A Study on Vibration Source Localization Using High-speed Vision

（高速ビジョンを用いた振動源定位に関する研究）

by

Mingjun Jiang

Graduate School of Engineering
Hiroshima University
April, 2017

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

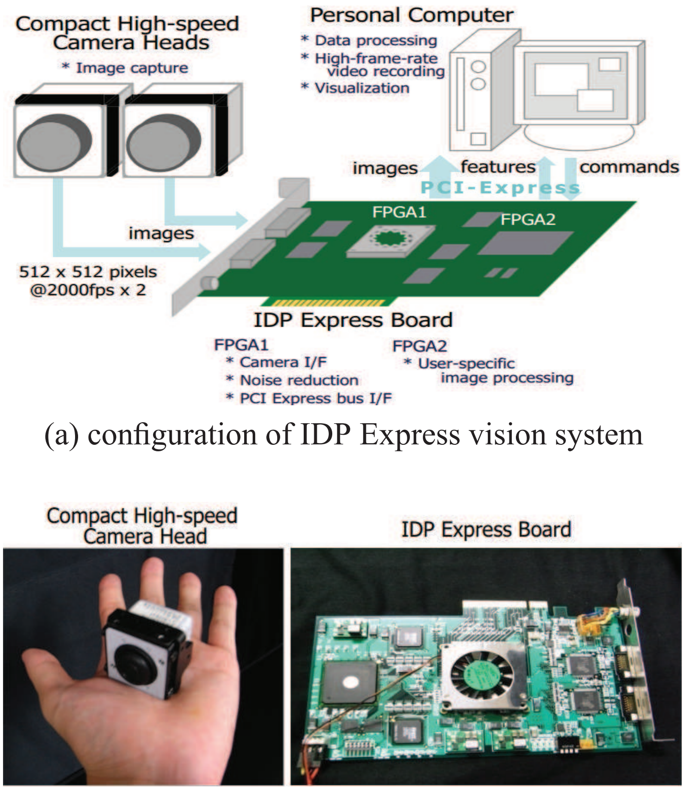
1.1.1 High-speed Vision

In the past decades, many kinds of vision systems have been applied to various fields, such as multimedia, industrial inspection, three-dimensional reconstruction, traffic system, and so on. Most of conventional vision systems with standard video signals (NTSC 30 fps / PAL 25 fps) are designed on the basis of the characteristics of the human eye, which implies that the processing speed of these systems is limited to the recognition speed of human eye. However, due to the bottleneck of sampling rate, the conventional vision systems are not applicable for high-speed phenomenon which cannot be observed with naked human eye, such as factory automation high-speed production line, hyper-human manipulation, and high-speed object tracking in robotics. Therefore, various high-speed vision systems that can operate at sound-level frame rate have been developed for various hyper-human applications.

The transmission speed limitation from photo-detectors (PD) to processing elements (PE) is the sampling rate bottleneck of conventional vision systems. To overcome this limitation, vision chips have been developed and execute real-time processes at a rate of 1000 fps or more by integrating sensors and processors compactly. Bernard et al. proposed an on-chip array of bare Boolean processors with half toning facilities and developed a 65×76 Boolean retina on a 50 mm^2 CMOS 2 μm circuit for the imager of an artificial retina [1]. Eklund et al. verified the near-sensor image processing (NSIP) concept, which describes a method to implement a two-dimensional (2-D) image sensor array with processing capacity in every pixel, and have fabricated and measured a 32×32 pixels NSIP [2]. Ishikawa et al. have developed a COMS vision chip for 1ms image processing and proposed the S^3PE(simple and smart sensory processing elements) vision chip architecture with each PE connected to a PD without scanning circuits [3, 4]. Komuro et al. proposed a dynamically reconfigurable single-instruction multiple-data (SIMD) processor for a vision chip and developed a prototype vision chip based on their proposed architecture, which has 64×64 pixels in a 5.4 mm 5.4 mm area fabricated using the 0.35 μm TLM CMOS process [5]. Ishii et al. proposed a new vision chip architecture specialized for target tracking and recognition, and developed a prototype vision chip using 0.35 μm CMOS DLP/TLM(3LM) process [6].

Many attempts to design real-time high-frame-rate (HFR) vision systems that can process images at hundreds of frames per second or more have been made by implementing image processing algorithms using parallel processing circuits on a field-programmable gate array (FPGA) board that is directly connected to an HFR camera head. Hirai et al. developed an flexibility FPGA-based vision system using the logic circuit to implement the image algorithm [7]. Watanabe et al. developed a high-speed vision system for real-time shape measurement of a moving/deforming object at a rate of 955 fps (256×256 resolution) [8]. Ishii et al. developed a high-resolution high-speed vision platform, H^3(Hiroshima Hyper Human) Vision, which can simultaneously process a 1024×1024 pixels image at 1000 fps and a 256×256 pixels image at 10000 fps by implementing image processing algorithms as hardware logic on a dedicated FPGA board [9]. In the latest two years, Ishii et al. developed a high-speed vision system called IDP Express, as shown in Figure 1.1., which can execute real-time image processing at a rate from 2000 fps (512×512 resolution) to 10000 fps (512×96 resolution), and high frame rate video recording simultaneously [10].

At present, high-speed vision systems can be used as robot sensors at hundreds of hertz or more; several applications of these systems have been also reported. Ishii et al. proposed a simple algorithm for high-speed target tracking using the feature of high speed vision, and realized target tracking on the 1 ms visual feedback system [11]. Nakabo et al.



(a) configuration of IDP Express vision system

(b) photo of IDP Express vision system

Figure 1.1: IDP Express high-speed imaging system

developed a 1 ms vision system, which has a 128×128 PD array and an all parallel processor array connected to each other in a column parallel architecture, for 1 ms cycle-time for visual servoing and applied it to high-speed target tracking [12]. Nakamura applied high-speed vision system to virtual stillness for beating heart surgery [13]. Nakabo et al. developed a 3D target tracking/grasping system, which are composed of a 1ms feedback rate using two high-speed vision systems called column parallel vision (CPV) systems and a robot hand arm [14]. Shiokata et al. proposed a strategy called ”dynamic holding” and developed a experimental robot dribbling using a high-speed multi-fingered hand and a high-speed vision system [15]. Mizusawa et al. used high-speed vision servoing to tweezers type tool manipulation by a three-finger robot hand [16]. Nie et al. developed a real-time scratching behavior quantification system for laboratory mice using high-speed vision [17]. Wang et al. developed an intelligent HFR video logging system to automatically detect high-speed unpredictable behavior and record video comprising images with dimensions of 512×512 pixels at a rate of 1000 fps [18]. Yang et al. proposed the concept of dynamics-based visual inspection for verification of the structural dynamics of vibrating object [19], and proposed a“modal radar” algorithm as a structural damage analyses methodology for modal testing that can localize and accurately quantify structural damage which is difficult to detect by appearance-based visual inspection of a single image [20]. Okumura et al. developed “1 ms Auto Pan-Tilt” technology, which can automatically control the camera’s pan-tilt angles with high-speed vision feedback information to always keep a object at the center of the field [21]. Ishii et al. developed an HFR laryngoscope that can measure the vibration distribution of a vocal fold in real time at hundreds of hertz [22]. Gao et al. proposed a novel light-section method that can accurately obtain a differential shape from a given reference 3D shape at a high frame rate by projecting a self-projected light pattern [23]. Ishii et al. developed a high-speed vision system that can be applied to real-time face tracking at 500 fps by using the GPU acceleration of a boosting-based face tracking algorithm [24]. Gu et al. proposed a 2000 fps multi-object feature extraction system based on FPGA implementation of the cell-based labeling algorithm, which is suitable for hardware implementation and only few memory is required for multi-object feature extraction [25]. Namiki et al. developed a novel air-hockey robot system that switches strategies according to the playing styles of its opponent by using the visual information received at the rate of 500 Hz [26]. Chen et al. developed an HFR camera-projector system that can acquire and process 512×512 depth images in real time at 500 fps and project computer-generated light patterns onto time-varying 3D scenes [27]. Gu et al. proposed a high-speed vision-based shape and motion analysis system for cells in microchannel flows, operates as a real-time inspection tool for cells flowing in micro channels at several meters per second, where their shapes are deformed corresponding to their mechanical properties [28]. With the development of the sampling rate and resolution, such high-speed vision systems can also observe the vibration distribution of an object excited at dozens or hundreds of hertz, which is too fast for the naked eye and standard NTSC cameras. Figures 1.2 shows a 512×512 image sequence of the explosion of a ballon, recorded with 2000 fps sampling rate vision system.

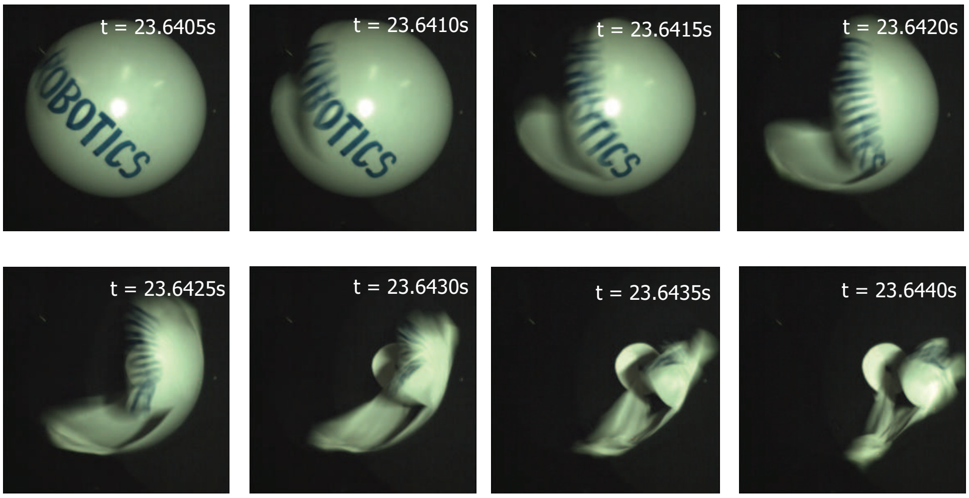


Figure 1.2: High-frame-rate image sequence of the explosion of a balloon

1.1.2 Object Tracking

Tracking the same object robustly against complex appearance variations is a significant task in the field of robot vision [29]. Many researchers have developed object tracking methods and systems that provide a visual representation to robustly describe the spatiotemporal characteristics of object appearance [30]. Object tracking methods using a global visual representation that reflects the global statistical characteristics of an image region to be tracked have been proposed on the basis of various global image features such as optical flows [31, 32, 33], color histograms [34, 35, 36], and texture histograms [37, 38, 39]. By encoding the object appearance information from the selected interest points in images, local-feature-based object tracking methods have also been proposed on the basis of local features such as scale invariant feature transform (SIFT) [40, 41], Haar-like features [42, 43], the histogram of oriented gradient (HOG) [44, 45, 46], and the local binary pattern (LBP) [47, 48, 49]. These appearance-based object tracking methods have been applied in various real-world applications such as traffic monitoring [50, 51, 52], video compression [53], and human-computer interaction [54, 55].

