected target (in blue) that is selected. The right image shows another surface (in red) is selected when the target (in blue) moves away from the previous selected surface.

The LDV can obtain audio signals at a large distance and not be affected by other sounds along the path. However, it is sensitive to the roughness and reflectance of the reflecting surfaces. Our approach is to use the background models of the surface on which the LDV measures the audio signals. Once GMM background models are constructed for all surface candidates, the foreground events can be discriminated from the background when using the right models. Then GMM-Bayesian Classifier (GMM-BC) is used to assign an unknown feature x to a class label $l_x = \text{argmax}\ _{c}p(x|c), c = 1,2, \ldots M$, where M is the number of classes, and $p(x|c)$ is the likelihood function of x. In this paper, we only have two classes for an input feature: background (no-event) or foreground (audio event).

The integration of the object detection using both the audio and video information is made at the decision level. There are four possible cases: 1) if a human target can be detected from both the audio and video, then the confidence of detection is enhanced; 2) if the target can only be detected using the audio information but cannot be observed from the video, then the decision is made based on the audio; 3) if the target is not detected in the audio, but there is an detection in the video, then the decision is made based on the video; 4) if neither audio nor video detects the target, no event is detected. In order to perform further data association between audio and video detection results, higher level understanding of the detected objects are needed, such as visual identification and voice identification of humans. This is our ongoing work.

5. EXPERIMENTS

5.1 Indoor audio background modeling and audio event detection

The first dataset was generated using an audio source of 40-second of President Obama's speech. The audio clip is annotated into the background parts and foreground parts for providing the ground truth data for evaluation. The source was continuously played and returning signals were detected by the LDV at 5 pairs of different surfaces (Fig. 5), each with and without retro-reflective tape treatment. The corresponding 10 background surface models were constructed using the segments within the audio clips that do not include foreground sounds using the previous discussed technique. The features were generated using the first 12 MFCC, with 40% overlap of each 0.01s window. All the collected clips were normalized and aligned at the same starting point so that the comparisons of different testing clips using various background models were possible. The true positive rates (TPRs) in Table 1 indicate the correction rates of correctly separating the foreground and the background of the 10 testing clips using various background models. It can be seen that using the right background model for the audio data collected from the corresponding background surface yields the highest detection rate. The zero TPR of a background model indicates the entire clip is completely classified as foreground (when using a wrong background model). Note that the parameters of the GMM-Bayesian Classifier are tuned so that only the background model corresponding to the right background surface can be employed for detecting acoustic events (outliers). When an unknown surface that does not have a background model has to be selected for acquiring audio signals, the background model of a similar surface may give many false alarms in audio event detection. We can easily solve this problem by relaxing the parameters for the background models. From Table 1, we observe that better results are obtained with retro-reflective for most materials, and better results for most metal objects (i.e., metal box, painted metal wall). This indicates that we should use retro-reflectance treatment whenever possible, otherwise should find the closest metal surfaces in order to do the best possible remote voice detection.

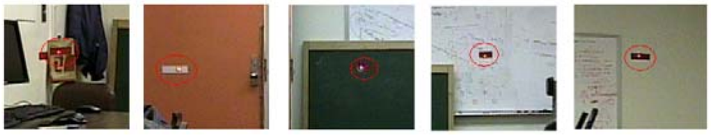

Figure 5. From the left to the right, the different surface materials are: metal box, door, chalkboard, whiteboard and wall. The red circle shows the retro-reflective tape

Table 1. TPR of test clips using updated background GMMs with aggregation

	bgGMMs \ Clip	1	2	3	4	5	6	7	8	9	10
1	Metal box with tape	0.8575	0.0000	0.0006	0.0000	0.0011	0.0000	0.0000	0.0000	0.0000	0.0000
2	Metal box without tape	0.0000	0.8041	0.0020	0.0006	0.0000	0.0025	0.0000	0.0116	0.0003	0.0059
3	Door with tape	0.0008	0.0017	0.8104	0.0000	0.0008	0.0000	0.0014	0.0000	0.0057	0.0006
4	Door without tape	0.0000	0.0003	0.0000	0.7841	0.0000	0.5300	0.0000	0.0023	0.0000	0.1819
5	Chalkboard with tape	0.0342	0.0000	0.0232	0.0000	0.6905	0.0000	0.0000	0.0000	0.0000	0.0000
6	Chalkboard without tape	0.0000	0.0003	0.0000	0.5047	0.0000	0.6674	0.0000	0.0003	0.0000	0.3775
7	Whiteboard with tape	0.0000	0.0000	0.0006	0.0000	0.0000	0.0000	0.8920	0.0000	0.2227	0.0000
8	Whiteboard without tape	0.0000	0.0113	0.0000	0.0003	0.0000	0.0011	0.0000	0.8206	0.0000	0.0003
9	Wall with tape	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.1467	0.0000	0.6993	0.0000
10	Wall without tape	0.0000	0.0028	0.0000	0.0749	0.0000	0.3372	0.0000	0.0000	0.0000	0.5302

5.2 Long range audio-visual event detection

The second experiment is performed at a 420-feet corridor for the long range audio-visual event detection. We collected the data at 120 feet, 300 feet, 420 feet with various surfaces selected for LDV measurements, including a utility box, a glass bottle, side walls, doors, a metal box, and etc. A person was reading an article along the corridor. Various cases were designed, including 1). Speaking before wall or door; 2) Walking into the scene while speaking; 3) Standing still while speaking; 4) Standing still while stopping talking; 5) Walking out of scene. In Fig. 6, experimental results for the 420-feet measurements are demonstrated. The red box shows a person speaking behind the wall therefore cannot be detected by visual scene analysis but through audio signal detection. The blue box shows the people detection in the camera view. The shaded portions of the audio stream show the foreground audio event (i.e., a person speaking). We are able to thus track a person as they walk through a part of a building even when we lose sight of them, and recapture them when they become visible or audible again.

