An Active Multimodal Sensing Platform for Remote Voice Detection

Yufu Qu, Tao Wang, *Student Member, IEEE*, and Zhigang Zhu, *Senior Member, IEEE*

***Abstract*—To improve the performance and the efficiency of Laser Doppler Vibrometers (LDVs) for long-range hearing, we design an active multimodal sensing platform that integrates a Pan-Tilt-Zoom (PTZ) camera, a mirror and a Pan-Tilt-Unit (PTU) to the LDV. With the assistance of the vision and active control components, the LDV can automatically select the best reflective surfaces, point the laser beam to the selected surfaces, and quickly focus the laser beam. For accomplishing these functions, distance measurement and sensor calibration methods are proposed using the triangulation between the PTZ camera and the mirrored LDV laser beam. Based on both the measured distances and the return signal levels of the LDV, a fast and automatic LDV focusing algorithm is designed. Furthermore, strategies of surface selection and laser pointing are designed for the platform to automatically point the laser beam to the designated surfaces. Experimental results are shown to validate the performance improvement of the LDV in remote automatic voice detection by using the active multimodal sensing platform.**

I. INTRODUCTION

Acoustic sensing and event detection is receiving a growing interest by the scientific community. It can be used for audio-based surveillance, including intrusion detection [1], abnormal situations detection in public places such as banks, subways, airports, and elevators [2], [3], underwater acoustic environment monitoring [4], and etc.. It can also be used as a complementary source of information for video surveillance and tracking [5], [6], where audio-visual integration has been successfully utilized in a wide range of security applications, such as automatic speech recognition, human activity recognition and human tracking. The audio-visual integration is also used in humanoid robots in order to response to a human's voice instructions and visual behaviors [7]. However, in these systems, the acoustic sensors are typically microphones that need to be placed close to the subjects of interests. Furthermore, these acoustic sensors need to be fixed on pre-determined places. If the tracking targets move out of their sensing ranges, they will not be able to obtain any signals. Even a microphone array can only cover a limited range. Parabolic microphones, which can capture voice signals at a fairly large distance in the direction pointed by the microphone, could be used for remote hearing and surveillance. But it is very sensitive to noise caused by the wind or the sensor motion, and all the signals on the way are captured. Therefore there is a great necessity to find a new type of acoustic sensor for remote voice detection. Recently, Laser Doppler vibrometers (LDVs) have been widely used in the inspection industry. Laser Doppler vibrometers such as those manufactured by Polytec and B&K Ometron can effectively detect vibration within two hundred meters with sensitivity on the order of 1μm/s. The audio signals of a target (human or other subject) could be acquired by capturing the vibration of a surface caused by the sound of the target next to the surface. Blackmon and Antonelli [8] have tested and shown a sensing system to detect and receive underwater communication signals by probing the water surface from the air using an LDV and a surface normal tracking device. In the past few years, our lab [9], [10] has studied the detection and processing of voice signals of people from large distances using an LDV. However, in our previous work, we have found that there are two tedious and difficult tasks in manually operating the LDV. First, a user has to manually adjust and point the LDV sensor head in order to aim the laser beam at a surface that well reflects the laser beam. In outdoor data collection, 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 user to aim the laser beam of the LDV at a distant target. Second, even if the laser beam is pointed to the surface, it takes quite some time to focus the laser beam; using a COTS Polytec LDV (OFV-505), it takes about 15 seconds for a signal focusing. And there is no guarantee that the signal returns will be able to capture the voice signals in need.

In this paper, we propose an active multimodal sensing platform to automate the remote voice detection. The platform integrates the LDV with a Pan-Tilt-Zoom (PTZ) camera, and a mirror on a Pan-Tilt-Unit (PTU). The integrated multimodal system can automatically detect reflective surfaces and aim the LDV laser beam through the analysis of the video images obtained by the PTZ. The distances and orientations of the reflective surfaces can then be measured by using stereo vision between the PTZ camera and the mirrored LDV laser beam. The emitted ray of the LDV can be quickly and automatically focused by using of both the surface distances and the LDV signal levels. The system also automatically points the laser beam to the designated surfaces.