Several unsupervised and semi-supervised object detection methods have been recently proposed to improve the localization accuracy in object tracking. These methods are based on spatio-temporal appearance cues across video frames such as max-path search [56, 57], tubelets [58], fast proposal [59], action tubes [60], bag of fragments [61], and stacked sequential learning (SSL) [62].

However, most appearance-based approaches assume that the target object is being tracked by identifying its spatial statistical pattern, and that the target object can be observed in a certain image region in which its spatial distribution represents its visual appearance. Several attempts at tracking low-resolution targets have been developed [63, 64]. However, appearance-based object tracking suffers from difficulties in handling complex real-world changes in object appearance, which are caused by factors such as illumination variation, lens defocus, shape deformation, and partial occlusion.

1.2 Outline of Thesis

This thesis is organized as 6 Chapters including this introduction.

Chapter 2 summarizes related works on acoustic sensor-based source localization, optical vibration detection, and real-time HFR vision..

In Chapter 3, I proposed a concept of vibration feature extraction with pixel-level digital filters, and the robustnesses of pixel-level vibration source localization against various appearance changes in HFR videos are analyzed in Chapter 4.

In Chapter 5, real-time vibration source tracking is implemented on a 1000 fps vision platform and in which latency effect on digital filters can be reduced by applying pixel-level digital filters to clipped region-of-interest (ROI) images.

Chapter 6, the final chapter, summarizes the contributions of this study and discuss the future work.

Chapter 2

Related Works

2.1 Acoustic Sensor-based Source Localization

Sound source localization is one of the important technologies in robot audition that are aimed to simulate human auditory sense in a robot for the detection and tracking of sound sources in the real world [65], and numerous methodologies have been proposed for enhancing sound source localization performance.

Simulating the frequency response of the human pinnae, which have direction-dependent filtering functions for incoming sound waves, monaural spectral cues at different frequencies have been used for monaural sound source localization using a single channel of sound[67, 68, 69].

Corresponding to human audition with the left and right ear, many binaural source localization methods that use interaural cues derived by differentiating the acoustic features at the left and right channels, such as interaural level difference (ILD) and interaural time difference (ITD), have been proposed for azimuth localization [70, 71, 72, 73]. Considering the acoustic reflections due to the robot head and outer ears, interaural cues have been expanded in the spectral domain for elevation localization [74, 75, 76]. Triangulation-based methods that use the azimuth directions estimated at two positions have been proposed for distance localization [77, 78, 79]. However, the binaural approach is limited in terms of source-localization accuracy when multiple sources are located in a noisy environment, owing to the number of sound channels.

To improve the performance and robustness of source localization, the microphone array approach using multiple microphones spatially organized along various geometries was proposed as an expansion of the binaural approach [80]. In the correlation methods [81, 82, 83, 84], the cross-correlation matrix among acoustic signals from multiple microphones is computed for indicating the distribution of the time delays of arrivals (TDOAs) at the microphones, and its correlation peaks are estimated for sound source localization. MUltiple-SIgnal-Classification (MUSIC) methods [85, 86, 87] have been proposed for multi-source-localization to estimate the direction of arrival (DOA) by computing the noise subspace with stochastic subspace identification and detecting the orthogonal peaks in the noise subspace and the steering vectors, which correspond to the true directions of the sound sources. Beamforming is a well-known signal processing technique used in sensor arrays for directional signal propagation and has been used for microphone array-based source localization [88, 89, 90] that can exploit the highest acoustic energy in all the estimated directions by setting different gains for each microphone.

These source localization methodologies have been applied to many applications for multi-speaker recognition and tracking, such as human-computer interaction [91, 92, 93], intelligent rooms [94, 95, 96], and mobile robot audition [97, 98, 99]. They have also been used in the fields of industry and transportation, in areas such as operational vibration in product machines [100, 101], automobile vibration tests [102, 103, 104, 105], running trains [106], and aircrafts in flight [107, 108]. Recently, several surveillance systems for flying drone detection [109, 110, 111, 112] have been developed in which the rotating frequencies at dozens or hundreds of Hertz are extracted from the acoustic signals that are emitted by drone propellers. Open source software frameworks and hardwares that integrate state-of-the-art source localization algorithms have been distributed, such as the HARK [113], ManyEars [114], embedded audition for robotics (EAR) [115], and Kinect SDK [116]. HARK is an open-source software for robot audition that uses multi-channel-based source localization. The localization of talking people with the standard deviation of 5 degrees at a distance of 1 to 15 m, have been reported on a telepresence system embedded with an 8-channel circular microphone array [117]. ManyEars is an open microphone-array system for beamforming-based source localization. An omnidirectional-microphone-array system with eight microphones has been reported, in which the angular deviation was better than one degree in localizing a source at a distance of 1.5 m [118]. EAR is an integrated auditory sensor with a linear array of eight microphones, a data-acquisition-board and an FPGA processing unit; it can localize a sound source with a precision of ±1 degree at the center of the microphone array [119]. Kinect SDK provides a solution that can localize a sound source within a range of ±50 degrees in front of the sensor; its error margins are 10 degrees or lower [120]. However, these sound source localization techniques still remain inaccurate when the microphones are distant from the source objects to be observed, because of the low directivity of sound propagation.

2.2 Optical Vibration Detection

In previous decades, many optical sensing methodologies were studied for remote monitoring of small vibration displacements [121].

Laser triangulation [122, 123] is a low-cost optical sensing method that detects the positions of beam spots or patterns projected on an object with optical sensors, such as a position sensitive device (PSD), and computes the distance between the optical sensor and the object to be observed via triangulation, and many laser-triangulation-based studies on structural vibration measurement have been reported [124, 125, 126]. For smaller vibration displacements, the Laser-Doppler-Vibrometer (LDV) system [127] is a well-known optical sensing method that can measure a small change in the optical path length along the beam axis at the level of the laser wavelength as the Doppler shift in the laser frequency, which is caused by the target's movement, and it has been used for small vibration measurements in many application fields from structure damage detection to biomedical dynamics sensing [128, 129, 130, 131]. Most of these optical sensors are designed for small displacement measurements at a single point, and displacement distribution can be obtained with their mechanical scans as a collection of small displacements at different points, which are sampled at different timings; it is not suitable for the dynamic analysis of non-stationary vibration distribution.

HFR video vibration analyses, for which vibration displacements at many points are captured at the same time, have also been reported as optical sensing methods for vibration distributions in the audio frequency range. The HFR-video-based approach allows the analysis of non-stationary vibration distributions, and it had been applied to various applications, such as structural vibration analysis [132, 133, 134] and human vocal fold vibration analysis [135, 136, 137]; the accuracies in most HFR-video-based displacement measurements are limited in terms of their image resolution, which is determined by the pixel pitch of the image sensor. To overcome this pixel-wise limitation in measurement accuracy, HFR-video-based small vibration analyses have been conducted with laser interferometry technologies, such as electronic speckle pattern interferometry (ESPI) [138, 139, 140, 141], for which small displacements at the laser-wavelength level are significantly magnified with the laser interferometric fringe patterns. However, most HFR-video-based analyses have been conducted for offline-captured short-term HFR videos and the vibration dynamics features, such as resonant frequency and source location, have not been used for real-time applications such as visual feedback-based target tracking.

Chapter 3

Concept of Vibration Features with Pixel-Level Digital Filters

3.1 Vibration-based Features

An image sensor is regarded as a collection of photo sensors, the number of which corresponds to the pixel number of the image sensor, and the image intensity at every pixel in an image can be considered as a time sequential signal. The temporal periodic changes in image intensity can be observed at pixels of vibration sources in images, depending on their vibration frequencies in the audio frequency range, when the frame rate of a vision system is sufficiently high to allow vibration measurement. Thus, an HFR vision system can localize spatiotemporal changes in image intensities as a vibration distribution by implementing digital filters, which is one of the basic operations in acoustic signal processing for the analysis of sound and vibration dynamics at all the pixels in images in order to pass signals in a specific band of frequencies for identifying and inspecting their vibration frequency and other properties. Fig. 3.1 shows the concept of HFR-vision-based vibration source localization presented in this study, wherein image features are calculated from vibration distributions using pixel-level digital filters.

Our concept of vibration-based image processing is very effective as a dynamics-based visual inspection technology in various applications fields; it is clearly different from conventional appearance-based visual inspection with spatial pattern recognition using single-frame-based image features. For example, it is difficult for the human eye to localize a small flying insect, such as a mosquito, under complex background conditions,

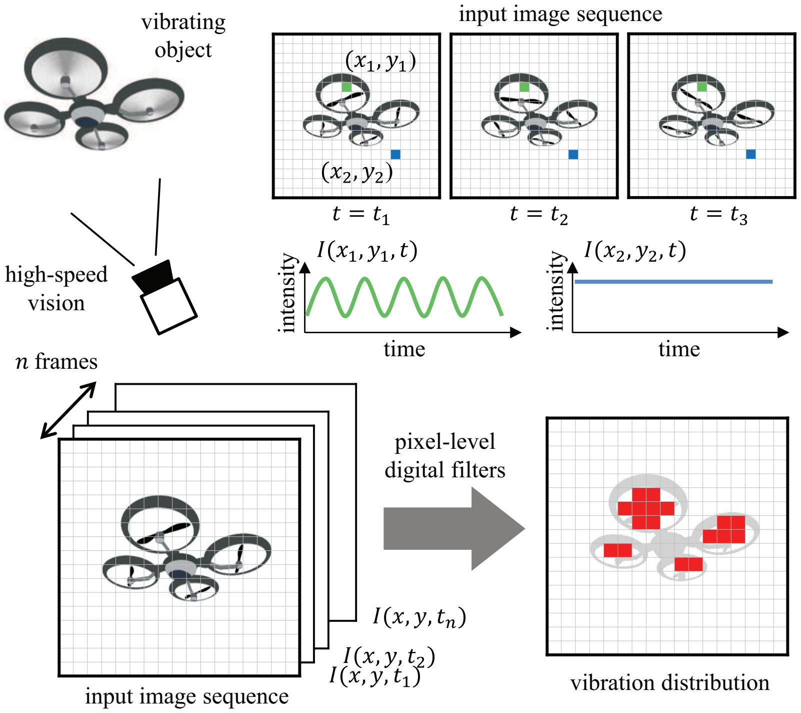


Figure 3.1: Concept of HFR-vision-based vibration source localization with pixel-level digital filters.

because the wings of such small insects beat at very high frequencies at hundreds of Hertz in the audio frequency range. Such rapid wing movements cannot be observed with the human eye, while the human ear can sense the presence of a small insect by detecting its buzzing sound as an acoustic signal. However, the human ear cannot identify its location accurately because of the low directivity of sound propagation; the directivity of sound is frequency dependent. If the periodic changes in the image intensities caused by the beating of the wings of a small insect are detected as vibration-based image features for their localization in an image, the small insect can be accurately localized and be continuously tracked under complex background conditions. In this study, we consider the real-time target tracking of an object vibrating at 100 Hz or higher, such as a flying drone with rotating propellers, to provide an example of HFR-vision-based vibration source localization using vibration-based image features, which can be simultaneously extracted with pixel-level digital filters from an HFR video.

