Vision-aided Laser Doppler Vibrometry for Remote Automatic Voice Detection

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

***Abstract*—For automating the remote voice detection using a Laser Doppler Vibrometer (LDV), we integrate a Pan-Tilt-Zoom (PTZ) camera, a mirror and a Pan-Tilt-Unit (PTU) with the LDV to form a multimodal sensing system. With the assistance of 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 and related image processing techniques in surface selection and laser pointing are designed. Experimental results are shown to validate the performance improvement of the LDV in remote automatic voice detection by using the multimodal system.**

***Index Terms*—Automatic voice detection, laser Doppler vibrometer, multimodal sensing, stereo vision**

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 targets move out of their sensing ranges, they will not be able to obtain any signals. 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 used in the inspection industry and other important applications concerning environment, safety and preparedness that meet basic human needs. Laser Doppler vibrometers such as those manufactured by Polytec [8] and B&K Ometron [9] can effectively detect vibration within two hundred meters with sensitivity on the order of 1μm/s. Whenever the LDV can obtain a return signal from a surface that is vibrated due to structural, environmental and other causes, it can detect the vibration signals. In bridge and building inspection, the non-contact vibration measurements for monitoring structural defects eliminate the need to install sensors as a part of the infrastructure [e.g.,10]. In security and perimeter applications, an LDV can be used for voice detection without having the intruders in the line of the sight [11], [12]. In medical applications, an LDV can be used for non-contact pulse and respiration measurements [13]. In search and rescue scenarios where reaching humans can be very dangerous, an LDV can be applied to detect survivors which are even out of visual sight. Blackmon and Antonelli [14] 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.

However, in most of the current applications, such systems are manually operated. In close-range and lab environments this is not a very serious problem. But in field applications, such as bridge/building inspection, area protection or search and rescue applications, the manual process takes a very long time to find an appropriate reflective surface, focus the laser beam and get a vibration signal; more so if the surface is at a distance of 100 meters or more. For example, using a COTS Polytec LDV (OFV-505), it takes about 15 seconds for the laser focusing on a surface.

In this paper, we integrate the LDV with a Pan-Tilt-Zoom (PTZ) camera, and a mirror on a Pan-Tilt-Unit (PTU). The

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

Y. Qu is with School of Instrumentation Science & Opto-electronics Engineering at Beijing University of Aeronautics and Astronautics (BUAA). This work was performed when he was a Postdoctoral Follow at the City College of New York.(email: qyf@buaa.edu.cn).

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

integrated multimodal system can automatically detect reflective surfaces and aim the LDV laser beam with the aid of 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 both the surface distance and the LDV signal level. The system can be viewed as a visual servo system to enhance the functionality of another measurement device (namely LDV), which adds a novel use to the spectrum of vision systems for Mechatronics applications such as multi robot localization [15], [16], [17], human-robot interaction [18], [7], to name a few. The vision add-ons also provide other sensing modalities (visual and range) that are usually important and necessary in inspection and surveillance applications.

This paper has three main contributions. First, an automation and real-time performance of the LDV measurements is achieved using computer vision techniques and multimodal signal feedback. Second, a novel stereo vision configuration is constructed with heterogeneous sensors that capture video and audio signals respectively, and a calibration method of the multimodal sensing system is proposed. Finally, a new multimodal sensing platform is designed that could obtain audio, video and 3D information using cost-effective add-ons, thus greatly expand the functionalities of the laser vibrometer for long-range sensing.

The rest of the paper is organized as follows. Section 2 introduces the basic principles, discusses issues of remote voice detection using an LDV sensor, and then gives an overview of our solution to these problems. In Section 3, and the methods for depth measurement and sensor calibration are proposed. In Section 4, the automatic LDV focusing algorithm is described. Section 5 discusses the automatic reflection surface selection and the laser beam pointing strategies. Section 6 presents some experimental results. Finally, we conclude our work and discus future work in Section 7.

II. Vision Aided Long Range Voice Detection

A. A novel sensor and the unmet needs

A 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 (Fig. 1), a coherent laser beam is divided into object and reference beams by 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. A detector converts this signal to a voltage fluctuation. The velocity v of the moving object is directly obtained by a digital quadrate demodulation method [19].

Most objects vibrate when wave energy (including voice waves) is applied on them. Though the magnitudes of the vibration caused by voice waves are very small (usually in nanometer level), this vibration can be detected by the LDV. The relation of voice frequency f, velocity v and magnitude m of the vibration is

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

Note that the velocity v is directly proportional to frequency f. 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.