The rest of the paper is organized as follows. Section 2 first introduces the basic principles and discusses important issues of remote voice detection using an LDV sensor, and then

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

Y. Qu, T. Wang and Z. Zhu are with the Department of Computer Science, City College, City University of New York, New York, NY 10031 USA (e-mail: {qu, twang, zhu}@cs.ccny.cuny.edu).

T. Wang and Z. Zhu are also with the Department of Computer Science, Graduate Center, City University of New York.

gives an overview of our solution to these problems. The configuration of the multimodal sensing platform and the methods for depth measurement and sensor calibration are presented in Section 3. In Section 4, the automatic LDV focusing algorithm is described. Section 5 discusses the automation reflection surface selection and the laser beam pointing strategies. Section 6 presents some preliminary experimental results. Finally, we conclude our work in Section 7.

II. Vision Aided Long Range Voice Detection

A. An novel sensor and the unmet needs

An LDV works according to the principles of laser interferometry. Measurements are made at the point where the laser beam strikes the structure under vibration. Most objects vibrate while wave energy (including that of voice waves) is applied on them. Though the magnitudes of the vibration caused by voice waves are very small (usually in nanometer levels), 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.

$$v = 2\pi f m \quad (1)$$

Note that the velocity v will be large with a large frequency f, even under a small magnitude m. The Polytec LDV sensor OFV-505 and the controller OFV-5000 we use in our experiments can be configured to detect vibrations under several different velocity ranges: 1 mm/s/V, 2 mm/s/V, 10 mm/s/V, and 50 mm/s/V, where V stands for velocity. For voice vibration of basic frequency range from 300 to 3000 Hz, we usually use the 1mm/s/V velocity range. The best resolution is 0.02 μm/s under the range of 1mm/s/V according to the manufacture's specification with retro-reflective tape treatment. Without the treatment, the LDV still has sensitivity on the order of 1.0 μm/s. This indicates that the LDV can detect vibration (due to voice waves) at a magnitude in nanometers without retro-reflective treatment or even picometer with retro-reflective treatment.

There are two important issues that have to be considered in order to use a LDV to measure the vibration of a surface caused by the voice of a subject at a large distance. First, the intensity of reflection laser beam back to the LDV should be sufficiently strong, otherwise the intensities of the reference beam and the object beam will have a big difference, and consequently, the contrast of interferometric fringe will be too low for detecting the subject's sounds. Second, the spot size of the LDV laser beam on the surface should be very small. If the laser spot is large, so will the number of scattering centers and the angular dependence of the path length differences in a given direction. As the speckles thus created have different phases, they will cause speckle noise in the vibrometer signal output. This speckle noise will have strong or even overwhelming negative effects on the acquired voice signals. This indicates that whenever a new surface is selected, the LDV must be re-focused. However, the built-in automatic focusing of the LDV (if any) is usually very slow. As an example, the Polytec 505 LDV takes about 15 seconds to focus the laser beam on the surface of a target. This will be very problematic if we need to constantly switch the LDV laser beam to different surfaces for target tracking or for area search, particularly for targets in a large distance. Therefore, both surface selection and fast focusing are crucial for the LDV to acquire high-quality long-range audio signals in these scenarios.