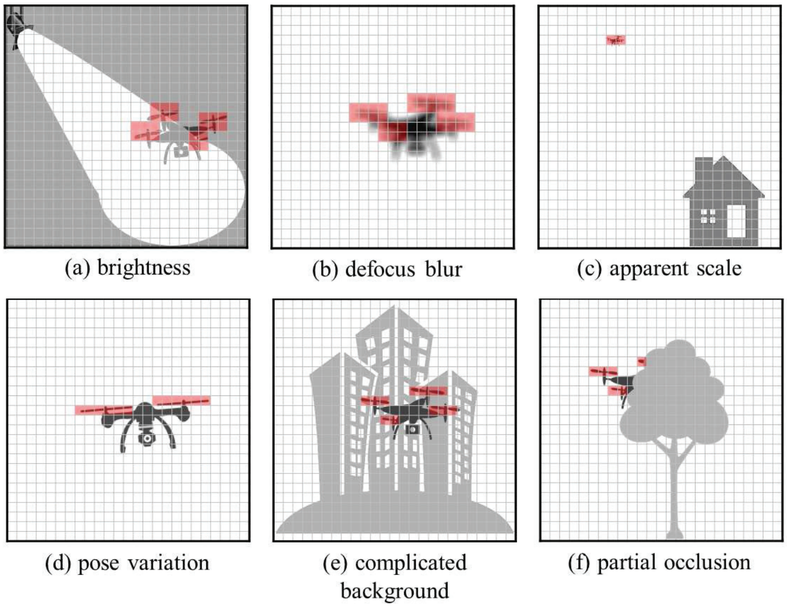


Figure 3.2: Robustness of vibration features with pixel-level digital filters.

3.2 Robustness of Vibration Features with Pixel-Level Digital Filters

Such a vibration feature can detect the temporal brightness variation in the audio-frequency range at every pixel on the premise that the input images are captured at a high frame rate.

Thus, it is very robust against the degradation of the image quality and the target's appearance variation especially when the frequency range of the vibration source is largely distant from that of background scenes, as illustrated in Figure 3.2, because it enables pixel-wise vibration source localization only by implementing band-pass filters at all the pixels in images without any spatial appearance representation. Such a very simple vibration feature with band-pass filters is suitable for real-time vibration source localization for drone tracking, where the operation frequency range of the drone's propellers is much higher than that of the temporal brightness changes at pixels around non-propeller regions in images. When a vibrating object such as a flying drone with rotating propellers is captured in low-quality images using a zoom camera at a very-long distance (and thus with limitations on the resolution of the lens and image sensor), the pixel-wise vibration feature can accurately localize the vibrating object in the low-quality images. This is despite the images being too spatially defocused or low-resolution for conventional appearance-based approaches to identify the target. Thus, in the design of vibration-object tracking systems, it is important to quantitatively verify the localization accuracy and detectability of such a pixel-wise vibration-feature under degraded video-shooting conditions (such as poor brightness, lens defocus, and low-resolution images) and confirm its robustness against object appearance variations (such as object pose variations, complex background scenes, and partial occlusions).

Chapter 4

Robustness Analysis of Vibration Features Against Appearance Changes in High-Frame-Rate Videos

4.1 Introduction

In this chapter, we investigate the effect of appearance variations on the detectability of vibration feature extraction with pixel-level digital filters for HFR videos. In particular, we consider robust vibrating object tracking, which is clearly different from conventional appearance-based object tracking with spatial pattern recognition in a high-quality image region of a certain size. For HFR videos of a rotating fan located at different positions and orientations and captured at 2000 or 300 frames per second with different lens or exposure time settings, we verify how many pixels are extracted as vibrating regions with pixel-level digital filters. The effectiveness of dynamics-based vibration features is demonstrated by examining the robustness against changes in aperture size and the focal condition of the camera lens, the apparent size and orientation of the object being tracked, and its rotational frequency, as well as complexities and movements of background scenes and motion blurs in captured videos. Tracking experiments for a flying multicopter with rotating propellers are also described to verify the robustness of localization under complex imaging conditions in outside scenarios.

4.2 Vibration Feature Extraction

Assuming that the input image of $N \times N$ pixels is captured at time t as $I(x, t)$, and the properties of a vibrating object are initially given, such as its center frequency f_0. The vibration feature to be evaluated in this study is calculated as follows:

(1) Pixel-level band-pass filter

The input image $I(\boldsymbol{x}, t)$ is filtered at every pixel $\boldsymbol{x} = (x, y)$ with a band-pass filter of the center frequency f_0 by adopting the following infinite impulse response (IIR) filter:

$$g(\boldsymbol{x}, t) = \sum_{s=0}^{p-1} b_s I(\boldsymbol{x}, t - s) - \sum_{s=1}^{p-1} a_s g(\boldsymbol{x}, t - s) \tag{4.1}$$

where p is the filter order and a_s, b_s are the tap coefficients. These parameters determine the center frequency and bandwidth of the filter.

(2) Amplitudes of filtered image intensities

To remove the offset values in the image intensities, the differences between the maximum and minimum values of $I(\boldsymbol{x}, t)$ and $g(\boldsymbol{x}, t)$ are computed at every pixel over a cycle of the target's vibration, $T_0 = 1/f_0$, for $t - T_0 \sim t$ as the following amplitudes of the image intensities at time t:

$$I_A(\boldsymbol{x}, t) = I_{max}(\boldsymbol{x}, t) - I_{min}(\boldsymbol{x}, t) \tag{4.2}$$

$$g_A(\boldsymbol{x}, t) = g_{max}(\boldsymbol{x}, t) - g_{min}(\boldsymbol{x}, t) \tag{4.3}$$

where the maximum and minimum values are calculated as follows:

$$I_{max}(\boldsymbol{x}, t) = \max_{t-T_0<t'\leq t} I(\boldsymbol{x}, t') \quad I_{min}(\boldsymbol{x}, t) = \min_{t-T_0<t'\leq t} I(\boldsymbol{x}, t') \tag{4.4}$$

$$g_{max}(\boldsymbol{x}, t) = \max_{t-T_0<t'\leq t} g(\boldsymbol{x}, t') \quad g_{min}(\boldsymbol{x}, t) = \min_{t-T_0<t'\leq t} g(\boldsymbol{x}, t') \tag{4.5}$$

(3) Moving averages of filtered amplitudes

The average amplitude value of the brightness of the input image in a certain interval ΔT_f and that of the intensity and the filtered image are calculated at every pixel as:

$$
\begin{aligned}
K(\boldsymbol{x}, t) &= \frac{1}{\Delta T_f} \int_{t-\Delta T_f}^{t} I_A(\boldsymbol{x}, t) dt && (4.6) \\
G(\boldsymbol{x}, t) &= \frac{1}{\Delta T_f} \int_{t-\Delta T_f}^{t} g_A(\boldsymbol{x}, t) dt && (4.7)
\end{aligned}
$$

where ΔT_f is set to several times the cycle time T_0.

(4) Vibration pixel localization

By thresholding the ratio of $G(\boldsymbol{x}, t)$ to $K(\boldsymbol{x}, t)$ with a threshold θ_2, the pixel $\boldsymbol{x}$ is judged to be a vibration pixel with the vibration component around the target frequency f_0 as follows:

$$
V(\boldsymbol{x}, t) = \begin{cases} 1 & \left(K(\boldsymbol{x}, t) > \theta_1 \text{ and } \dfrac{G(\boldsymbol{x}, t)}{K(\boldsymbol{x}, t)} > \theta_2 \right) \\ 0 & \text{(otherwise)} \end{cases} \tag{4.8}
$$

where the pixel $\boldsymbol{x}$ is judged to be ambiguous and not extracted when the average amplitude $G(\boldsymbol{x}; t)$ is lower than a threshold θ_1.

In this study, we focus on offline quantitative verification of the accuracy and detectability in localizing a vibration source such as a flying drone with rotating propellers by using HFR videos, whereas we evaluate the execution times of our algorithm on a personal computer (PC) in calculating the above-mentioned processes of (1)~(4) toward future real-time implementation. Table 4.1 summarizes the execution times for our algorithm for different image sizes. Here we used a PC with an ASUSTek SABERTOOTH X79 mainboard, Intel Core i7-4820K @ 3.70 GHz CPU, 8GB memory, and two 16-lane PCI-e 2.0 buses with Windows 7 Enterprise 64-bit OS, and the filter order was set to $p = 4$, which is the same parameter used in the experiments in Sections4.4. The execution time for our algorithm increased in proportion with the total number of image pixels. In the case of real-time software execution, the operable frame rates of a vision system are 6143, 1517, 372, 96, 25, and 6 fps for images with different sizes of 64 × 64, 128 × 128, 256 × 256, 512 × 512, 1024 × 1024, and 2048 × 2048 pixels, respectively. Low

Table 4.1: Execution times on PC.

image size	64 × 64	128 × 128	256 × 256	512 × 512	1024 × 1024	2048 × 2048
exec time	0.16 ms	0.66 ms	2.69 ms	10.47 ms	39.78 ms	157.38 ms

resolution images can only be processed by software in real time at thousands of fps, whereas our algorithm should be accelerated for real-time processing of higher resolution images at high frame rates by implementing parallel processing logics of our algorithm on specific accelerators such as FPGAs (Field Programmable Gate Arrays) and GPGPUs (General-Purpose Graphic Processing Units).

4.3 Experiments for a Rotating Fan

We extracted the vibration features from HFR videos captured with different lens settings to consider the robustness under the following seven imaging conditions.