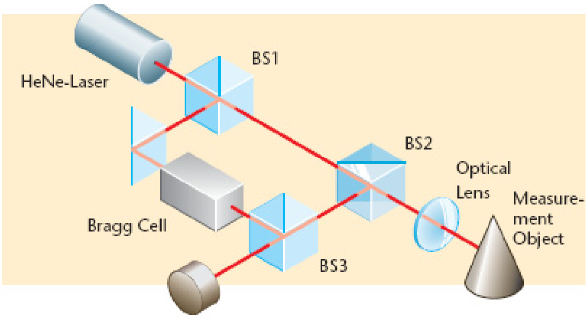


Fig. 1. The modules of the Laser Doppler Vibrometer (LDV)

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 the reflected 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 be the number of scattering centers and the angular dependence of the path length differences in a given direction. Since 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 target(s) 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.

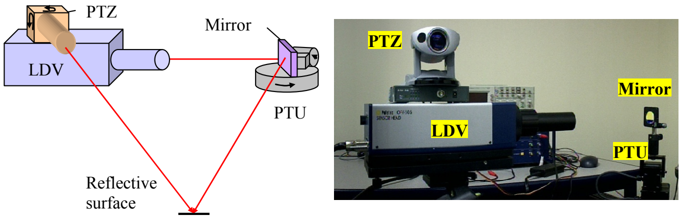


Fig. 2. System setup of the vision-aided LDV. Left: an illustration of the system configuration; Right: a real experimental setup.

B. Our solution

To solve the above two problems in using the LDV for voice detection, we have designed a multimodal system that integrates the LDV with a PTZ camera and a small mirror mounted on a PTU (Fig. 2). 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 work flow of the multimodal system includes the following six steps.

Step 1.*Target detection*. The system detects moving objects via PTZ video analysis, and then candidates of possible human or vehicle targets are generated.

Step 2. *Surface selection.* For each moving target or a set of targets, surrounding surfaces are selected for pointing the LDV laser beam to measure the vibration caused by the sounds of the target.

Step 3. *Laser pointing.* The platform controls the PTU to turn the LDV laser beam to point to each selected surface.

Step 4. *Distance measurement.* The distance between the LDV and the surface is calculated by using the triangulation between the LDV laser beam and the PTZ camera.

Step 5. *LDV automatic focus.* Based on the measured distance and the return signal levels of the LDV, the LDV laser beam is automatically focused.

Step 6. *Voice detection.* The LDV captures vibration signals to the host computer to decode the signals into acoustic signals for further enhancement and analysis.

Since the LDV voice measurements require the reflective surfaces to be stationary, we do not direct point the laser to the moving targets. Object tracking should be performed by the PTZ camera. Whenever a moving target is detected, the LDV will point to a background surface close to the moving target and try to get some sound signals (if any). The LDV can "hear" the sound when the moving target is within a certain distance to the background surface, usually more sensitive than a microphone as if it was setup on the surface. When the moving target is out of the range, a new surface should be selected. The purpose of the automated LDV pointing is to facilitate this adaptation.

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

III. Distance Measurement and Calibration

There are four hardware components in our vision-aided LDV platform: LDV, PTZ, mirror, and PTU. In order to enable the system to measure various ranges of surfaces with sufficient accuracy, we need to geometrically calibrate the platform. The following issues of the system make 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 intersects with the mirror. 4) The focal lengths of the PTZ camera change with its zooms. In the following, we first build the geometric model of the system to measure the distances, and then present our method in calibrating the platform.

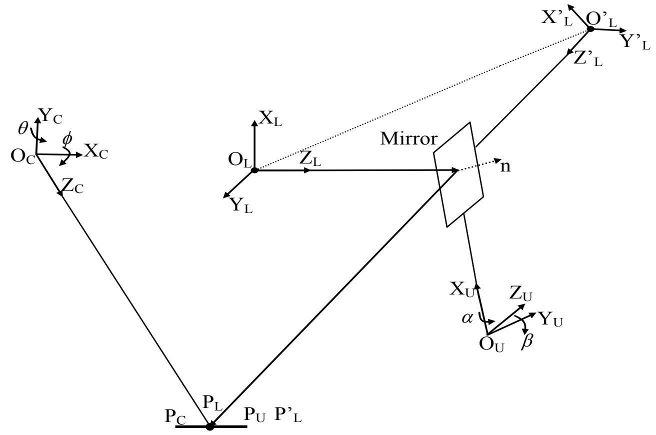


Fig. 3. Coordinate systems of the vision-aided LDV platform

A. Geometric model

The vision-aided LDV 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. 3).

