Remote audio/video acquisition for human signature detection

Article · June 2009

DOI: 10.1109/CVPRW.2009.5204294

3 authors, including:

Yufu Qu
Beihang University (BUAA)
42 PUBLICATIONS **232** CITATIONS

SEE PROFILE

Remote Audio/Video Acquisition for Human Signature Detection

Yufu Qu[1], Tao Wang[1,2] and Zhigang Zhu[1,2]

[1]Department of Computer Science, The City College of New York
138th Street and Convent Avenue, New York, NY 10031

[2]Department of Computer Science, The CUNY Graduate Center
365 Fifth Avenue, New York, NY 10016

Email: {qu, twang, zhu}@cs.ccny.cuny.edu

Abstract

To address the challenges of non-cooperative, large-distance human signature detection, we present a novel multimodal remote audio/video acquisition system. The system mainly consists of a laser Doppler virbometer (LDV) and a pan-tilt-zoom (PTZ) camera. The LDV is a unique remote hearing sensor that uses the principle of laser interferometry. However, it needs an appropriate surface to modulate the speech of a human subject and reflect the laser beam to the LDV receiver. The manual operation to turn the laser beam onto a target is very difficult at a distance of more than 20 meters. Therefore, the PTZ camera is used to capture the video of the human subject, track the subject when he/she moves, and analyze the image to get a good reflection surface for LDV measurements in real-time. Experiments show that the integration of those two sensory components is ideal for multimodal human signature detection at a large distance.

Keywords – Multimodal sensing, human recognition at a distance, audio-visual integration.

1. Introduction

Remote human signature detection and identification is becoming increasingly important in non-cooperative and hostile environments for applications such as large area surveillance, perimeter protection, and search and rescue in the fields. A few systems [1, 2] have been reported for this goal, particularly with the rapid improvements of color and infrared (IR) cameras and the corresponding algorithms for monitoring subjects at a large distance. Although video technologies (including visible and IR) have a great advancement in human detection at a large distance, there is still a serious limitation in non-cooperative and hostile environments because of intentional camouflages and natural occlusions. Audio information, another important data source for identifying humans, has not yet been well explored for remote human signature detection. Zotkin, and et al [3] and also Zou and Bhanu [4] have reported the integrations of visual and acoustic sensors. But in these systems, the acoustic sensors (microphones) need to be close to the subjects in monitoring. Parabolic microphones, which can capture voice signals at a fairly large distance in the direction pointed by the microphone, can be used for remote hearing and surveillance. But it is very sensitive to noise caused by wind or sensor motion, and all the signals on the way get captured.

Commercial Laser Doppler Vibrometer [LDV] such as those manufactured by Polytec [5] and B&K Ometron [6] can effectively detect vibration within two hundred meters with sensitivity on the order of 1μm/s. Larger distances could be achieved with the improvements of sensor technologies and the increasing demands of applications. The voice signals of a human could be acquired by capturing the vibration of a target surface that is caused by the speech of the person next to the target. Li, and et al [7] have presented their results in detecting and processing voice signals of people from large distances using an LDV. However, in their work, a user has to manually adjust the LDV sensor head in order to aim the laser beam at a surface that well reflects the laser beam, which has been proven to be a tedious and difficult task. In addition, it is very hard for the user to see the laser spot at a distance above 20 meters, and so it is extremely difficult for the human operator to aim the laser beam of the LDV at a distant target. The goal of this paper is to integrate the LDV with a pan-tilt-zoom (PTZ) camera, which can aid the LDV in finding a reflection surface automatically, and consequently, both video and audio signals can be captured synchronously for multimodal human signature detection, and further for human identification /verification with both audio and video information.

The organization of the paper is as the following. Section 2 introduces the basic principles and important issues of remote audio/video detection using an LDV sensor and a PTZ camera. The remote audio/video acquisition system is presented in Section 3. In Section 4, the method of video detection and analysis is presented. Section 5 discusses the automation reflection surface selection for audio detection. Section 6 presents some preliminary experimental results. Finally, we conclude and discuss our work in Section 7.