4.3.1 Image Intensity

Several 512 × 512 videos of a rotating fan were captured at 2000 fps with different aperture values, which were adjusted to simulate various image intensities. We applied pixel-level digital filters to these videos to analyze the robustness of the proposed vibration-based localization method under brightness variations.

Figure 4.1 illustrates the video shooting conditions. Three 13-cm-diameter fans with three blades were set at a distance of 20 m in front of the camera against a black background. The center fan was the target, rotating at 37 revolutions per second (rps), and the other two fans were rotating at 44 rps and 26 rps (left and right of the camera view, respectively). These acted as obstacles to the tracked vibration motion. We used a zoom lens with an adjustable focal length and maximum aperture of 16~160 mm and F2.0, respectively. We fixed the focal length to 90 mm, giving a measurement area of 1600 × 1600 mm for 512 × 512 pixels at a distance of 15 m in front of the camera head, where one pixel corresponds to 3.1 mm^2. The tap coefficients a_s, b_s of the pixel-level digital filters were set to operate as band-pass filters with center frequencies of f_0 = 110 Hz and

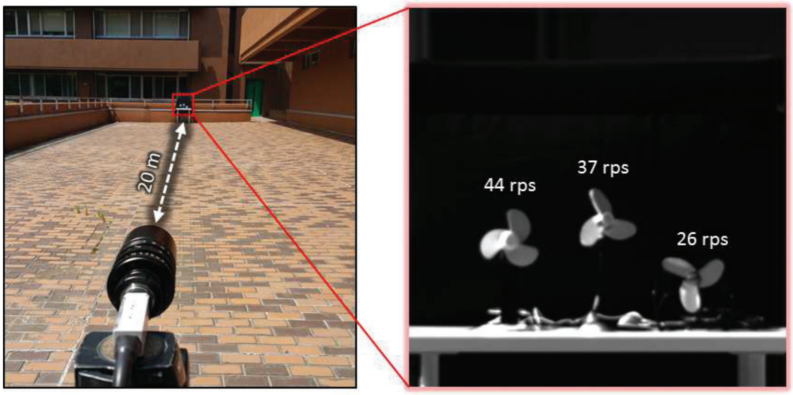


Figure 4.1: Overview of HFR video shoot.

half-widths of 10 Hz. The parameters were set to $p = 4$, $\Delta T_f = 36$ m·s, and $T_0 = 1/f_0 =$ 9 m·s. The thresholds θ_1 and θ_2 for vibration region extraction were set to 30 and 0.5, respectively. These parameters were also used in the experiments reported in the rest of this section.

The aperture value was gradually adjusted from F2.0 to F10.0 with a properly varying interval to darken the images. Figure 4.2a shows five input images of 512 × 512 pixels illustrating the tendency of darkening. Figure 4.2b,c show the moving average distributions of the amplitude of the input images and pixel-wise filtered images, respectively. With the weakening of the image intensity, the amplitude of both the input images and filtered images decreased in the vibration area. However, in Figure 4.2d, the amplitude ratio distributions of filtered images to input images remain roughly uniform under variations in image intensity. The vibration regions were steadily extracted by thresholding these ratio values in our proposed algorithm, as shown in Figure 4.2e.

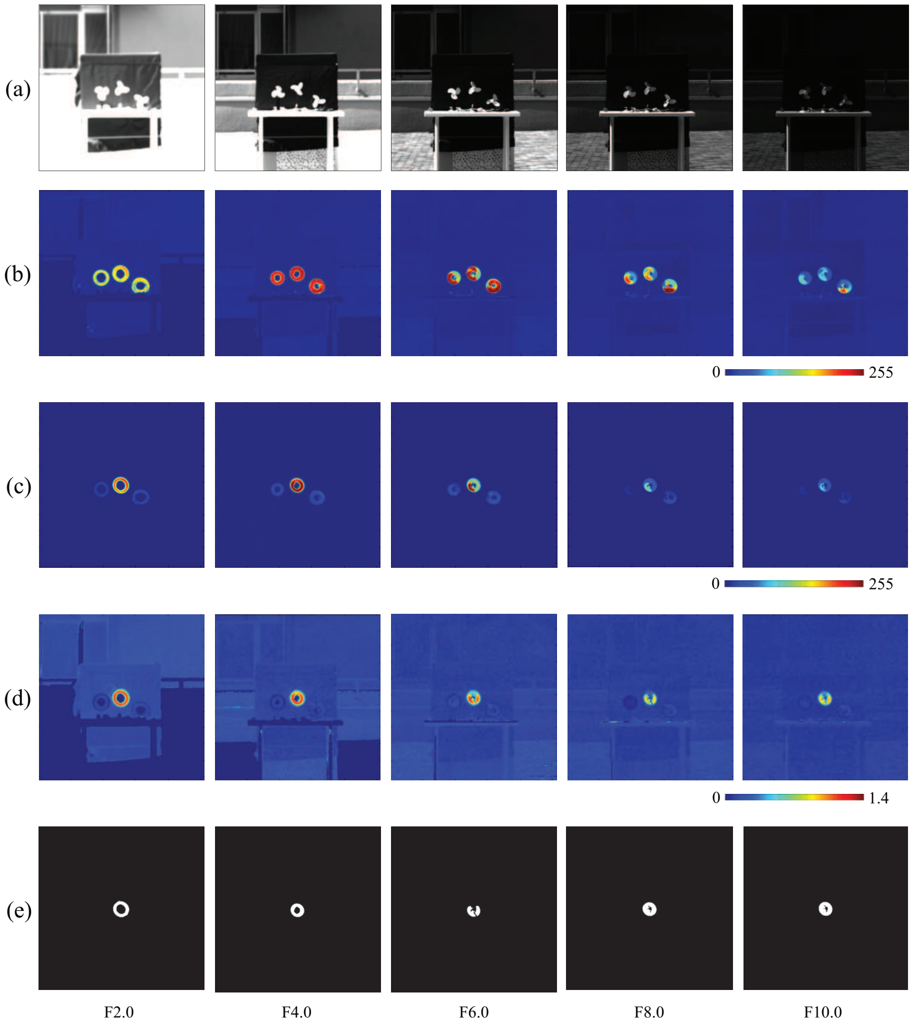
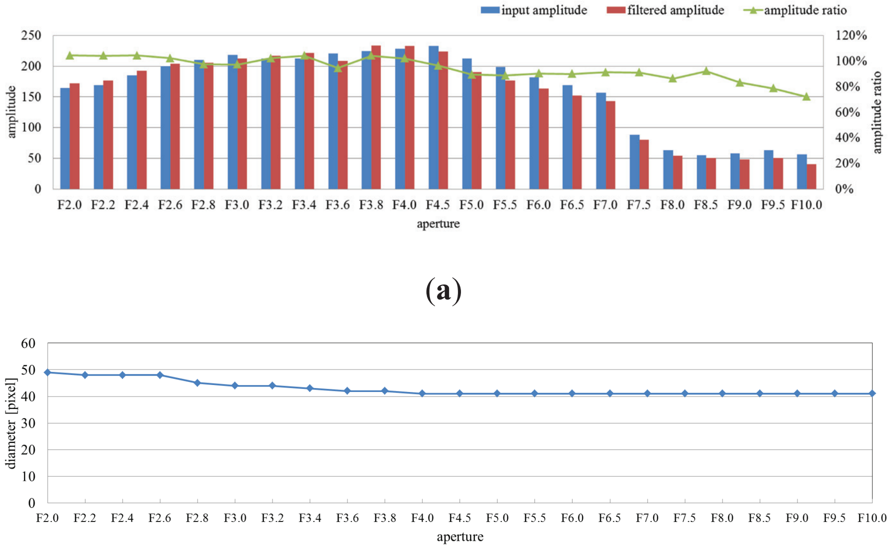


Figure 4.2: (a) Input images; amplitude of (b) input; and (c) pixel-wise filtered images; (d) amplitude ratios; (e) extracted vibration features.

The averaged values of the input and filtered images' amplitude and their ratio in the extracted pixels are shown in Figure 4.3a. The diameters of the extracted vibration region are shown in Figure 4.3b From these figures, we can observe that, although the two amplitudes changed under image intensity variations, the ratios remained between 80% and 110%, and the diameters of the extracted vibration region corresponded to the size of the fan in the captured images (except for exceptional cases containing oversaturated images).



(a)

(b)

Figure 4.3: Averaged amplitudes and extracted region sizes with aperture variation. (a) Averaged amplitudes of input and pixel-wise filtered images and their ratios on the extracted pixels; (b) diameters of extracted vibration region.

4.3.2 Defocus Blur

To analyze the robustness of the proposed vibration extraction method when the vibration source is out of focus, we captured several 512 × 512 videos of three rotating fans at 2000 fps with different focus distances. The three fans and their rotation speeds were as described in Section 4.3.1. In this experiment, they were located 5 m in front of the camera lens. The focal length and aperture value were fixed at 50 mm and F6.0, respectively. For such settings, the measurement area was 790 × 790 mm for 512 × 512 pixels at a distance of 5 m in front of the camera head, where one pixel corresponds to 1.5 mm^2. The focus distance was gradually extended from 1.5 m to an infinite distance by adjusting the lens setting.

Figure 4.4a shows the 512 × 512 input images contaminated by blur of different intensities. Figure 4.4b,c show the moving average distributions of the amplitude of input images and pixel-wise filtered images, respectively. In both cases, the amplitudes on the extracted pixels became greater when the focus distance was set around the camera-object distance and vice versa. As shown in Figure 4.4d, the ratio distributions of the input to filtered amplitudes on the extracted pixels remained roughly uniform at different focus distances, and these were utilized to extract clean vibration regions in Figure 4.4e.

The averages of the input and filtered amplitude and their ratio on the extracted pixels are shown in Figure 4.5a, and the diameters of the extracted vibration region is shown in Figure 4.5b. From these figures, we can observe that, although the two amplitudes change significantly with variations in the focus distance, the ratio values remained between 70% and 80%. The diameters of the extracted vibration region correspond to the size of the fan in the captured images when the focus depth was set around the camera-object distance, and increased when the images were contaminated by the lens blur.