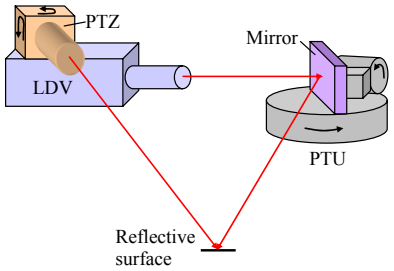


Fig. 1. System setup of the active multimodal sensing platform

B. Our solution

To solve the above two problems in using the LDV for voice detection, we design a system that integrates the LDV with a PTZ camera and a mirror mounted on a PTU. The setup of our active multimodal sensing platform is shown in Fig. 1. The direction of the laser beam of the LDV is controlled by the reflection mirror. The images captured by the PTZ camera are analyzed for detecting objects and the surrounding surfaces which are possibly good candidates for laser pointing. The triangulation method between the LDV laser beam and the PTZ camera is used to obtain the distances of those surfaces adaptively.

There are four major issues in order to obtain high-quality acoustic signals using the LDV: 1) determine an appropriate surface with sensible vibration and good reflection index; 2) measure the precise distance of the surface from the LDV; 3) focus the LDV automatically and rapidly; and 4) point the laser beam to the selected surface. In the following, we will start with the methods in distance measurement and system calibration, and then present an algorithm for automatic focusing the LDV. After that, we will propose our approach in reflection surface selection, and then describe our strategies in automatically pointing the laser beam to the selected surface.

III. Distance Measurement and Calibration

There are four hardware elements in our multimodal sensing platform: the LDV, the PTZ, the mirror and the PTU. In order to enable the system to measure ranges of surfaces with sufficient accuracy, we need to geometrically calibrate the platform. The following features of the system make such a calibration complicated. 1) The PTZ is an array sensor, whereas the LDV is a point sensor, even though they obey the

same perspective projection geometry. 2) Both the PTZ and the PTU undergo pan and tilt rotations when working. 3) The rotation center of the PTU is not on the point where the laser beam interacts with the mirror. 4) The focal length of the PTZ camera changes with its zooms. In the following, we will first build the geometrical model of the system to measure distances, and then present our method in calibrating the platform.

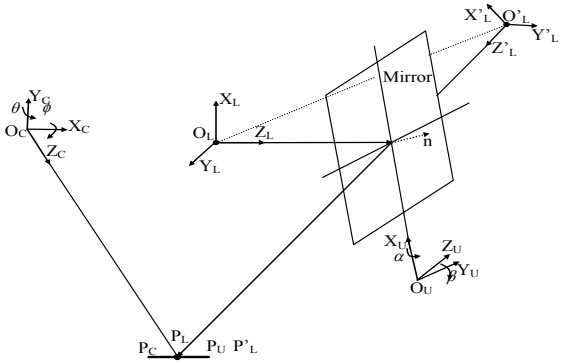


Fig. 2. Coordinate systems of the active multimodal sensing platform

A. Geometric model

The main task of the geometric modeling is to build the relationship between different coordinate systems and determine the unknown parameters need to be estimated by calibration. This multimodal sensing platform consists of four coordinate systems: the LDV coordinate system ($O_L X_L Y_L Z_L$), the PTZ coordinate system ($O_C X_C Y_C Z_C$), the PTU coordinate system ($O_U X_U Y_U Z_U$), and the mirrored LDV coordinate system ($O_L^{'} X_L^{'} Y_L^{'} Z_L^{'}$) (Fig. 2).

Given a 3D point P represented in the $O_L X_L Y_L Z_L$ as P_L, in the $O_C X_C Y_C Z_C$ as P_C, in the $O_U X_U Y_U Z_U$ as P_U, and in the $O_L^{'} X_L^{'} Y_L^{'} Z_L^{'}$ as P'_L, the relationship between the LDV and the PTZ is defined as:

$$P_L = R_C P_C + T_C = R_{C0} R_{C\theta} R_{C\phi} P_C + T_C \tag{2}$$

where R_C and T_C are the rotation and the translation matrices of the PTZ in the $O_L X_L Y_L Z_L$, with $T_C = \left(T_{Cx}\ T_{Cy}\ T_{Cz}\right)^T$. The relationship between the LDV and the PTU is defined as:

$$P_L = R_U P_U + T_U = R_{U0} R_{U\alpha} R_{U\beta} P_U + T_U \tag{3}$$

where R_U and T_U are the rotation and the translation matrices of the PTU in the $O_L X_L Y_L Z_L$, with $T_U = \left(T_{Ux}\ T_{Uy}\ T_{Uz}\right)^T$. Note that both R_C and R_U contain 3 rotation matrices: an initial rotation matrix (R_{C0} for the PTZ and R_{U0} for the PTU) which needs to be calibrated, a pan rotation matrix ($R_{C\theta}$ for the PTZ and $R_{U\alpha}$ for the PTU) and a tilt rotation matrix ($R_{C\phi}$ for the PTZ and $R_{U\beta}$ for the PTU). According to the principle of mirroring, we obtain the relationship between the "mirrored" LDV and the PTU as:

$$P_L^{'} = R_U R_{LR} P_U + T_U \tag{4}$$

where R_{LR} is the rotation matrix that converts a right hand coordinate system to a left hand coordinate system. From (3) and (4), we obtain the relationship between the LDV and the mirrored LDV as:

$$P_L = R_U R_{LR} R_U^T \left(P'_L - T_U\right) + T_U \tag{5}$$

Using the well-known triangulation method [11] between the PTZ ray and the mirrored laser ray, we can obtain:

$$\begin{aligned} & t_c R_C K^{-1} p_I - t_l R_U R_{LR} R_U^T P'_{L0} + c\left(R_C K^{-1} p_I \times R_U R_{LR} R_U^T P'_{L0}\right) \\ & = T_U - T_C - R_U R_{LR} R_U^T T_U \end{aligned} \tag{6}$$

where K is the intrinsic parameter matrix of the PTZ camera, p_I is the projection of the 3D point in the image plane, $P'_{L0} = \begin{pmatrix} 0 & 0 & 1 \end{pmatrix}^{\mathrm{T}}$ is the unit vector on the Z axis in the $O_L^{'} X_L^{'} Y_L^{'} Z_L^{'}$, t_c is the scale factor of the camera ray from the PTZ and the P; t_l is the scale factor of the laser ray from the mirrored LDV to the P. The parameter c is the minimal distance between a point on the ray $O_L^{'} P_L^{'}$ and a point on the ray $O_C P_C$, which ensures that the distance can still be measured even though the two rays do not intersect. If the system has been calibrated, i.e., if K, R_{C0}, R_{U0}, T_U and T_C are known, and the pan and tilt angles of both the PTZ and PTU are given, there are only three unknowns t_c, t_l and c in (6), which can give exactly three linear equations to obtain values of the three unknowns. Then the 3D coordinates of the point P in LDV coordinate system is obtained by

$$P_L = t_c R_C K^{-1} p_I + T_C + \frac{c}{2}\left(R_C K^{-1} p_I \times R_U R_{LR} R_U^T P'_{L0}\right) \tag{7}$$

The distance will be use to adjust the focus parameters of the LDV.

B. Calibration

There are total 28 unknowns within the five matrices K, R_{C0}, R_{U0}, T_U and T_C that characterize the sensor geometry. The intrinsic parameters of the PTZ camera (in K) are first estimated using a well-known calibration technique [12]. Then the extrinsic parameters are estimated by combining (2) and (5) to eliminate P_L, as

$$R_C P_C = R_k \left(P'_L - T_U\right) + \left(T_U - T_C\right) \tag{8}$$

where $R_k = R_{U0} R_{U\alpha} R_{U\beta} R_{LR} {R_{U\beta}}^T {R_{U\alpha}}^T {R_{U0}}^T$. However, (8) is non-linear and very complicated, particularly due to the rotation matrix R_k. To simplify the calibration, we further assume the initial rotation matrix R_{U0} of the PTU is an identity matrix, which can be satisfied by adjusting of the mirror in initial setup procedure. With R_{U0} = I, R_k is known since the pan and tilt angles of both the PTZ and PTU are