2. Remote A/V Detection: an Overview

In this section, we introduce the basic principle and important issues of remote audio/video (A/V) detection using a PTZ camera and a LDV sensor.

The use of the PTZ camera is to acquire visual information of targets at a large distance, obtaining a suitable image resolution of the targets with its zoom capability while keeping those targets inside the field of view (FOV) using its pan/tilt capabilities. Here, the targets include both human subjects and their surrounding objects. However, solely using visual information is not sufficient to recognize or to identify human subjects at distances; moreover, it is well known that robust visual recognition is challenging due to variations of illuminations, changes of background, shadowing, objects occlusions, segmentation errors, low spatial and temporal video resolutions, and perspective distortions. Therefore, adding audio information should improve the performance of human signature detection at distance. However, the microphone-type acoustic sensors are sensitive to noises caused by sensor and wind motion, and more unfortunately, they usually cannot detect audio signals from a long distance. Therefore, we propose to use a novel sensor – the laser Doppler vibrometer (LDV) – to detect voice signals at a large distance. There are two advantages of using LDV instead of a microphone. First, only voice signals that cause vibrations of the surface that the laser beam of the LDV focuses on can be captured and heard. Second, it can detect sound in a long distance.

The LDV works according to the principles of laser interferometry. Measurements are made at the point where the laser beam strikes the structure under vibration. In the Heterodyning interferometer (Figure 1), a coherent laser beam is divided into object and reference beams by a beam splitter BS1. The object beam strikes a point on the moving (vibrating) object and light reflected from that point travels back to beam splitter BS2 and mixes (interferes) with the reference beam at beam splitter BS3. If the object is moving (vibrating), this mixing process produces an intensity fluctuation in the light as

$$I_1 = \frac{1}{2}A^2\left\{1-\cos\left[2\pi\left(f_B+\frac{2v}{\lambda}\right)t\right]\right\} \quad (1)$$

A detector converts this signal to a voltage fluctuation. Most objects vibrate while wave energy (including that of voice waves) is applied on them. Though the vibration caused by vibration, this vibration can be detected by the LDV. The relation of voice frequency f, velocity v and magnitude m of the vibration is as the following.

$$v = 2\pi fm \quad (2)$$

As seen from the above principle of the LDV, There are two important issues to be considered in order to use a LDV to measure the vibration of a target caused by human voices. First, the target vibrates with the voices. Second, the intensity of reflection laser beam back to the LDV should not be very weak, otherwise the contrast of interferometric fringe will very low for detecting human voices. Therefore, it is crucial to find an appropriate surface close to the human subject who is in talking so that both requirements can be met. So the PTZ camera will be used not only to obtain visual information of human subjects (the first type of targets), but also to aid the detection of surfaces of vibrating and reflecting targets in the environment (the second type of targets).

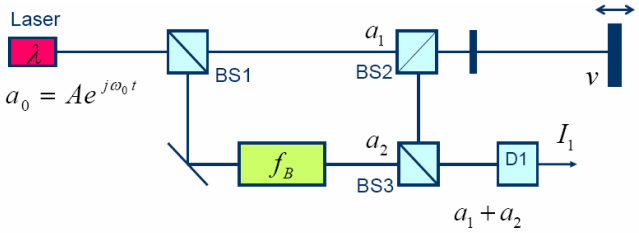


Figure 1: The modules of the Laser Doppler Vibrometer (LDV)

3. A Remote A/V Acquisition System

The schematic setup of our remote audio/video acquisition sensors system is shown in Fig. 2. It consists of a LDV sensor on a Pan-Tilt Unit (PTU), a PTZ camera and a personal computer (PC).

The LDV from Polytec [5] includes a controller OFV-5000 with a digital velocity decode card VD-6 and a sensor head OFV-505. The sensor head of the LDV uses a HeNe red laser with a wavelength of 633.8 nm and is equipped with a super long-range lens. It sends the interferometry signals to the controller, which is connected to the computer via an RS-232 port. The controller box processes signals received from the sensor head of the LDV, and then output signals to computer using S/P-DIF output.