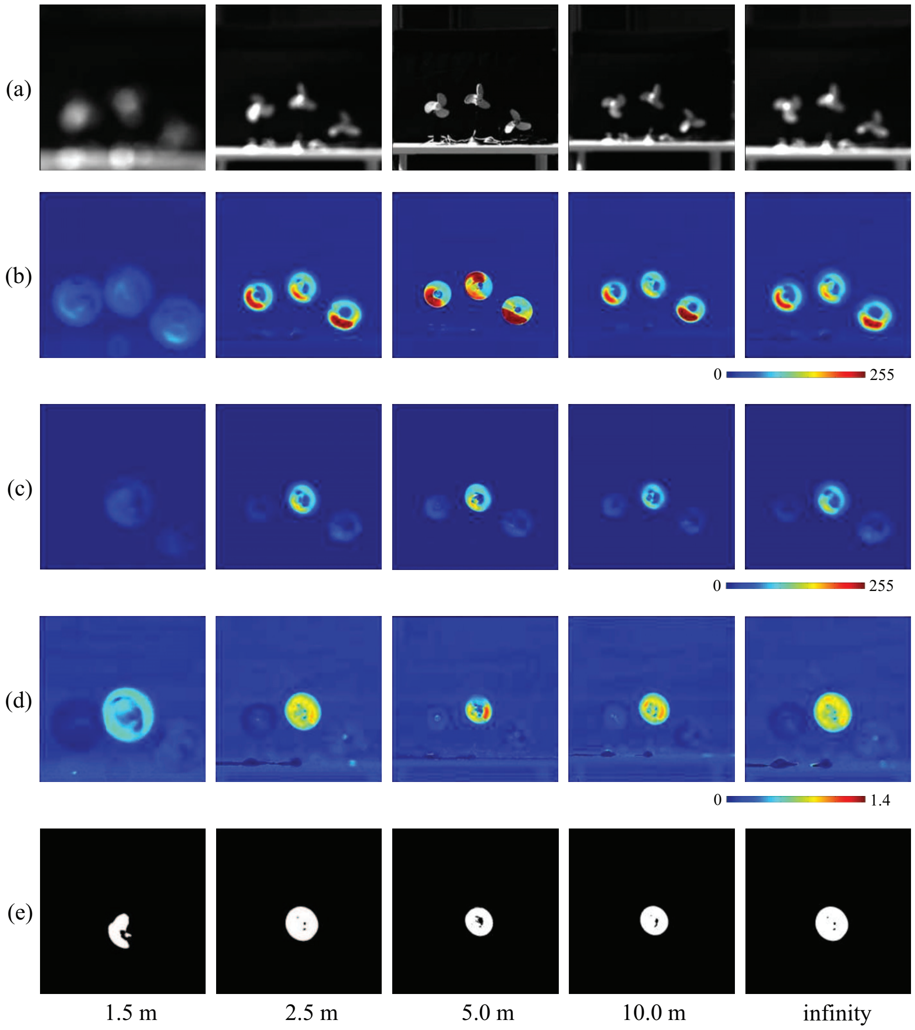


Figure 4.4: (a) Input images; amplitude of (b) input; and (c) pixel-wise filtered images; (d) amplitude ratios; (e) extracted vibration features.

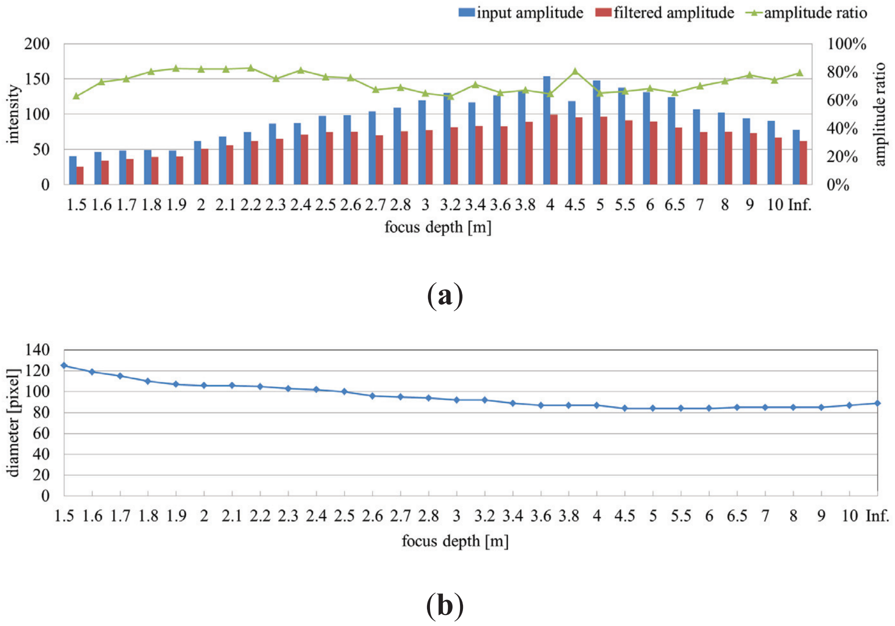


Figure 4.5: Averaged amplitudes and extracted region sizes with focus distance variation. (a) Averaged amplitudes of input and pixel-wise filtered images and their ratios on the extracted pixels; (b) diameters of extracted vibration region.

4.3.3 Apparent Scale

To analyze the robustness of the proposed vibration extraction method when the vibration source is located sufficiently remotely that it is difficult to recognize its appearance from images, we captured several 512 × 512 videos of rotating fans at 2000 fps with different focal lengths. The overall arrangement, including the camera, three fans, and their rotating speed and background, was the same as described in Section 4.3.1, i.e., the distance from the camera to the object was 20 m. The lens aperture was fixed to F5.0 and its focus distance was adjusted to give perfect focus. We gradually adjusted the focal length from 20 mm to 160 mm to simulate changes in the vibration source's apparent scale in the images.

Figure 4.6a shows the input 512 × 512 images of three rotating fans, whose apparent scale is increasing with the focal length. Figure 4.6b,c illustrate the moving average distributions of the amplitude of the input and pixel-wise filtered images, respectively. Although the two amplitudes differed while the focal length was increasing, the ratio distributions remained similar (see Figure 4.6d). Figure 4.6e shows the extracted regions given by thresholding the amplitude ratio of every pixel.

Figure 4.7a quantifies the tendency of the averaged input and filtered images' amplitude and their ratio distribution on the extracted pixels throughout the image-capture procedure. Although the two amplitudes change significantly, the ratio values remained around 80%. The diameters of the extracted vibration region correspond to the increasing size of the fans in the captured images in Figure 4.7b.

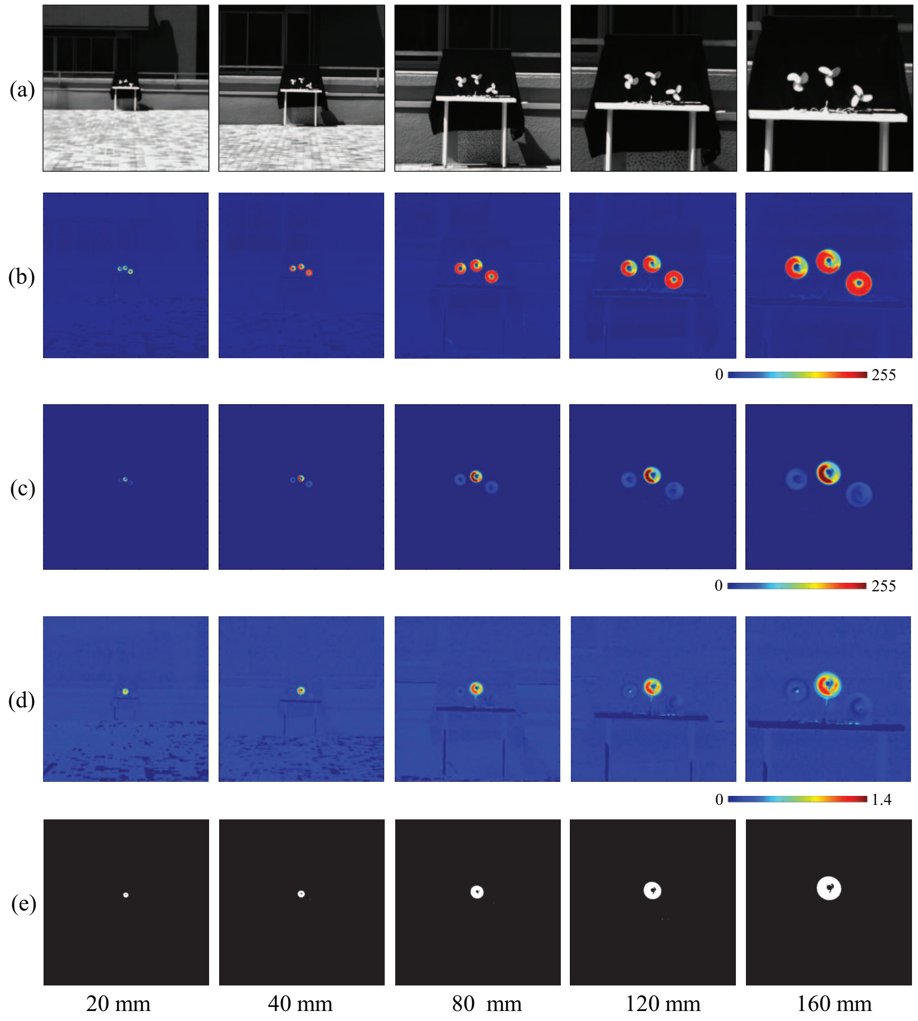
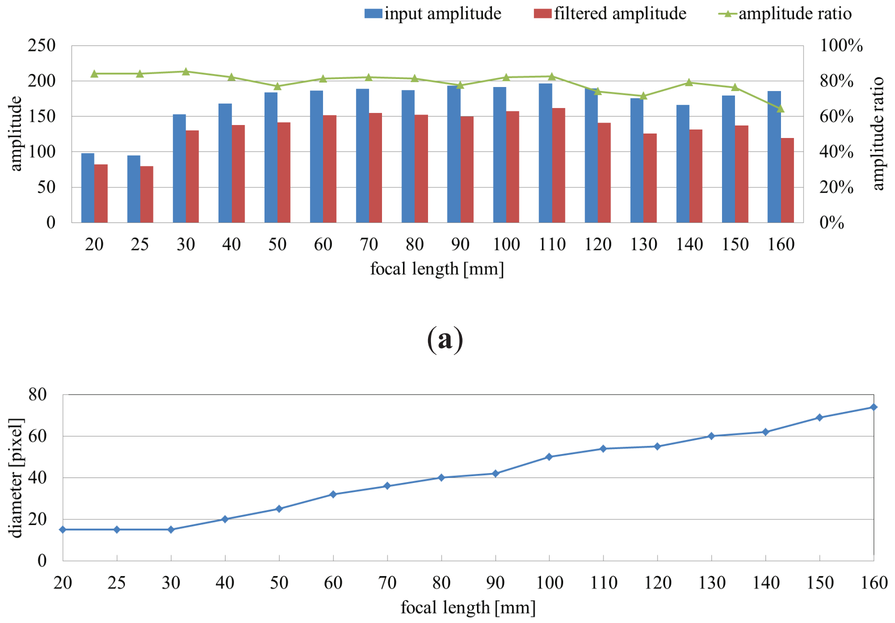


Figure 4.6: (a) Input images; amplitude of (b) input; and (c) pixel-wise filtered images; (d) amplitude ratios; (e) extracted vibration features.