given. In addition, because the variables $P'_L - T_U$ and $T_U - T_C$ are not independent, we pre-measure the distance between the fore lens of the LDV and the laser point on the mirror. Initially, the distance is set as a constant and then will be refined later. Consequently, the number of the unknown parameters in (8) is reduced to 14. Thus, given n 3D points, we can build 3n linear equations that include n+14 unknown (n for the Z coordinates of the n laser points in the mirrored LDV coordinate system; note their X and Y coordinates are zeros). In other words, to solve the problem, at least 7 points are needed. However, it is still hard to estimate all the n+14 unknowns simultaneously due to the linear system's sensitivity to noise. Therefore, we estimate the extrinsic parameters in four sub-steps. First, we adjust the PTZ camera so that the matrix R_{C0} can be initialized as an identity matrix, and both T_{Cy} and T_{Uy} are zeroes. By doing that we only need to solve (8) to find the values of T_{Cx}, T_{Cz} and T_{Ux}. Second, with the initial values of these three parameters, and till assuming the matrix R_{C0} is an identity matrix, we solve T_{Cy} and T_{Uy}. Third, we refine R_{C0} after we obtain the two translation vectors T_C and T_U .Finally, we can further refine the distance between the LDV to the PTU using to the relation between the distances and focus steps of the LDV (which is discussed in Section 4).

IV. Automatic Focusing of the LDV

The built-in automatic focus function of the Polytec LDV in our lab uses a passive focusing method. When the LDV accepts the automatic focusing command to focus to a reflective target at a distance, it tries all the focus steps of the full range (0-3300) and obtains the corresponding signal levels. Then it returns to the step position with the maximal signal level. However, since the range of all steps is large (from 0 to 3300), analyzing signal levels for such a large step range both takes long time (15 second) and may also have the problem of multiple peaks. Therefore we have designed an intelligent automatic focusing algorithm based on both surface distances and signal levels. The work includes two parts: calibrating the relation between the focus step parameters versus distance, and automatic fast focus with feedback of both the distance and signal level information.

A. Focus-step and distance relation

According to Gaussian lens equation, if the distances from the object to the lens and from the lens to the image are S_1 and S_2 respectively, they are related by

$$\frac{1}{S_1}+\frac{1}{S_2}=\frac{1}{f} \tag{9}$$

where f is the effective focal length of the lens of the LDV. The image distance S_2 is constant. When the object distance S_1 changes, the focal length of the LDV has to change for obtaining a focused image of the laser spot. The relation of the focal length and the object distance can be calculated by (9). Due to the lack of the intrinsic parameters of the LDV, we calibrate the relation experimentally. We measure the distances between fore lens of the LDV and the reflection surface, meanwhile acquiring the focus step values by using the built-in automatic focus function of the LDV to achieve laser beam focusing. We have found the relation between distances and focus steps for the Polytec OFV-505 LDV through this experiment method. Fig. 3 shows the relation curve.