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

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

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

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

where R_U is the rotation matrix and T_U is the translation vector of the PTU in $O_L X_L Y_L Z_L$, with $T_U = \left(T_{Ux}\ T_{Uy}\ T_{Uz}\right)^T$. Note that each of R_C and R_U is the multiplication of 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:

$$P_L^{'} = R_U R_{LR} P_U + T_U \qquad (4)$$

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

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

Using the well-known triangulation method [20] between the PTZ ray and the mirrored laser ray, we can obtain:

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

where K is the matrix containing the intrinsic parameters of the PTZ camera, p_I is the projection of the 3D point in the image plane, $P'_{L0} = \left(0 \quad 0 \quad 1\right)^T$ is the unit vector on the Z axis, representing the reflected laser ray in the mirrored LDV system $O_L^{'} X_L^{'} Y_L^{'} Z_L^{'}$, t_c is the scale factor of the camera ray from the PTZ and the point P; t_l is the scale factor of the laser ray from the mirrored LDV to the point 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 Eq. (6), which can give exactly three linear equations to obtain values of the three unknowns. Then the 3D coordinates of the point P_L in LDV coordinate system is obtained by

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

Once the coordinates of P_L is calculated, the distance between the point and the LDV laser source (along the laser rays) can be estimated and used for laser beam focusing. The distance is the Z'_L coordinate of the point represented in the mirrored LDV system.

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's geometry. The intrinsic parameters of the PTZ camera (in *K*) are first estimated using a well-known calibration technique [21]. Then the extrinsic parameters are estimated by combining Eq. (2) and Eq. (5) to eliminate P_L, as

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

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, Eq. (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 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 will be refined later. Consequently, the number of the unknown parameters in Eq. (8) is reduced to 14. Thus, given n 3D points, we can build *3n* linear equations that include *n+14* unknown (14 for the extrinsic sensor parameters, and *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 not a robust approach to estimate all the n+14 unknowns simultaneously due to the sensitivity of the linear system to the 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. Second, we find the values of T_{Cx}, T_{Cz} and T_{Ux} by solving Eq. (8). Third, we solve T_{Cy} and T_{Uy} using the calculated values of T_{Cx}, T_{Cz} and T_{Ux}. Fourth, 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 the relation between the distances and focus steps of the LDV (which will be discussed in Section IV).

C. Depth error

The depth error of the measurement comes from two sources after the calibration. The first source of error is the difference of the estimated laser position in the 2D image and the true 3D laser location projected onto the 2D image, denoted as $\partial\gamma_1$. The second error is the angular error in turning the PTU, denoted as $\partial\gamma_2$. According to principle of triangulation, the absolute error of depth measurement ∂D is given by

$$\partial D = \frac{D^2}{B}\partial\gamma \tag{9}$$

where D is distance of the target point to fore lens of the LDV (i.e., the Z coordinate of the point P'_L), B is baseline (the distance between O_C and O'_L), $\partial\gamma$ is the sum of $\partial\gamma_1$ and $\partial\gamma_2$. The first error term $\partial\gamma_1$ can be calculated as

$$\partial\gamma_1 = \frac{\Delta x}{f} \tag{10}$$

where Δx is the pixel difference between the true laser beam position and its estimated position in the image plane, and f is the focal length in pixel. The distance error estimation will be used in automatic laser focusing based on distance measures.

IV. Automatic Focusing of the LDV

The built-in automatic focusing 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 range of steps takes a long time (about 15 seconds) and may also have the problem of multiple peaks. Therefore, we design an intelligent automatic focusing algorithm based on both target distances and signal levels. This work includes two parts: calibrating using the relation between the focus steps and target distances, and automatic fast focusing with feedback from both the distance and signal level information.

A. Focus-step and distance relation

According to Gaussian lens equation, the relationship between the distance from the target (i.e. the reflective surface) to the lens D, and the distance from the lens to the image d is defined as

$$\frac{1}{D} + \frac{1}{d} = \frac{1}{f_L} \tag{11}$$

where f_L is the effective focal length of the lens of the LDV. A super-long range lens OFV-SLR (f_L = 200 mm) is used in the LDV, and the possible stand-off distance D of the target is from 1.8 meters to over 300 meters. While the focal length f_L is constant when the target distance D changes, the image distance of the LDV has to change for obtaining a focused image of the laser point. In the LDV, this is achieved by changing the focus step S (from 0 to 3300 digital steps).The relation of the image distance and the target distance can be calculated by Eq. (11). Due to the lack of the intrinsic parameters of the LDV, particularly the relation between the image distance and the focus steps (0 – 3300), we calibrate the relation experimentally. We measure the distances between fore lens of the LDV and reflective surfaces (targets) at various distances (from 2.13 to 200 meters), meanwhile acquiring the focus step values (from 893 to 2962) when using the built-in automatic focus function of the LDV to achieve laser beam focusing. Those corresponding values between the distances and the focus steps for the Polytec OFV-505 LDV are shown as red circles in Fig. 4. Similar to a zoom camera, we assumed the monotone increasing non-linear relation between the image distance and the focus step is as follows:

$$d = e^{-a\cdot S+b} \tag{12}$$

where a and b are two constant values that are determined via curve fitting. Substituting Eq. (12) into Eq. (11), we obtain the relation between the focus step and the target distance:

$$S = \frac{1}{a}[\ln(\frac{1}{f_L} - \frac{1}{D}) + b] \tag{13}$$

where we have found $a = 4.68\times10^{-5}$, $b = -1.4697$ when we set the focal length $f_L = 0.2$ meters for the Polytec OFV-505 LDV that we used. The fitted curve is shown in black line on top of the measured data in red circles in Fig. 4. Given the data we have measured, the fitted curve applies from 1.54 meters to 300 meters for the distance, and from 0 to about 3000 for the steps. This is consistent with the manufacture's specification.

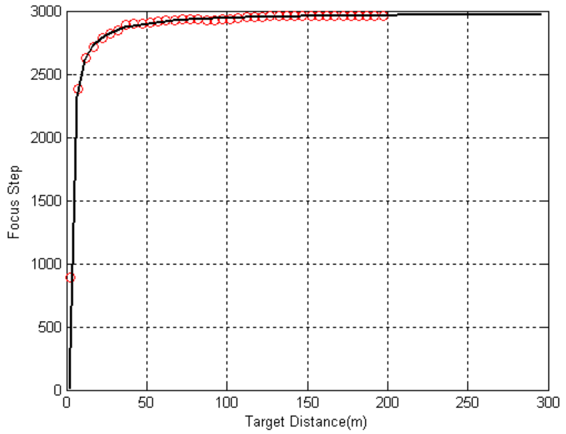


Fig. 4. Focus-step and distance relation (The fitted curve is shown in black line on top of the measured data in red circles)

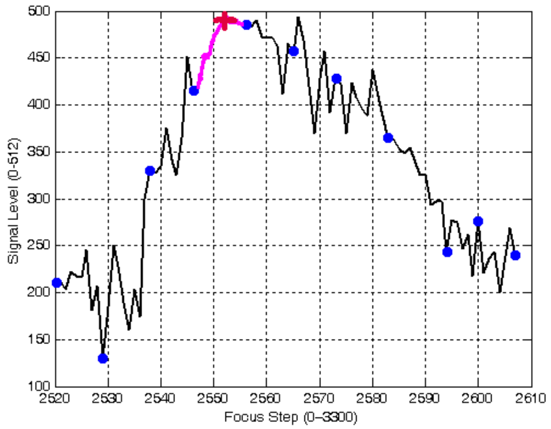


Fig. 5. Fast automatic LDV focusing (Horizontal axis: focus steps between 0 – 3300; vertical axis: signal levels between 0-512)

B. Multi-scale automatic focusing algorithm

In theory, giving a distance of the reflective surface, the corresponding LDV focus step (within the range from 0 to 3300 for the LDV we used) can be calculated. However, there are two problems. First, the distance measurement 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 measured distance and signal return level.

First, the distance D of the laser point is measured using the triangulation between the LDV laser beam and the PTZ camera, using Eq. (7). Then a focus step S is calculated by using Eq. (13) given the distance D.

Second, an offset s is set for defining a searching range $\left[S-s, S+s\right]$ for the best focus step. By calculating the derivative of the Eq. (13) wrt D, and also using the distance error in Eq. (9), we can obtain the offset s of the automatic focusing as a function of the distance D, as

$$s = \partial S = \frac{1}{aB\left(\frac{1}{f_L} - \frac{1}{D}\right)}\partial\gamma \tag{14}$$

Third, in the searching range, n discrete points are defined (the solid dots in Fig. 5) to control the LDV and n corresponding signal levels are recorded (this is the coarse scale focus-level search).

Fourth, through analyzing these signal levels, a smaller search range (the thick segment of the curve in Fig. 5, about 10-20 steps) is obtained. Then every focus step in this smaller range is searched and the corresponding signal level is obtained. This is the fine scale focus-level search. Lastly, the focus step that returns the best signal level is selected to focus the LDV (the cross mark in Fig. 5).

V. Surface Selection and Laser Pointing

The performance of the acquired LDV signals is mainly affected by the vibration and the reflection properties of the surface. As we have noted, most of surfaces vibrate with voice; however, it is hard to properly select a good surface in a large distance. The use of the PTZ camera can help to determine several possible surface candidates. The selection of the reflective surface is based on its distance to the moving target and its color. Right now the selection is a heuristic process so it might not to be optimal. We use a “hypothesis and test” procedure by first picking up a few smooth surfaces close to the target and choose the one that provides the best signal returns. This will be useful for both automatic monitoring and interactive measuring since the vision add-ons speed up the surface selection.