The PTZ camera - Cannon VC-C50i has a 720×480 focal plane array and an auto-iris zoom lens that can change from

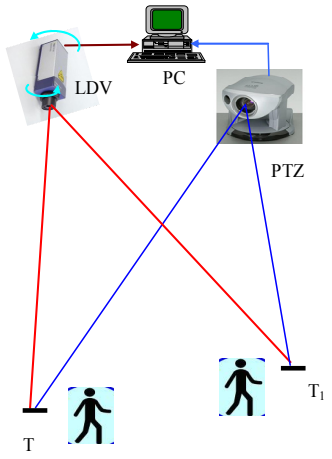


Figure 2: Schematic setup of the remote audio/video acquisition sensor system

3.5mm to 91mm (26× power zoom). The pan angle of the PTZ is ±100º with rotation speed 1-90º per second and the tilt angle of it is from -30º to +90º with rotation speed 1-70º per second. The camera can expand the sensitivity into the infrared range, making the camera possibly work in night condition.

The PTU is PTU-D46-17 of Directed Perception, Inc. It has a pan range from -159º to +159º and a tilt range from -47º to +31 º. Its rotation resolution is 0.013 º and max rotation speed is 300º/s.

The LDV which mounts on the top of the PTU collects voice signals and the PTZ captures image signals. Both audio and video signals are sent to the PC, via the S/P-DIF and 1394 interfaces of the PC, respectively. The PC sends commands via RS-232 ports to control, the PTU, the PTZ, and the LDV. The other function of the PC is to analyze the received signals to detect audio/video human signatures.

The process flow diagram is shown in Fig. 3. The LDV and PTZ capture voice and image signals synchronously and in real-time send them to the PC to analyze the signals. If the system detects moving objects, then candidates for possible human targets are generated for human signature detection. For each moving target, or a set of moving targets, the system analyzes their background surrounding to select an appropriate reflection surface, and then control the LDV laser beam to point to the surface. Upon the analysis of the return signals of the LDV, the LDV captures vibration signals and sends them to the PC. The system analyzes the vibration signals and tries to extract voice signals of a subject from them. Meanwhile, visual information is also obtained, such as faces, bodies or gaits of the human. If the target moves, the PTZ tracks the target and aid the LDV to re-select a reflection surface and to collect voice signals. Note that the orientations of the LDV and PTZ are controlled separately since they need to aim to and track different targets, human subjects and reflection surfaces, respectively. A patent [8] has been filed on vision-aided LDV voice detection, and automatic LDV focusing based on distance measurement, signal strength

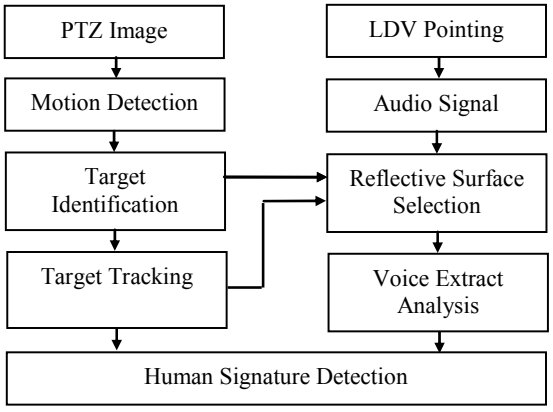


Figure 3: Diagram of remote audio/video acquisition

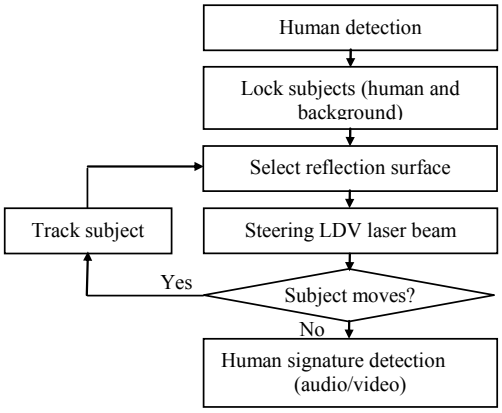


Figure 4: Flow chart of video analysis

and laser spot image information.