(a)

(b)

Figure 4.7: Averaged amplitudes and extracted region sizes with focal length variation. (a) Averaged amplitudes of input and pixel-wise filtered images and their ratios on the extracted pixels; (b) diameters of extracted vibration region.

4.3.4 Orientation

We analyzed the robustness of detection of the proposed vibration extraction method to changes in the orientation of the vibration source. For this experiment, several 512 × 512 videos of fans rotating at 37 rps were captured at 2000 fps from different orientations. The focal length, focus distance, and aperture were set to 50 mm, 4 m, and F5.0, respectively. The measurement area was 600 × 600 mm for 512 × 512 pixels at a distance of 5 m in front of the camera head, where one pixel corresponds to 1.2 mm^2. The fan was mounted on a goniometer to measure its rotation degree, and was located 4 m in front of the camera. The initial rotation plane was 0° with respect to the camera axis, and the angle was gradually increased to 90° at intervals of 5°.

Figure 4.8a shows the input 512 × 512 images at different orientations towards the camera lens. Figure 4.8b,c show the moving average distributions of amplitude of the input images and pixel-wise filtered images, respectively. Figure 4.8d shows the two amplitudes' ratio distributions, and Figure 4.8e shows the extracted vibration regions.

The averages of the input and filtered images' amplitude and their ratio on the extracted pixels are shown in Figure 4.9a, and the minor axis tendency of the extracted vibration region is shown in Figure 4.9b. From these figures, we can observe that the two amplitudes changed slightly with the rotation, whereas the ratio values remained relatively stable at around 85%. The minor axis of the extracted vibration region corresponds to the size of the fan in the captured images throughout the process.

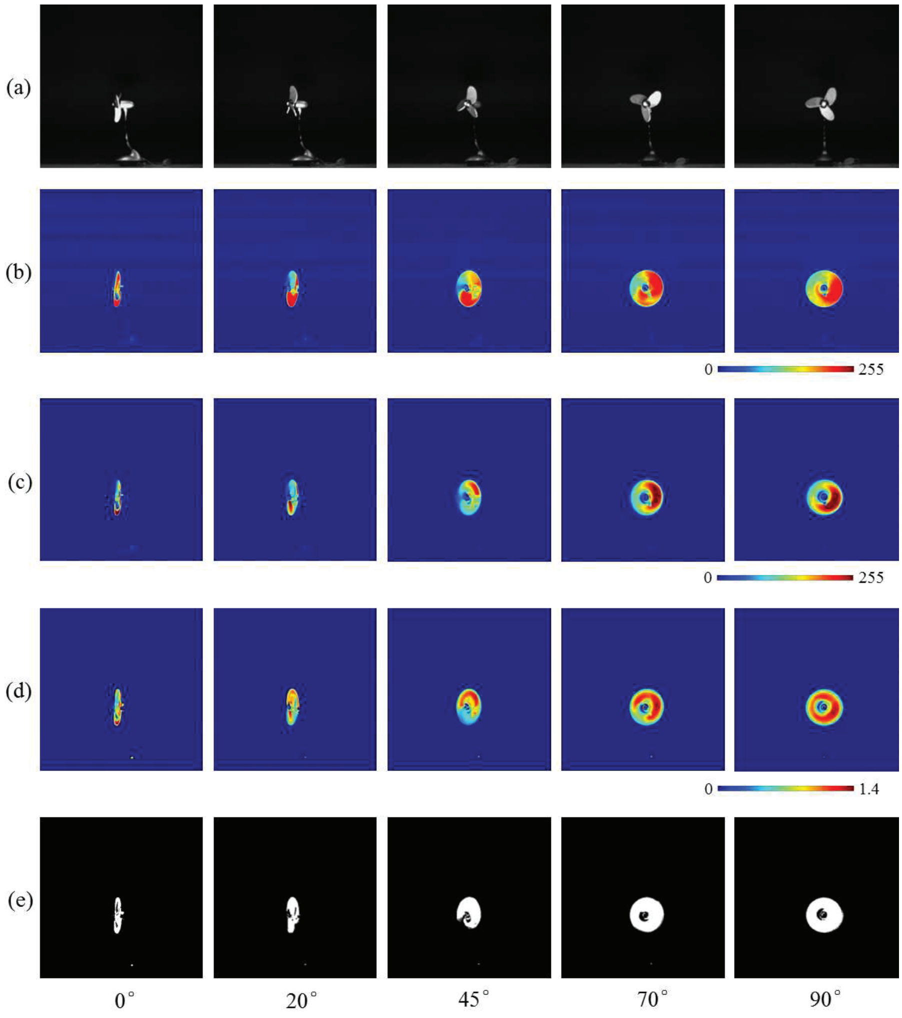
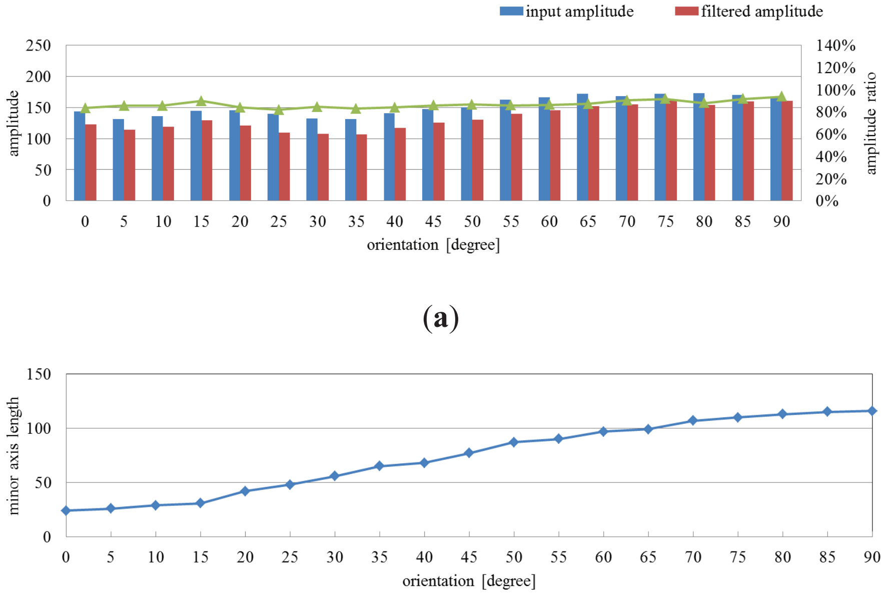


Figure 4.8: (a) Input images; amplitude of (b) input; and (c) pixel-wise filtered images; (d) amplitude ratios; (e) extracted vibration features.



(a)

(b)

Figure 4.9: Averaged amplitude values and extracted region sizes with orientation variation. (a) Averaged amplitudes of input and pixel-wise filtered images and their ratios on the extracted pixels; (b) minor axis lengths of extracted vibration region.

4.3.5 Rotation Speed

We analyzed the frequency range of the proposed vibration extraction method by caturing several 512 × 512 videos of rotating fans at 2000 fps with different rotation speeds. The three fans used in this experiment were as described in Section 4.3.1; the rotation speed of the center fan was gradually increased from 26 rps to 44 rps in intervals of 1 rps, whereas those of the fans on the left and right were fixed at 44 rps and 26 rps, respectively. The distance from the camera to the object was 5 m. The focal length and aperture value were fixed at 50 mm and F1.4, respectively. The measurement area was 790 × 790 mm for 512 × 512 pixels at a distance of 5 m in front of the camera. The tap of coefficients and other parameters of the pixel-level band-pass filters were the same those in Section 4.3.1; their center frequencies and half-widths were 110 Hz and 10 Hz, respectively.

Figure 4.10a shows the 512 × 512 input images with different rotation speeds from

31 rps to 43 rps. Figure 4.10b,c show the moving average distributions of amplitude of the input images and pixel-wise filtered images, respectively. Although the variation of the amplitudes of the input images was small in relation to the rotation speed, those of the extracted pixels around the center three-wing fan became greater when its rotation speed approached 37 rps, whose triple frequency almost corresponds to the center frequency 110 Hz of the band-pass filters. Figure 4.10d shows the ratio distributions of the two amplitudes, and Figure 4.10e shows the extracted vibration regions.