A. Automatic 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.

One application of our system is to have the user to pick up the surfaces by looking at the images of the PTZ camera. Here we discuss a simple automatic method for surface selection based on surface smoothness. Once a human target is found, the background image is segmented using a well-known mean-shift color segmentation method [22]. Usually the large regions are the smooth, flat regions that could be good candidates for laser reflection. In addition to their sizes, we also select regions that have strong red components (e.g. red, white or pink colors instead of green) so the laser reflection would be good. 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 is pointed to the center C_k of the kth region to capture the best audio signals among all of these regions.

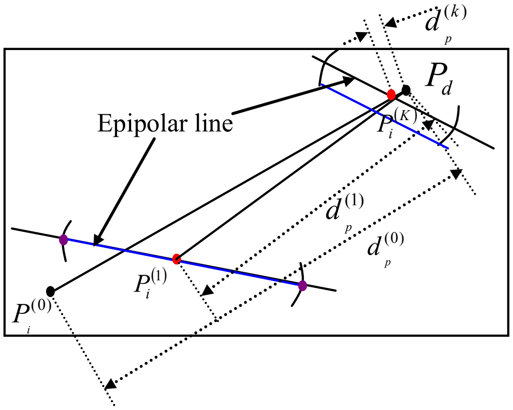


Fig. 6. Incremental LDV laser pointing

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 $P_i^{(0)}$ (which has distance measurement) 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 achieve an accurate aim of the laser beam to the destination point due to (1) the viewpoint changes (from the PTZ to the PTU); (2) off-center rotations of the mirror; and (3) 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. 6 illustrates the incremental laser pointing and detection strategy. Here, we first describe the laser pointing algorithm; the laser point detection method will be discussed in the following sub-section. At every laser pointing step, k = 1, 2, …N, 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 change of the scene. For example, for a scene of a possible 10-meter maximum range change, we would set m to 20 so that each rotation of the PTZ will roughly go over 0.5 meters in distance. 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 searching 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 Eq. (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 (within a few pixels, e.g. 2), the laser pointing procedure is complete. Then the system focuses the LDV laser beam and turns the PTZ to the center of 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. 6). Note that the value of m is unchanged, so the rotation angles of the PTU will be smaller and smaller from step to step. This is a typical visual servo problem, and the algorithm works when the depth changes are smooth; but we cannot prove the convergence in a general case when abrupt changes in depths occur.

C. Laser point detection method

A single laser point from the LDV has various speckle patterns when captured in images (Fig. 7). It is important to find the precision center position (in sub-pixel) of the laser point in the image in order to measure the distance of surface or move the laser point to the correct location. A complex speckle pattern may contain several bright regions spreading around the center region. Flood fill and morphological techniques [23] are used to obtain a solid laser speckle region. First, a seed point is selected based on the epipolar constraint (in automatic searching; see Section V.B) or by a manual click (in interactive mode). Second, the flood fill method is applied to find the region of the laser spot and obtain an initial centroid. Third, we start from this central region to find, in its surroundings, all laser speckle regions which might not be connected to it. The histogram that counts number of laser-like (red) pixels is constructed to determine the threshold for the binary result of the laser region(s). Fourth, the erosion and dilation operations are performed to obtain a solid laser point region, and then a second centroid of laser point (based on this extended solid region) is calculated. Finally, the initial and second centroids are compared. If the distance between the two centroids is less than a threshold (usually 1 pixel in our experiments), the laser position is calculated by averaging the two centroids. Otherwise, the laser point is reselected by repeating the above steps.

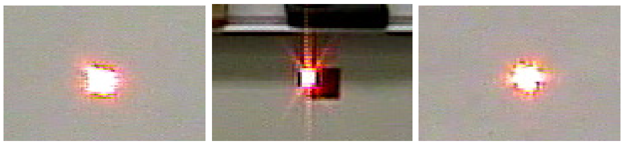


Fig. 7. Laser point displayed in image

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 results by measuring the depth accuracy. The second experiment demonstrates our fast automatic focusing technique. The third one is to test position repeatability of the laser point searching. The fourth experiment shows some laser pointing results. A video demo can be found at [24] 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 0.013°, the baseline between the PTZ and the mirrored LDV laser projection is about 1000mm. The focal length of the PTZ changes from 938 pixels to 27450 pixels. The purpose of using a zoom camera is to obtain the appropriate FOVs and image resolutions for various distances. We have calibrated the PTZ camera at various zoom factors, but in the paper we only showed some results. Assuming the image error is 1 pixel, and the maximum zoom of the PTZ is used, the equivalent angular error is $3.643° \times 10^{-5}$ according to Eq. (10). Under these conditions, the depth errors at various distances calculated according to Eq. (9) are shown in Table 1.