4. Video Detection and Analysis

The first step of video detection is to discriminate moving targets from its background which is continuously updated from the PTZ camera. Any conventional shaped-based and/or motion-based feature extractions methods [9] and [10] are capable of detecting human targets visually. When a possible human subject is found in the surveillance perimeter, an appropriate reflection surface need be found to guide the LDV pointing direction for acquiring audio signals of the human subject. The flow chart of video analysis is shown in Fig. 4. When a human subject is found, the PTZ first locks on to the human subject and zooms in to obtain a clear image of the subject. The ideal video images should include both the human subject and certain portion of its background to search for reflection surface. Second, the system pre-processes and segments the color image to get a reflection surface region for the laser of the LDV to point on. The suitable image region should be reflective - smooth in color, better with large red components, close to the human subject in order to vibrate with human voice, and large enough to aim the laser on. Finally, the system controls the LDV laser beam to point onto the reflection surface in order to capture voice signals; this will be further described in the next section. At the same time, the PTZ camera tracks the human subject in real-time. If the subject moves, the system controls the PTZ to pan, tilt or zoom to keep the tracked subject large enough and within its FOV. Based on the signal levels of the LDV vibration signals and the closeness of the reflection surface to the tracked subject, the system judges whether the reflection surface needs to be re-selected. If the answer is yes, the system re-obtains a better reflection surface. As a result, the optimal audio and video signals can be acquired at the same time for multimodal human signature detection.

5. LDV Voice Detection

5.1. Reflection surface selection

The selection of reflection surfaces for LDV signals collection is important since it is a major factor that determines the quality of acquired vibration signals. There are two basic requirements for a good surface: vibration to the voice energy and reflectivity to the HeNe laser. We have found that almost all natural objects vibrate more or less with normal sound waves. Therefore, the key technique in finding a good reflection surface is to measure its reflectivity. Based on the principle of the LDV sensor, the relatively poor performance of the LDV on a rough surface at a large distance is mainly due to the fact that only a small fraction of the scattered light (approximately one speckle) can be used because of the coherence consideration. A stationary, highly reflective surface usually reflects the laser beam of the LDV very well. Unfortunately, the body of a human subject does not have such good reflectivity to obtain LDV signals unless (1) it is treated with retro-reflective materials; and (2) it can keep relatively still to the LDV. Therefore, we tested two approaches. One approach is to set a reflection surface on the human target. For example, if a human subject can be treated with a small piece of retro-reflectance tape, which has very high reflectance, we can aim the laser beam to the tape and then track the tape while the human body moves, by analyzing the video of the PTZ camera. Another approach is to select a background object nearby to the human subject in order to detect speech signals. Typically, a large and smooth color-segmented background region can be the selection for the LDV pointing location.

Fig. 5 illustrates the result of color segmentation of an image into multiple regions, including the human subject and its surrounding background. The point H is the center of the human subject, whereas the point B_i is the center of one of the background regions. A point L_i is selected on the line HB_i with a pre-defined distance from the boundary of the human subject H (a few pixels are set in our current experiments). Then the system controls the LDV to point its laser beam to the position L_i (which is as close as possible to H) and to obtain a signal level S_i of the return signals. Similarly, we obtain a series of signal levels S_1, S_2 …Sn, when the laser beam points to other neighboring regions to the human subject. Comparing the values of these signal levels, the kth region is selected, which has the maximum signal level S_k. Then the LDV points to the corresponding location L_k of the kth region to capture the best audio signals among all of these neighboring regions.