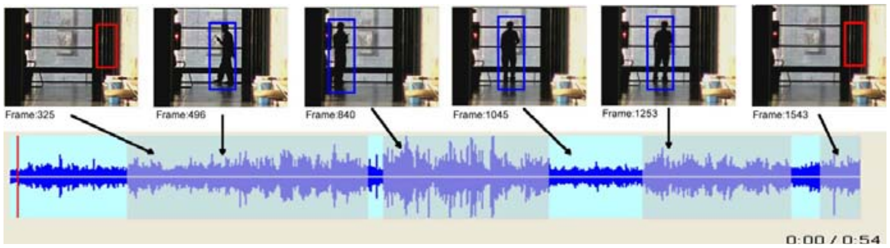

Figure 6. Audio-video integration in a 420-feet corridor

6. CONCLUSIONS

We presented a novel long range audio-visual event detection approach using an LDV and a PTZ camera. The system framework and the calibration between multiple sensor modalities were described. The audio background modeling is based on GMMs. The models capture the vibration properties of various reflecting surfaces measured by the LDV, and are important to obtain robust foreground audio event. The evaluation on different surfaces was presented. The processing of the audio and video detection and their integration was described with long-range experiments. In the future, we will further develop techniques to performance target identification by exploring higher level features of both audio and video data.

ACKNOWLEDGMENTS

The work of the first and second authors is supported in part by AFOSR under Award #FA9550-08-1-0199, ARO under a DURIP Award #W911NF-08-1-0531, NCIIA under an E-Team Award (#6629-09) and by NSF under grant No. CNS-0551598

REFERENCES

[1] Tian, Y., Senior, A. W., Hampapur, A., Brown, L., Shu, C., and Lu, M., "IBM smart surveillance system (S3): event based video surveillance system with an open and extensible framework," *Machine Vision and Applications*, 2008.

[2] Boiman, O. and Irani, M., "Detecting irregularities in images and in video," In *Proc. IEEE International Conference on Computer Vision*, pp. 1985-1988, Beijing, China, Oct. 15-21, 2005.

[3] Zotkin, D. N., Raykar, V. C., Duraiswami, R., and Davis, L. S., "Multimodal tracking for smart videoconferencing and video surveillance," In Z. Zhu and T. S. Huang (Eds.), *Multimodal Surveillance: Sensors, Algorithms, and Systems*. pp. 141-175, Norwood, MA: Artech House, 2007.

[4] Cristani, M. Bicego, M., and Murinon V., "Audio-visual event recognition in surveillance video sequences," *IEEE Transactions on Multimedia*, vol. 9. no. 2, pp. 257-267, February, 2007.

[5] Dedeoglu, Y., Toreyin, B. U., Gudukbay, U., and Cetin, A. E., "Surveillance using both video and audio," In P. Maragos et al. (Eds.) *Multimodal Processing and Interaction*, Springer, ch. 6. Science+Business Media, LLC, 2008

[6] Radhakrishnan, R. and Divakaran, A., "Generative process tracking for audio analysis," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP' 06), vol. 5, May, 2006

[7] Polytec Laser Vibrometer, http://www.polytec.com/. Last visit: 2009

[8] Li, W., Liu, M., Zhu. Z. and Huang, T. S., "LDV remote voice acquisition and enhancement," ICPR 2006: 262-265.

[9] Zhu, Z. and Li, W., "Integration of laser vibrometry with infrared video for multimedia surveillance display," AFRL Report: http://www-cs.ccny.cuny.edu/~zhu/LDV/FinalReportsHTML/CCNY-LDV-Tech-Report-html.htm, 2005.

[10] Qu, Y., Wang, T. and Zhu, Z., "Remote audio/video acquisition for human signature detection", in CVPR'09 Biometrics: 66-71, 2009

[11] Mei, C. and Rives, P., "Calibration between a central catadioptric camera and a laser range finder for robotic applications," in *Proceedings of ICRA06*, Orlando, May, 2006.

[12] Wasielewski, S. and Strauss, O., "Calibration of a multi-sensor system laser rangefinder/camera," Proc. of the Intelligent Vechicles Symposium, pp. 472-477, 1995.

[13] Zhang, Z., "A flexible new technique for camera calibration," *IEEE Trans. On Pattern Analysis and Machine Intelligence*, vol. 22, no. 11, pp. 1330-1334, 2000.

[14] Yang, S. Y. and Hsu, C. T., "Background modeling from GMM likelihood combined with spatial and color coherency", *IEEE ICIP*, 2006