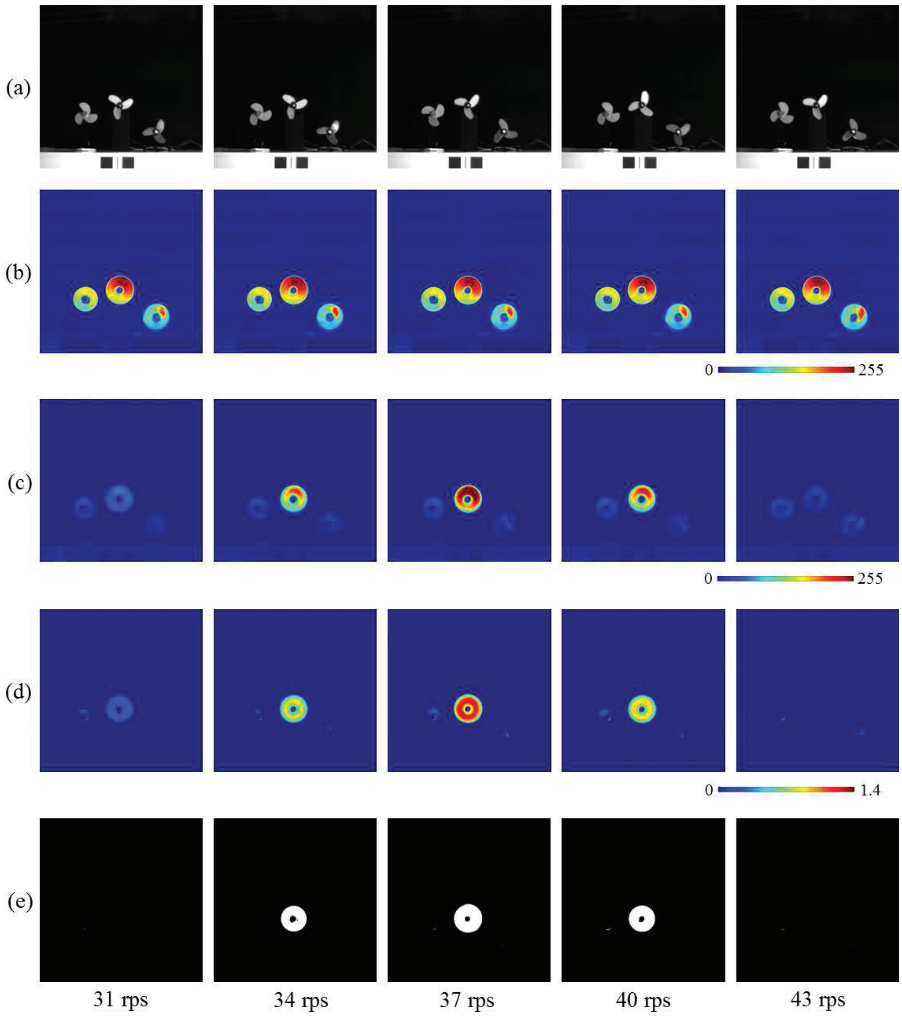
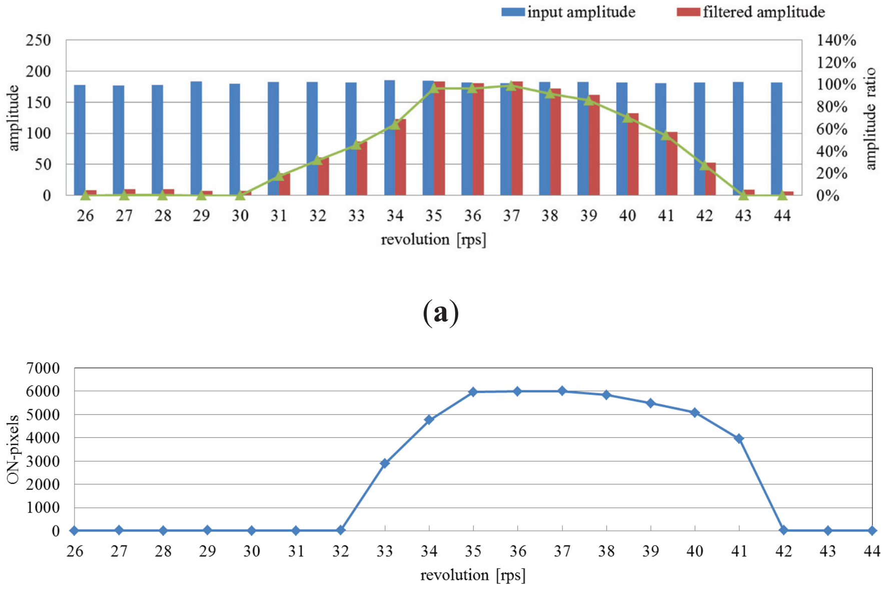


Figure 4.10: (a) Input images; amplitude of (b) input; and (c) pixel-wise filtered images; (d) amplitude ratios; (e) extracted vibration features.

The average amplitude of the input and filtered images and their ratio on the specified pixels around the center fan are shown in Figure 4.11a when the rotation speed of the center fan was changed from 26 to 44 rps; the brightness was periodically changed from 78 to 132 Hz, according to the three wings of the fan. Here the specified pixels around the center fan were set to equal those of the extracted ones when the rotation speed was 37 rps. The number of the extracted pixels as vibration regions is shown in Figure 4.11b. Thus, the pixels around the center fan were distinctly extracted as vibration regions when its rotation speed was within 33 rps from 41 rps, which corresponds to the brightness changes in the frequency range from 99 to 123 Hz. It highly corresponds to the center frequency of 100 Hz and the half-width of 10 Hz of the pixel-level band-pass filters used in this experiment.



(a)

(b)

Figure 4.11: Averaged amplitude values and number of extracted pixels with rotation speed variation. (a) Averaged amplitudes of input and pixel-wise filtered images and their ratios; (b) number of extracted pixels as vibration region.

4.3.6 Moving Fan

We analyzed the robustness of the proposed vibration extraction method when a rotating fan moves against a complicated background scene. We captured 512 × 512 videos of a moving rotating fan for 1.5 s at 2000 fps with the environment illustrated in Figure 4.12. A 37-rps-rotation fan, whose size and rotation speed was the same as those used in Section 4.3.1, was installed on a linear slider. The distance from the camera head to the fan was 2 m. By controlling the slider mechanically, the fan moved alternatively in the right and left directions with an amplitude of 30 cm at a cycle of 1.5 s. A wallpaper patterned with many three-blade propellers, whose shape, size, and color were the same as those of the rotating fan, was used as a spatial jamming pattern in this experiment, because it is very difficult to distinguish the rotating fan from these patterns in a single image. The focal length and aperture value of the lens were 25 mm and F1.4, respectively. The measurement area was 500 × 500 mm^2 for 512 × 512 pixels at a distance of 2 m in front of the camera, where one pixel corresponds to 1 mm^2.

Figure 4.13a shows the input of 512 × 512 images for 1.2 s, taken at intervals of 0.3 s. The translation speeds of the fan were 0.00, 0.96, 0.00, −0.40, and −0.60 m/s at time t = 1.1, 1.4, 1.7, 2.0, and 2.3 s, respectively; the positive/negative signs indicate the movements in the right/left direction. Figure 4.13b,c show the moving average distributions of the amplitude of the input and pixel-wise filtered images, respectively. Both the moving average values in (b) and (c) became larger at the pixels around the moving fan, whereas the moving average distributions of the pixel-wise filtered images were slightly dilated in the direction opposite to the movement direction of the fan, because of the latency effect in the digital filter. Figure 4.13d shows the two amplitudes' ratio distributions, and Figure 4.13e shows the extracted vibration regions. These regions excluded the pixels around the three-blade-fan patterns on the background wallpaper, and they only involved those around the moving fan. Several pixels around the fan were not detected, because of the close similarity of the brightness around its blades with that of the background three-blade-patterns. Thus, the brightness changed very little with time when the fan was passing over the background patterns.

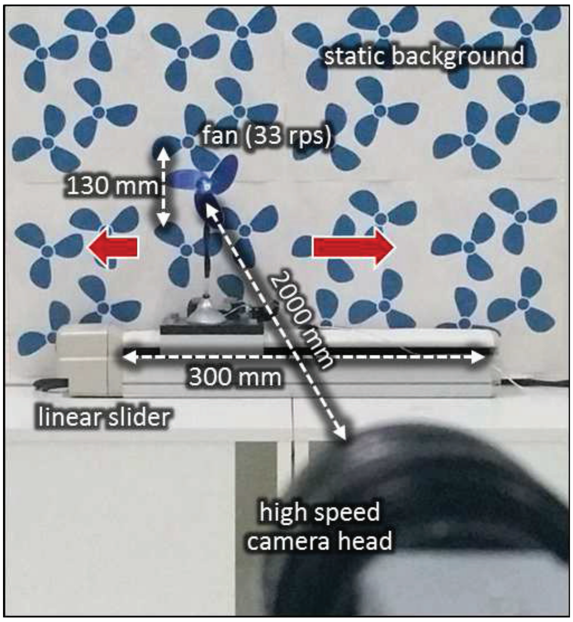


Figure 4.12: Moving fan against three-blades-patterned background.

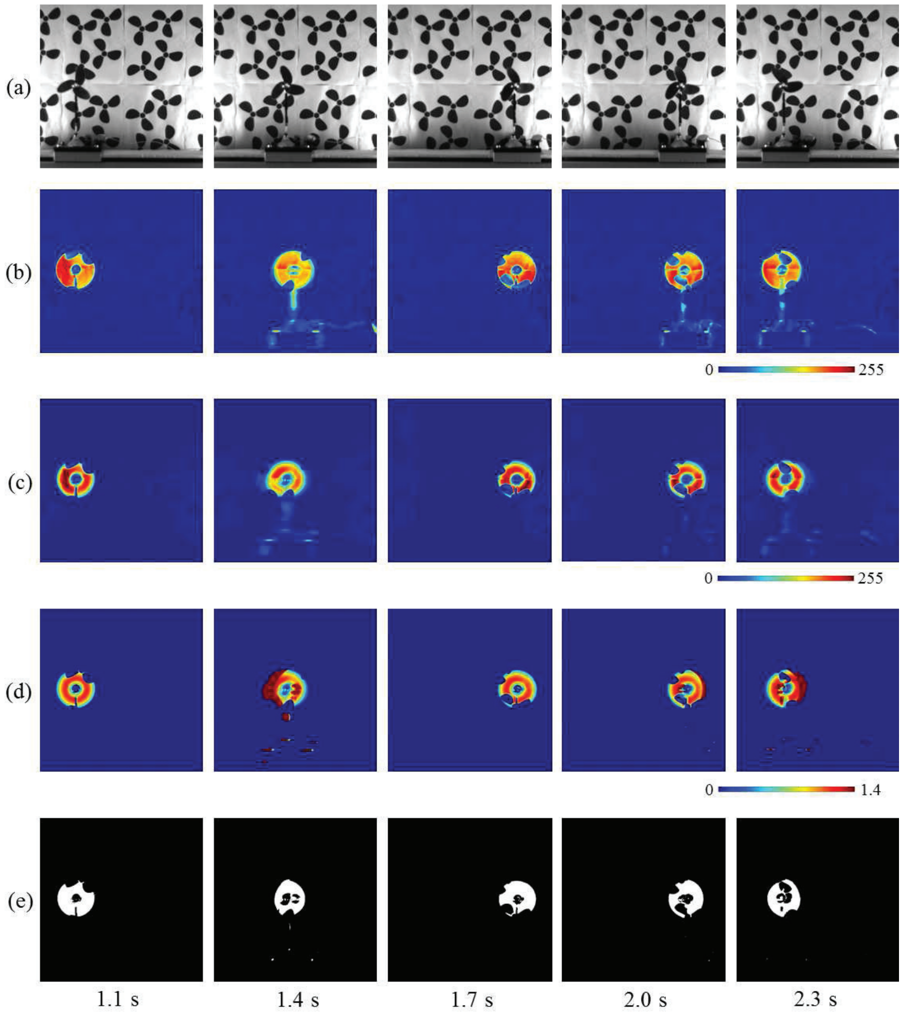
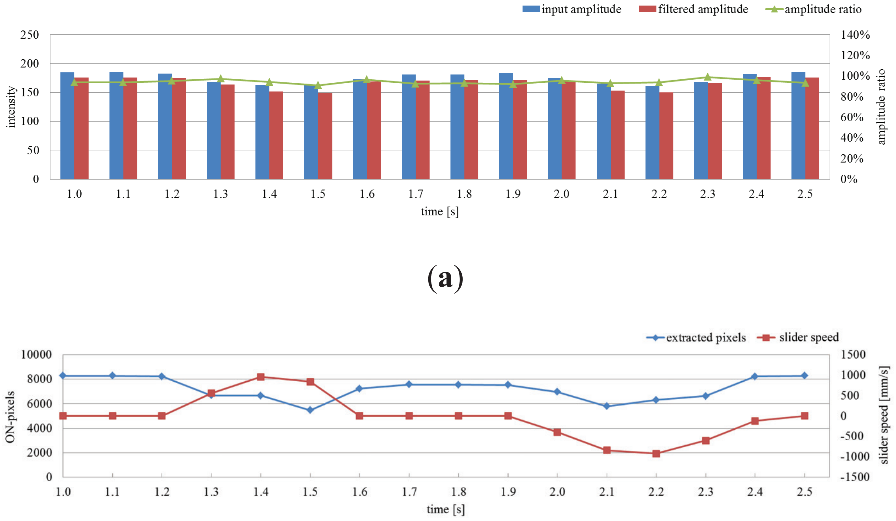


Figure 4.13: (a) Input images; amplitude of (b) input; and (c) pixel-wise filtered images; (d) amplitude ratios; (e) extracted vibration features.