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

To verify the accuracy of the measuring depth, we test the vision-aided LDV platform on various objects. Here we only show one example using a planar object (a whiteboard). The captured points (blue circles) in the image and the re-projected points (white crosses) from 3D to 2D of the object are shown in Fig. 8. The plane is placed about 6.2m to the LDV, the focal length of the PTZ is 6456 pixels. To verify the accuracy of 3D measurements, a plane is fitted on these 3D points, and the average distance of these points to the plane is calculated as the actual distance measurement error (9.41mm). The theoretical depth error is 8.72mm using Eq. (9) given the distance and the PTU and PTZ measurement resolutions. The actual error is very close to the theoretical one. The 3D reconstruction of the plane is shown in Fig. 9, where the red circles are the points with positive errors and the blue crosses are the points with negative errors. From the automatic focusing algorithm which sets a range for searching the maximum signal level, the distance measurement is sufficiently accurate for the automatic focusing of the LDV.

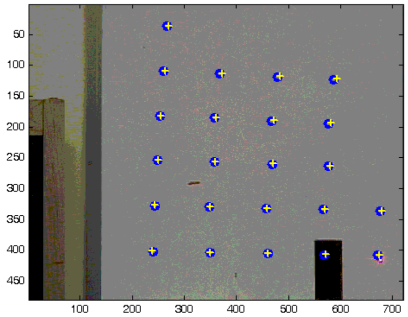


Fig. 8. 3D points selection and re-projection on a whiteboard

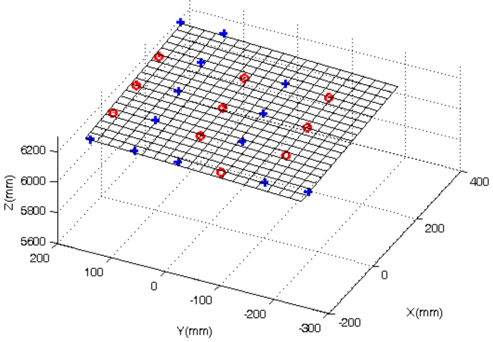
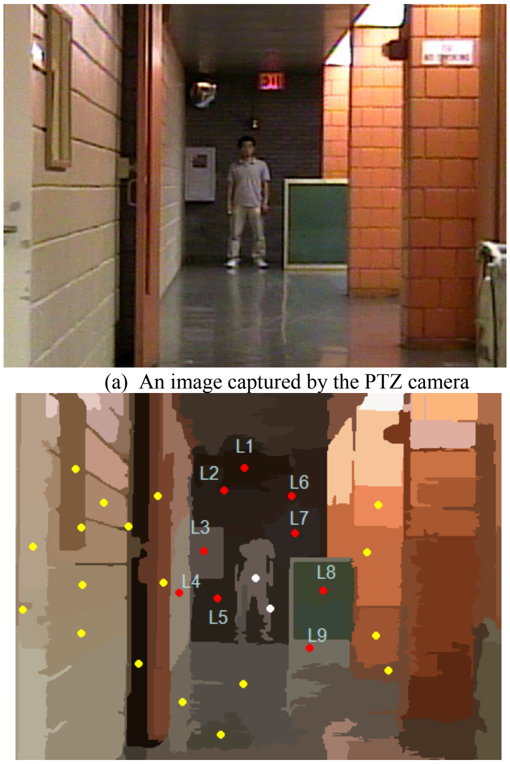


Fig. 9. 3D reconstruction of the plane



(a) An image captured by the PTZ camera

(b) Image segmentation and surface selection

Fig. 10. Experiment result of laser pointing from 30 meters

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 searching range of our intelligent focusing is 165 times smaller than the built-in focusing of the LDV. That is to say, if we embedded our intelligent focus function into the controller of the LDV, the focusing speed will improve 165 times faster (ideally). Even when the current platform spends quite some time to read and write 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 step positions using our technique are the same as, if not better than the built-in focus function of the LDV.

Table 2. Comparison in times and search ranges of built-in focus and our intelligent focus (Focus step/ range: 0 to 3300; Signal level: 0 to 512)

Distance (meters)	Built-in Focus Time: 15s		Our intelligent focus			
	Signal level	Focus step	Time(s)	Signal level	Focus step	Search range
10	512	2581	1.0	512	2578	20
20	512	2769	1.0	512	2768	19
30	512	2837	1.0	512	2838	18
40	512	2886	1.0	512	2889	18
60	489	2912	1.0	492	2913	18
80	463	2931	1.0	461	2931	18
120	410	2939	1.0	407	2938	18
140	377	2957	1.0	381	2958	18

C. Laser Point Searching

To test repeatability of the laser point searching method, we test three different image shapes of the laser point (Fig. 7). We randomly clicked a seed point multiple times close to the laser speckle to obtain the centroid of the laser point to observe the consistency of the localization. The experiment results are shown in Table 3. Experiments show that the standard deviation of the laser point searching is less than 0.5 pixels. The laser location accuracy satisfies the requirements for the laser pointing and distance measurement.