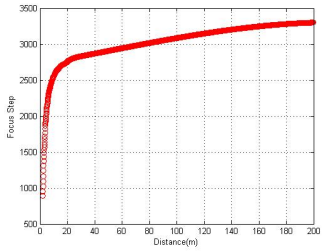


Fig. 3. Focus-step and distance

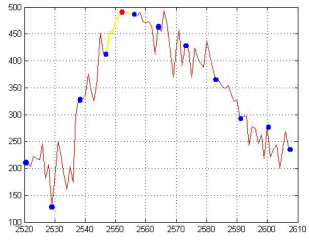


Fig. 4. Fast automatic LDV focusing

B. Automatic focusing algorithm

In theory, given a distance of the reflective surface, we will have a corresponding LDV focus step (within the range from 0 to 3300 for the LDV we used). However, there are two problems. First the distance measurements may not be accurate. Second, the step-distance correspondence is not fine enough for accurate focusing. Therefore, we design an automatic multi-scale focusing algorithm based on the measurement distance and signal return level. First, the distance D of the laser spot is measured using the triangulation between the LDV laser beam and the PTZ camera, using (7), and the corresponding LDV focus step S is found via the interpolation of the focus-step and distance ($S-D$) relation. Second, we set a small search range of the focus steps $[S-s, S+s]$. In our experiments, the range is set according to both ($S-D$) relation and depth measurement accuracy. Third, in the small search range, n discrete points are defined (the blue points in Fig.4) to control the LDV and to read n signal levels. By analyzing these signal levels, the platform obtains an even smaller search range (the yellow line in Fig. 4) which should include the maximum signal level. Fourth, the focusing position of the LDV is tuned to every focus step in this smaller range and each signal level is obtained. Finally by analyzing the curve of the signal levels, the best focus step is selected to focus the LDV (the red point in Fig. 4).

V. Surface Selection and Laser Pointing

The performance of the acquired LDV signals is mainly affected by the vibration and the reflection property of the surface. As we have noted, most of surfaces vibrate with

voice; however, it is hard to properly select a good surface in large distance. The use of the PTZ with image segmentation algorithm can help to determine several possible surface candidates. Then our techniques can be applied to point the LDV laser beam and focus on the target quickly.

A. Surface selection

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 due to the coherence requirement. A stationary, highly reflective surface usually reflects the laser beam of the LDV very well.

Once a human target is found, the background image is segmented into several regions which include the human object and several possible surfaces. Only those surface regions close to the human region are selected. Given n set of those regions with centers: $C_1, C_2, \ldots, C_n$, a series of signal levels $S_1, S_2 \ldots, S_n$ are obtained from the LDV when the laser beam point to those regions. Comparing the values of these signal levels, the kth region which has the maximum signal level S_k is selected. Then the LDV points to the center C_k of the kth region to capture the best audio signals among all of these regions (please refer to Fig. 6 for an example).

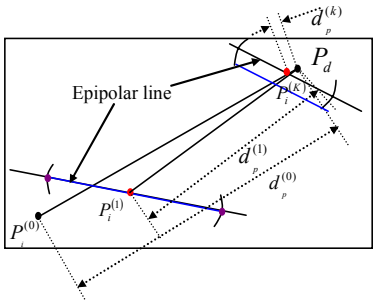


Fig. 5. Incremental LDV laser pointing

B. Laser pointing algorithm

Upon the selection of a surface, a destination point P_d on the surface is chosen for re-pointing the laser ray onto it. Since the distance of the point to the LDV is unknown before the laser is pointed to it, we cannot calculate the exact pan and tilt angles of the PTU system in order to turn the laser ray to that point. However, we could give a rough estimation of the angles based on the pixel displacement $d_p^{(0)}$ in the PTZ image between the current laser point $P_i^{(0)}$ and the destination point P_d. The pan and tilt angles viewed from the PTZ camera's viewpoint can be accurately calculated based on this displacement. Since the laser beam is reflected from the mirror mounted on the PTU at a different location, the pan angle α and the tilt angle β of the PTU are roughly set as halves of the pan and tilt angles as seen from the PTZ camera. However turning the PTU by these angles will not enable an accurately aim of the laser beam to the destination point due to the viewpoint changes (from the PTZ to the PTU), off-center rotations of the mirror, and the difference in the distances from LDV to the two points (P_i and P_d). Therefore, we design an incremental laser pointing algorithm. Fig. 5 illustrates the incremental laser pointing and searching strategy. At every laser pointing step, k = 1, 2, …, the total pan and tilt angles are roughly estimated based on the previous pixel displacement $d_p^{(k-1)}$, but the PTU is only rotated by a small pan angle α/m and a small tilt angle β/m, to an intermediate point $P_i^{(k)}$ (the original point is denoted as $P_i^{(0)}$. The value of m is set based on the range of the scene. The current laser point $P_i^{(k)}$ is searched in the PTZ image on an epipolar line given by the known rotation angles of the PTU, whereas the search range along the epipolar line is calculated based on the distance of the previous laser point $P_i^{(k-1)}$) to the LDV that can be estimated using (7), and the range of the scene. After finding the current laser point $P_i^{(k)}$, a new pixel displacement $d_p^{(k)}$ between the current laser point $P_i^{(k)}$ and the destination point P_d is calculated. If the pixel displacement $d_p^{(k)}$ is smaller than a threshold, then the laser pointing procedure is completed. Then the system focuses the LDV laser beam and then turns the PTZ to center the destination point P_d. Otherwise, a new set of pan and tilt angles of the PTU is estimated based on the current pixel displacement $d_p^{(k)}$, and another incremental laser pointing step is performed, until the threshold is reached at step K (Fig. 5). Note that the value of m keeps unchanged, so the rotation angles of the PTU will be smaller and smaller from step to step.