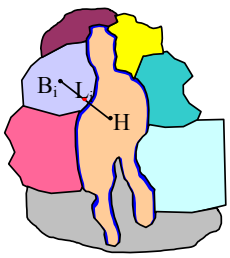


Figure 5: Illustration of the LDV pointing location

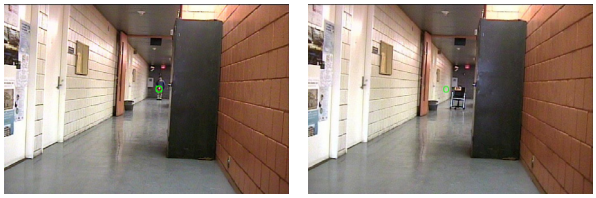
5.2. Voice signal enhancement

For the audio signals of human voices, the frequency range is about 300 Hz to 3 KHz. However, the frequency response range of the LDV is much wider than that. Even if we have used the on-board digital filters, we still get signals that include troublesome large, slowly varying components corresponding to the slow but significant background vibrations of the targets. The magnitudes of the meaningful acoustic signals are relatively small, adding on top of the low frequency vibration signals. This prevents the intelligibility of the acoustic signals by human ears. On the other hand, the inherent "speckle pattern" problem on a normal "rough" surface and the occlusion of the LDV laser beam (by passing-by objects) introduce noises with large and high-frequency components into the LDV measurements. This creates very high and loud noise when we directly listen to the acoustic signals. Therefore, we have applied a Gaussian bandpass filter to process the vibration signals captured by the LDV. In addition, the volumes of the voice signals may change dramatically with the changes of the vibration magnitudes of the target due to the changes of speech loudness (shouting, normal speaking, whispering) and the distances of the human speakers to the target. Therefore, we have also designed an adaptive volume function to cope with this problem.

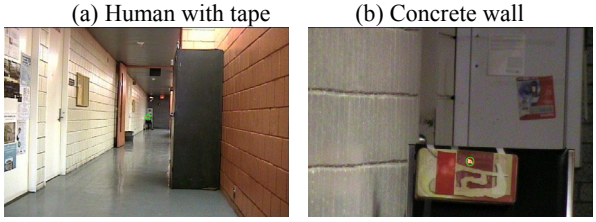
6. Experimental Results

6.1. Comparison on different reflection surfaces

In order to obtain good voice signals generated by human subjects in a long distance, we have performed a few experiments in the corridor close to our lab to analyze and compare the signal levels of the LDV and the audio quality of the acquired signals. The surfaces we tested include human body with retro-reflective tape treatment, stationary objects (wall, white board, metal cake box, etc.) with retro-reflective tape treatment, and those stationary objects without such treatment, such as concrete wall, metal cake box, red paper and white paper pasted on a utility box, steel file cabinet and black leather chair back. The objects without retro-reflective treatment are "natural" (uncooperative), and exist in environments. We tested both their vibration and reflection properties for acquiring human voice signals. In all the experiments, the LDV velocity range is set to 1 mm/s, and the human subject was

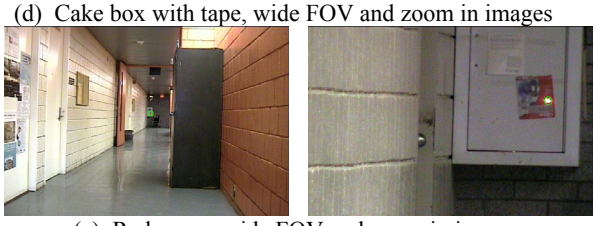


(a) Human with tape (b) Concrete wall

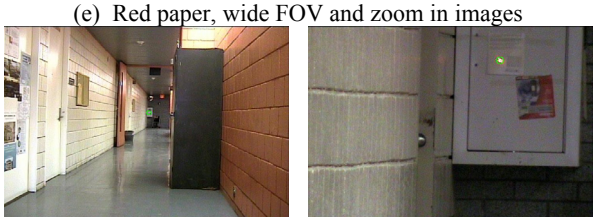


(c) Metal cake box, wide FOV and zoom in images

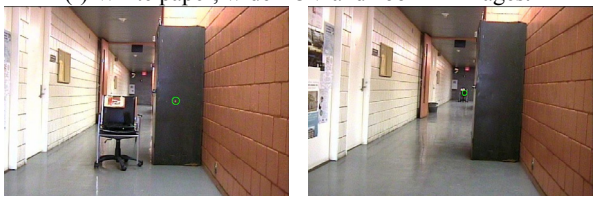
(d) Cake box with tape, wide FOV and zoom in images



(e) Red paper, wide FOV and zoom in images



(f) White paper, wide FOV and zoom in images.