Table 3. Results of laser point searching (K =1 to 6 is the number of clicks, and points (x and y)and errors are measured in average pixel locations)

K / Point		1	2	3	4	5	6	error
Point 1	x	264.83	264.53	264.36	265.12	265.30	264.79	0.35
	y	135.70	135.47	135.26	135.59	135.98	135.44	0.25
Point 2	x	437.25	437.17	438.06	437.90	437.59	437.37	0.36
	y	364.91	364.54	364.32	364.37	364.13	364.08	0.30
Point 3	x	359.77	359.61	359.03	360.21	359.82	359.39	0.40
	y	216.58	216.15	216.26	216.77	216.99	217.27	0.41

D. Laser Pointing

We performed an experiment in the corridor outside our lab to verify our laser pointing method. The retro-reflective tape is put on a surface 30 meters away (Fig. 10 (a)). When a human target was detected in the environment, the system segments the image into multiple color regions (Fig. 10 (b)). Among the regions (with dots) selected, some are on the moving targets (dots in white), some are too far away (dots in yellow and not labeled), and some are close to the target (dots in red and labeled from L1 to L9). Then the system automatically pointed the LDV to those regions close to the target, measured their distances (in meters) to the LDV, focused the laser beam, and read the signal return levels (from 0 to 512). Table 4 shows the results of distance measurements, focus steps, and the signal return levels from those surfaces. Among those returns, the metal box gave the best signal return, and the audio signals collected were intelligible. We also compared the results of our system with the "ground truth" data: distances by manual measurement, LDV focus steps and signal return levels by using the built-in full-range auto focus function, and found that our system gave very accurate and robust results.

Table 4. Surface selection and laser pointing (Units: Range – meters; Step - from 0 to 3300; Signal Level – from 0 to 512)

Label	Surface	Measurements			Ground Truth		
		Distance	Step	Level	Distance	Step	Level
L1	Wall	30.26	2786	12	30.63	2845	14
L2	Mirror	28.56	2758	12	30.63	2757	12
L3	Metal box	30.26	2745	18	30.32	2839	21
L4	Side wall	-	-	-	23.16	2410	11
L5	Wall	31.43	2410	11	30.63	2757	11
L6	Wall	29.20	2734	11	30.63	2734	12
L7	Wall	28.25	2732	12	30.63	2734	12
L8	Chalkboard	29.88	2750	11	27.74	2764	12
L9	Woodside	26.14	2642	12	27.74	2581	12

VII. Conclusion and Discussions

In this paper, we have proposed a vision-aided LDV system to improve the performance and the efficiency of the LDV for automatic remote hearing. The platform integrated a PTZ camera, a mirror, and a PTU to the LDV. The PTZ camera not only could assist the LDV to point the laser beam to a better reflective surface and obtain optimal audio signals, but could also measure distance of the target by using triangulation of the PTZ camera, the mirror, and the LDV. Based on the measured distance, we have designed a fast automatic focusing technique which is 15 times faster than the built-in automatic focus function of the LDV. This automated acoustic sensing platform can be used in various applications, such as remote hearing and voice detection in large area surveillance, perimeter protection in high security area, and search-and-rescue in disaster.

The key contributions of this paper are not just a multimodal sensing system for acquiring complementary information (audio, video and 3D), but also the intelligent sensing integration that significantly increases the performance of individual sensors, and adds new functionalities. This is particularly important for long-range surveillance and inspection applications. The LDV provides long-range acoustic detection capability to the video sensor; whereas the vision techniques enable the real-time and automatic operation of the LDV to acquire target surfaces, measure their distances for automatic laser focusing, and track targets. In addition, 3D information obtained through the cooperation of the two heterogeneous sensors is also very important for target detection and identification.