VI. Experimental Results

We have carried out a few experiments on various aspects of the vision aided automated vibrometry.The first experiment is to verify our calibration result by measuring the depth accuracy. The second experiment demonstrates our fast automatic focusing technique. The third one shows some surface selection results. A video demo can be found at [13] showing the main functions of our vision-aided intelligent LDV voice acquisition system.

A. Distance measurement

In our multimodal system, the angle resolution of the PTU is 0.0129° and the baseline between the PTZ and mirrored LDV laser projection equals to 1000mm. The focal length of the PTZ changes from 938 pixels to 27450 pixels, and the rotation error of the PTU is 0.013 degrees. Assuming the image error is 1 pixel, and the maximum zoom of the PTZ is used, the depth errors at various distances are shown in Table 1. The error table will be used for automatic focusing.

Table 1. Theoretical estimates of depth accuracy

Depth (m)	5	10	30	50	100	200
Absolute error (m)	0.01	0.03	0.24	0.65	2.62	10.46
Relative error (%)	0.13	0.26	0.78	1.21	2.62	5.23

B. Automatic focusing

In this experiment we verify our automatic focusing technique by placing a reflecting surface (with a

retro-reflective tape) at different distances, and compare the results with those with the built-in focus function of the LDV.

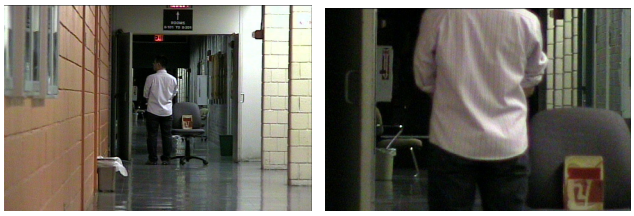
The comparison results are shown in Table 2. The search range of our intelligent focusing is 165 times smaller than the built-in focusing of the LDV. Ideally, if we could embed our intelligent focus function into the controller of the LDV, the focusing speed will improve 165 times. Even when the current implementation spends quite some time in reading and writing separate commands to the serial port of the LDV in the current configuration, our intelligent focusing only takes about 1.0 seconds, which is about 15 times faster than the built-in focus function of the LDV. Experiments show that the returning signal levels and focus positions using our technique are the same as or even better than the built-in focus function of the LDV.