(g) Steel file cabinet (h) Black leather chair back

Figure 6: Various reflection surface experiments
(The laser spot is within a green circle in each image)

about 30m to the LDV sensor head.

We placed a Sony IC player near the LDV pointing surface and play a short clip of the Obama weekly address as the baseline source clip. When the audio player played the address repetitively, the LDV sensor captured the voice signals, followed by audio signal enhancement including band-pass filtering (300 – 3000 Hz) and adaptive volume changes. The PTZ camera views of both wide FOV and zoom images of each surface are shown in Fig. 6. Table 1 listed the signal levels (from 1 to 512) and video qualities of the LDV signal acquisition of the aforementioned objects, listed into four groups: stationary objects with retro-reflective tape treatment, human body with tape treatment, natural objects with medium signal returns, and natural objects with poor signal returns. Typical audio clips corresponding to these four groups, before and after audio signal enhancements, can be played following the links in our supplemental materials [11]. From the experiment results, we have the following observations:

Table 1 Experiment results on different surfaces

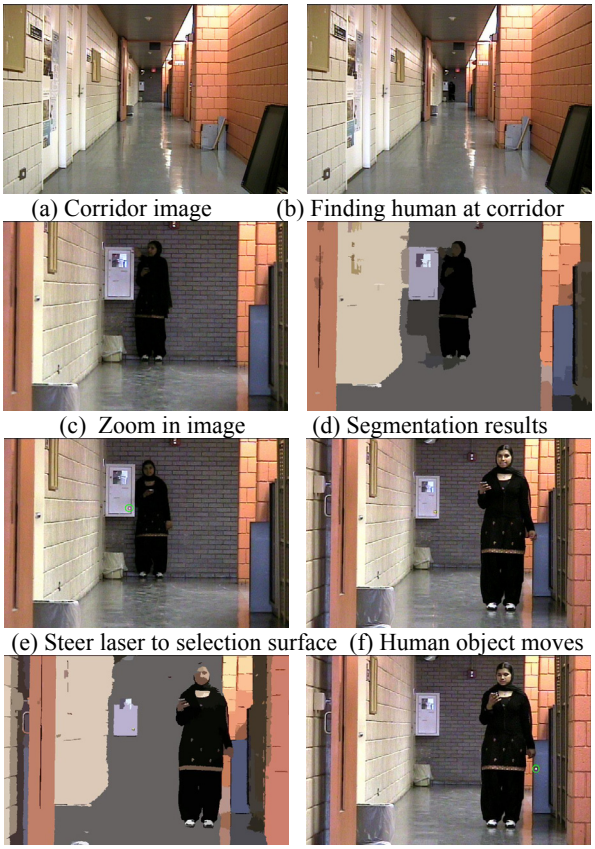
Surfaces	Signal level (1-512)	Classes of surfaces	Audio quality
Stationary objects with tape	512	Cooperative Stationary	Good
Human body with tape	512	Cooperative Moving	Noisy
Red paper	28	"Natural" Medium return	Okay
Metal cake box	25		
Steel file cabinet	15		
White paper	1	"Natural" Weak return	Poor
Black leather	1		
Concrete wall	1		

(1) If a retro-reflectance tape is placed on the body of a human subject, the voice signals (either of the speech of the person, or a play of an audio clip) can be acquired. However, this almost places a requirement that the human subject is cooperative. Moreover, the movements of the human target during speech may cause the laser beam to briefly stray off the tape, therefore large noises could be heard in some places. As such, even placing a retro-reflectance tape on the body of the human body, the moving human body is not an ideal selection.

(2) Searching a good vibration and high reflectance surface around the human target, even without retro-reflective tape treatment, is a favorable scheme to detect audio signals in a non-cooperative environment. We have found that if the signal level is medium (above 10 in the range of 1 to 512), we can always obtain intelligible audio signals even if the signals have various degrees of noise (after the signal enhancement process). In the objects we have tested, all of still subjects with the retro-tape treatment, and some of the natural objects - the red paper, the metal cake box and the steel file cabinet all have good signals returns, whereas the concrete wall, the white paper and black leather chair back do not perform well.