However, the current design and system has some limitations. First, in system configuration, since the laser cannot be pointed to the mirror at the center of the rotation of the PTU, the calibration procedure is sophisticated. Therefore, the laser pointing for finding a new surface via the PTZ camera is not very straightforward due to the fact that there is not a common center of projection for the laser beams controlled by the PTU. This drawback will be eliminated in our next prototype by using a PTU that allows a mirror to be mounted in its rotation center. Second, there is the chicken or the egg problem in the laser focusing and laser point detection. In order to focus the laser quickly on a target surface using the distance measurement from triangulation, we will have to assume the laser spot can be seen and detected in the PTZ camera images, which is hard when the laser beam is not focused. In our current implementation, we use an iterative method that could fail when the laser beam sweeps across two surfaces with significant depth change. Finally, there is also a tradeoff between the 3D accuracy and visibility. In order to obtain sufficient and accurate distance measurement over a large distance (>100 meters), the baseline between the camera and the mirrored LDV system has to be large. However, this will increase the difficulty in coordinating the two sensors to point the laser beam to a designated surface location; in the worst case, the laser spot might be occluded by some closer objects and therefore is not in the field of view of the PTZ camera. We are looking into these problems in the design of our next generation long-range audio-video sensing platform by using a pair of stereo cameras that can measure distance without actually seeing the laser spot in images.

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] Polytec Laser Vibrometer, http://www.polytec.com/. Last visited March 2010.

[9] Ometron Systems. http://www.ometron.com/. Last visited March 2010.

[10] A. Z. Khan, A.B. Stanbridge and D.J. Ewins, "Detecting damage in vibrating structures with a scanning LDV," Optics and Lasers in Engineering, 32(6), 1999:583-592.

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

[12] 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.

[13] P. Lai, et al, "A robust feature selection method for noncontact biometrics based on Laser Doppler Vibrometry," IEEE Biometrics Symposium, 2008: 65-70.

[14] 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.

[15] H.Y. Chen, D. Sun, and J. Yang, "Global localization of multirobot formations using ceiling vision SLAM strategy," Mechatronics, 19(5), 2009:617-628.

[16] .H.Y. Chen, D. Sun, J. Yang, and J. Chen, "SLAM Based Global Localization for Multi-robot Formations in Indoor Environment," IEEE/ASME Transactions on Mechatronics, 14(5), 2009.

[17] S. Shair, J. H. Chandler, V. J. Gonz´alez-Villela,R. M. Parkin,and M. R. Jackson, "The Use of Aerial Images and GPS for Mobile Robot Waypoint Navigation," IEEE/ASME Transactions on Mechatronics, 13(6), 2008: 692-699.

[18] E. Kim, K. Hyun, S. Kim and Y. Kwak, "Improved Emotion Recognition with a Novel Speaker-Independent Feature," IEEE/ASME Transactions on Mechatronics, 14(3), 2009:317 – 325.

[19] C. B. Scruby and L. E. Drain, "Laser Ultrasonics Technologies and Applications," Madison Avenue, New York:Taylor & Francis, 1990.

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

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

[22] D. Comanicu and P. Meer, "Mean shift: a robust approach toward feature space analysis,". *IEEE Trans. Patten Analysis and Machine Intelligence*, May 2002

[23] G. Bradski and A. Kaebler. Learning OpenCV: computer vision with the OpenCV library. Sebastopol , CA:O'Reilly Media Inc, 2008.

[24] 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.

Yufu Qu received his MS and PhD degree in instrument science and technology from Harbin Institute of Technology in 2001 and 2004, respectively, China. He joined Beijing University of Aeronautics and Astronautics (BUAA) in 2004, and is currently an Associate Professor in the School of Instrumentation Science and Opto-electronics Engineering at BUAA. Since 2008, he has been visiting in the Visual Computing Laboratory in the Department of Computer Science of City University of New York, USA, as a Postdoctoral Follow. His research interests include multimodal sensor design and integration, optical sensing, precise inspection and machine vision.

Tao Wang received his B.S degree in Computer Science from Stony Brook University, New York, in 2002, and the M.Eng degree in Civil Engineering from Cornell University, New York, in 2004. He is now a Ph.D. student at the Graduate Center of City University of New York.

Since 2006, he has been a research assistant in the City College Visual Computing Laboratory, working on multimodal sensor design and integration, and video surveillance. He is a student member of IEEE.

Zhigang Zhu received his B.E., M.E. and Ph.D. degrees, all in computer science from Tsinghua University, Beijing, China, in 1988, 1991 and 1997, respectively. He is currently a Full Professor in the Department of Computer Science, the City College and the Graduate Center, at the City University of New York. He is Director of the City College Visual Computing Laboratory (CcvcL), and Co-Director of the Center for Perceptual Robotics, Intelligent Sensors and Machines (PRISM) at CCNY. Previously he has been an Associate Professor at Tsinghua University, and a Senior Research Fellow at the University of Massachusetts, Amherst. His research interests include 3D computer vision, multimodal sensing, virtual / augmented reality, video representation, and various applications in education, environment, robotics, surveillance and transportation. He has published over 100 technical papers in the related fields. He is a senior member of the IEEE, a senior member of the ACM and an associate editor of the Machine Vision Applications Journal.