Table 2. Comparison of built-in focus and our intelligent focus methods

Depth (feet)	Built-in Focus Time: 15s Range: 3300		Our intelligent focus			
	Signal level	Focus position	Time(s)	Signal level	Focus position	Search range
30	512	2581	1.0	512	2578	20
60	512	2769	1.0	512	2768	19
90	512	2837	1.0	512	2838	18
120	512	2886	1.0	512	2889	18
180	489	2912	1.0	492	2913	18
240	463	2931	1.0	461	2931	18
360	410	2939	1.0	407	2938	18
420	377	2957	1.0	381	2958	18

C. *Surface Selection*

We have performed an experiment in the corridor outside our lab to verify our laser pointing method. The retro-reflective tape is put on a surface 300 feet away (Fig. 6a). The original image was zoomed using 26x optical zoom of the PTZ in order to obtain a clear and large human image (Fig. 6b). Then the image was segmented into several homogenous color regions in order to find the best reflection surface, which is close to the human object (Fig. 6c). The yellow points are the centers of those regions. One of the center points with the strong signal is selected for pointing the LDV laser beam (Fig 6d). We have done a lot of experiments for surface selection with different backgrounds, targets and environments. Here we only demonstrate one. In almost all experiments we can select a good or optimal surface that can be used for long-range hearing.



(a) Initial image (b) Zoom-in image

(c) Segmentation & selection results (d) laser ray to selected surface

Fig. 6. Experiment results of surface selection

VII. CONCLUSION

In this paper, we have proposed an active multimodal sensing platform to improve the performance and the efficiency of the LDV for long-range hearing. The platform integrated a PTZ camera, a mirror and a PTU to the LDV. The PTZ camera not only could assistant the LDV to point the laser beam to a better reflective surface and obtain optimal audio signal, but also could measure distance of the target by using of triangulation of the PTZ camera, the mirror and the LDV. Based on the measured distance, we designed a fast automatic focus technique which is 15 times faster than the built-in automatic focus function of the LDV. This intelligent multimodal sensing platform can be used in various applications, such as remote and large area surveillance, perimeter protection in important place, and rescue in disaster.

REFERENCES

[1] C. Zieger, A. Brutti and P. Svaizer, "Acoustic based surveillance system for intrusion detection," IEEE ICVSBS'09: 314-319.

[2] C. Clavel, T. Ehrette and G. Richard, "Events detection for an audio-based surveillance system," IEEE ICME'05 :1306-1309.

[3] R. Radhakrishnan, A. Divakaran and A. Smaragdis, "Audio analysis for surveillance applications," IEEE WASPAA'05:158-161.

[4] L. Antonelli and F. Blackmon, "Experimental demonstration of remote, passive acousto-optic sensing," J. Acoust. Soc. Am., 116(6): 3393-403, Dec. 2004.

[5] M. Cristani, M. Bicego and V. Murino, "Audio-visual event recognition in surveillance video sequences," IEEE Trans. Multimedia, 9(2): 257-267, Feb. 2007.

[6] Y. Dedeoglu, B. U. Toreyin, U. Gudukbay and A. E. Cetin, "Surveillance using both video and audio," in Multimodal Processing and Interaction: Audio, Video, Text, P. Maragos, A. Potamianos and P. Gros Eds., New York: Sringer, 2008: 143-156.

[7] E. S. Neo, T. Sakaguchi, and K. Yokoi, "A natural language instruction system for humanoid robots integrating situated speech recognition, visual recognition and on-line whole-body motion generation", IEEE/ASME AIM'2008:1176-1182.

[8] F. A. Blackmon and L. T. Antonelli, "Experimental detection and reception performance for uplink underwater acoustic communication using a remote, in-air, acousto-optic sensor," IEEE J. Oceanic Egnineering, 31(1), Jan. 2006: 179-187.

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

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

[11] E. Trucco and A. Verri, Introductory Techniques for 3-D Computer Vision. Upper Saddle River, NJ:Prentice Hall, 1998.

[12] J. Y. Bouguet, Camera calibration toolbox for Matlab. CalTech, http://www.vision.caltech.edu/bouguetj/calib_doc/index.html, June, 2008.

[13] Y. Qu, T. Wang and Z. Zhu, Vision aided laser Doppler vibrometer (demo), http://visionlab.engr.ccny.cuny.edu/~qu/VaLDV-demo.wmv, January, 2010.