6.2. Preliminary results on A/V acquisition

Some preliminary experimental results on remote audio/video tracking and adaptive acquisition have been obtained on the same corridor outside our lab. The corridor was thought (Fig. 7 (a)) as a "non-cooperative"



(a) Corridor image (b) Finding human at corridor
(c) Zoom in image (d) Segmentation results
(e) Steer laser to selection surface (f) Human object moves
(g) Surface in new position (h) Steer laser to a nearby surface

Figure 7: Experiment results of remote A/V
(The laser spots are within the green circles in (e) and (h))

environment since all the objects are naturally placed there. When a human subject was detected in the scene (Fig. 7 (b)), the system locked in the human target and the PTZ camera zoomed in to get a clearer image (Fig. 7 (c)), with both the person's face and some surrounding objects for projecting the LDV laser beam. Then the image was segmented into several homogenous color regions in order to find the best reflection surface next to the human subject (Fig. 7 (d)). Finally, the system controlled the LDV to point the best reflectance surface, and captured audio signals (Fig. 7 (e)). For comparison, we still used the short clip of the Obama weekly address listed in Table 1 to simulate the speech of the human subject. The original audio clip captured by the LDV and the enhanced audio clip can be played in our supplemental materials [11]. When the human subject moved to another poison, the system first tracked the human target and obtained the image (Fig. 7 (f)). Then, the captured image was analyzed and processed to acquire its neighbor regions (Fig. 7 (g)). Finally, the LDV laser beam was redirected to the region with the best reflectance (Fig. 7 (h)), and the audio signals was acquired and processes.

7. Conclusions and Discussions

We present a novel remote audio/video acquisition sensing system for human signature detection. A PTZ camera is used to remotely capture video signals for both visual information collection and the guidance of the laser vibrometry sensor to remotely obtain the corresponding audio signals. We have mainly focused on the use of PTZ images to intelligently track human subjects and then to automatically select the best reflection surface for LDV listening. Thus, we not only enhance the remote surveillance from solely visual surveillance to truly multimodal surveillance with both audio and video capabilities, but also advance the efficiency and performance of the LDV sensor for remote video detection.

Ongoing work includes the automatic focusing based on the calibration of the PTZ camera and the LDV for using stereo triangulation to obtain distance information of a reflecting surface to rapidly bring the laser beam into focus so that optimal acoustic signal can be acquired [8]. Further experiments will be performed in a real non-cooperative environment and longer distance (more than 200m) to verify the performance of the presented remote audio/video acquisition system. Using the collected data, we are also interested in performing human identification and/or verification based on long-range video (face, gait, etc) and audio (speech) recognition.

8. Acknowledgments

This work is supported by the AFOSR Discovery Challenge Thrusts (DCTs) under Award No. FA9550-08-1-0199 and by the NSF Computing Research Infrastructure program under Grant No. CNS-0551598.

References

[1] http://www.z-shiny.com.cn/enGLN-810418E.asp
[2] X. Li, G. Chen, Q. Ji, B., Erik. A non-cooperative long-range biometric system for maritime surveillance, ICPR 2008
[3] D. Zotkin, R. Duraiswami, H. Nanda, L. Davis, Multimodal tracking for smart videoconferencing, IEEE ICME 2001
[4] X. Zou and B. Bhanu, Tracking humans using multimodal fusion, OTCBVS'05
[5] Polytec Laser Vibrometer, http://www.polytec.com/
[6] Ometron Systems. http://www.imageautomation.com/
[7] W. Li, M. Liu, Z. Zhu and T. S. Huang, LDV remote voice acquisition and enhancement, ICPR 2006
[8] Z. Zhu, Y. Qu and T. Wang, Vision-aided automated vibrometry, U.S. Prov Patent Appl. No. 61/163,169, March 25, 2009
[9] B. Leibe, E. Seemann, and B. Schiele. Pedestrian detection in crowded scenes, CVPR2005
[10] L. Zhao and L. Davis. Closely coupled object detection and segmentation, ICCV 2005
[11] http://www-cs.ccny.cuny.edu/~zhu/LDV2009