The average amplitude of the input and filtered images and their ratio on the extracted pixels are shown in Figure 4.14a for 1.5 s, and the number of extracted pixels as vibration regions and the translation speeds of the fan are shown in Figure 4.14b. When the rotating fan was moving alternatively in the right and left directions, the ratio remained at around 90% whereas the two amplitudes slightly changed. Here the number of extracted pixels decreased around t = 1.5 and 2.1 s when the translation speed of the fan increased. Because of the latency effect in the digital filter; the vibration features were not extracted at the pixels around the side of the rotating fan opposite to its movement direction as illustrated in Figure 4.13e. Nevertheless, these results apparently indicate the robustness of the proposed vibration extraction method when a rotating fan moves against a complicated background.



(a)

(b)

Figure 4.14: Averaged amplitude values and number of extracted pixels when a rotating fan moves. (a) Averaged amplitudes of input and pixel-wise filtered images and their ratios on the extracted pixels; (b) number of extracted pixels as vibration region and slider speeds.

4.3.7 Moving Background

We analyzed the robustness of the proposed vibration source extraction method when observing a rotating fan against a moving background scene. The experimental setting, which includes the distance from the camera to the fan, the lens parameters, the background pattern, and the moving speed of the linear slider, was similar as that used in Section 4.3.6, except that the 37-rps-rotating fan was fixed and the three-blades-patterned wallpaper was installed on a linear slider to enable the background wallpaper to move in the right and left directions at a cycle time of 1.5 s.

Figure 4.15a shows the input 512 × 512 images. The background moved at speeds of 0.32, 0.64, 0.00, −0.8, and 0.00 m/s at time $t = 1.1, 1.4, 1.7, 2.0$, and 2.3 s, respectively.

Figure 4.15b,c show the moving average distributions of the amplitude of the input and pixel-wise filtered images, respectively. Due to the movement of the background wallpaper, the moving averages in (b) had certain values at the pixels around the edges of the three-blades-patterns, whereas those in (c) became larger only at the pixels around the rotating fan. Figure 4.15d shows the ratio distributions of the two amplitudes, and Figure 4.15e shows the extracted vibration regions. The extracted regions did not include the pixels around the edges of the three-blades-patterns, and they involved only the pixels around the fan. This means that its neighboring pixels were not always detected for the same reason described in Section 4.3.6.

The average amplitude of the input and filtered images and their ratio on the extracted pixels are shown in Figure 4.16a for 1.5 s, and the number of extracted pixels as vibration regions and the speeds of background wallpaper are shown in Figure 4.16b. The two amplitudes slightly fluctuated, whereas the ratio remained at around 90% when the background wallpaper was moving alternatively in the right and left directions. The number of extracted pixels slightly fluctuated because several pixels around the rotating fan were not extracted as illustrated in Figure 4.15e, where the blades of the fan and the moving three-blades-patterns overlapped. Nevertheless, these results apparently indicate the robustness of the proposed vibration extraction method for a rotating fan against a moving patterned background.

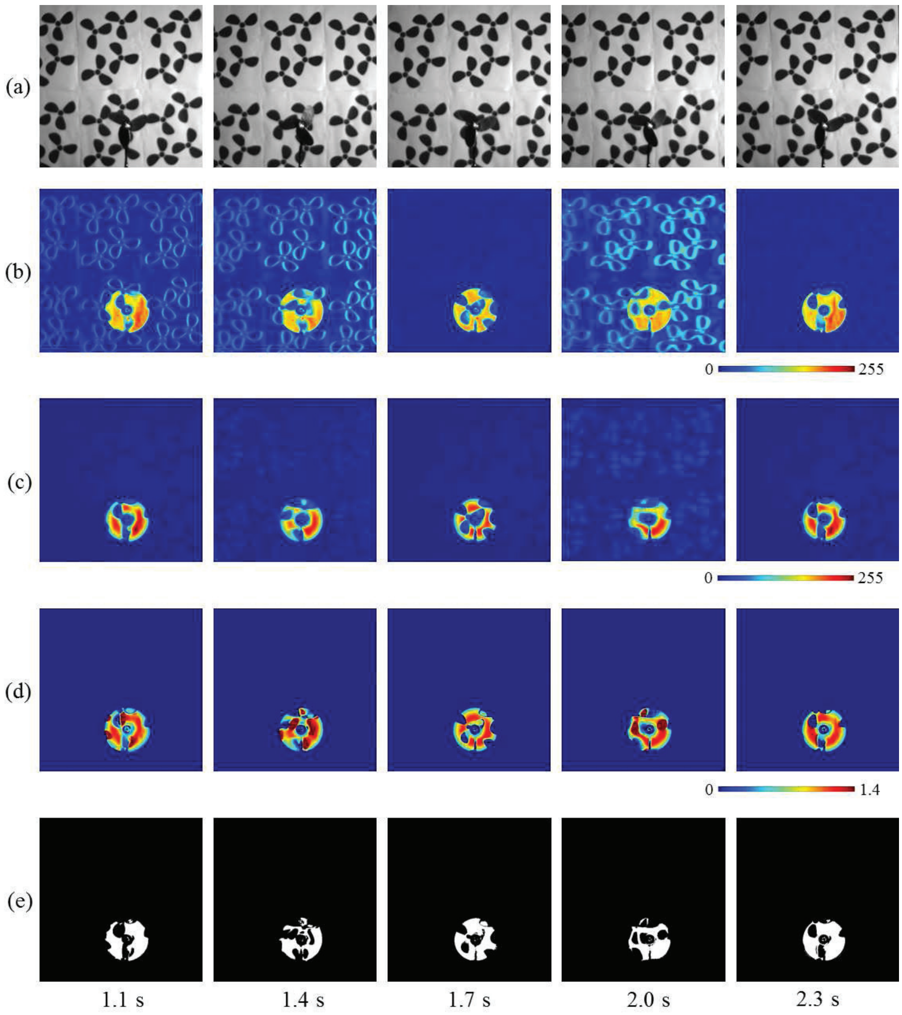
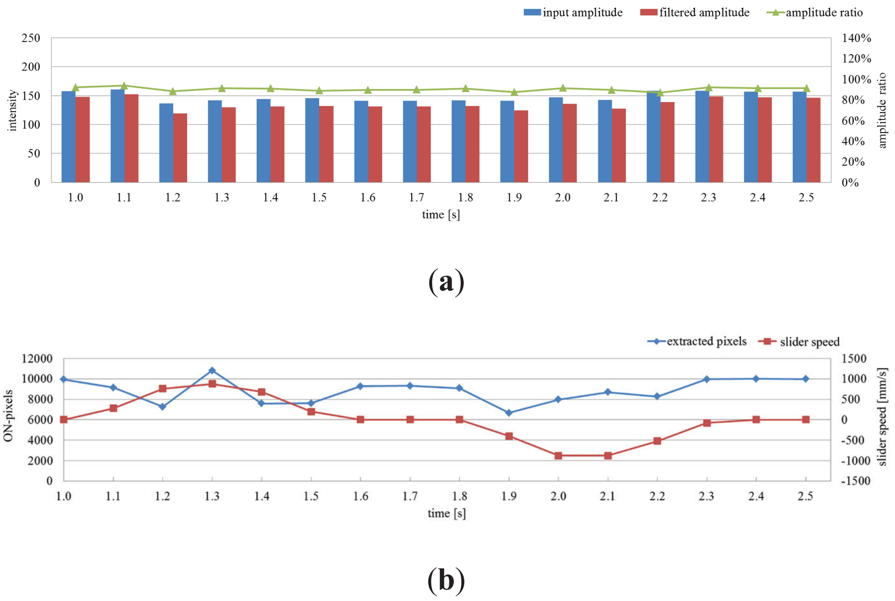


Figure 4.15: (a) Input images; amplitude of (b) input; and (c) pixel-wise filtered images; (d) amplitude ratios; (e) extracted vibration features.



(a)

(b)

Figure 4.16: Averaged amplitude values and number of extracted pixels with moving background. (a) Averaged amplitudes of input and pixel-wise filtered images and their ratios on the extracted pixels; (b) number of extracted pixels as vibration region and slider speeds.

4.3.8 Motion Blur

Several 648×488 videos of a target rotating fan were captured in 300 fps using different exposure times that were adjusted to create motion blurs of different intensity in images. We implemented pixel-level digital filters on these videos to investigate the robustness of vibration-based localization against the motion blur of a vibrating object by analyzing the quantified results.

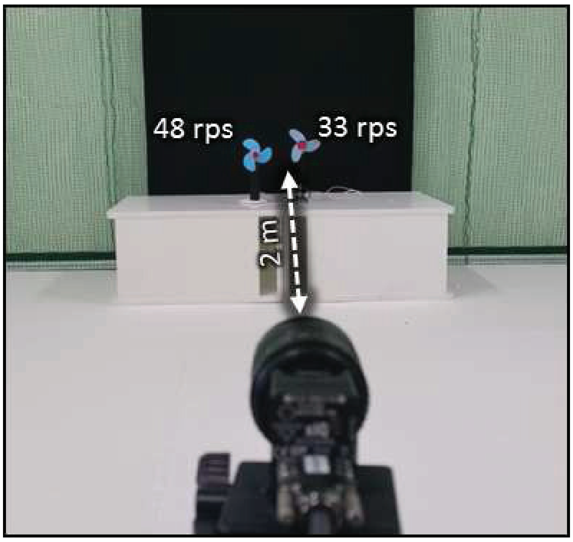


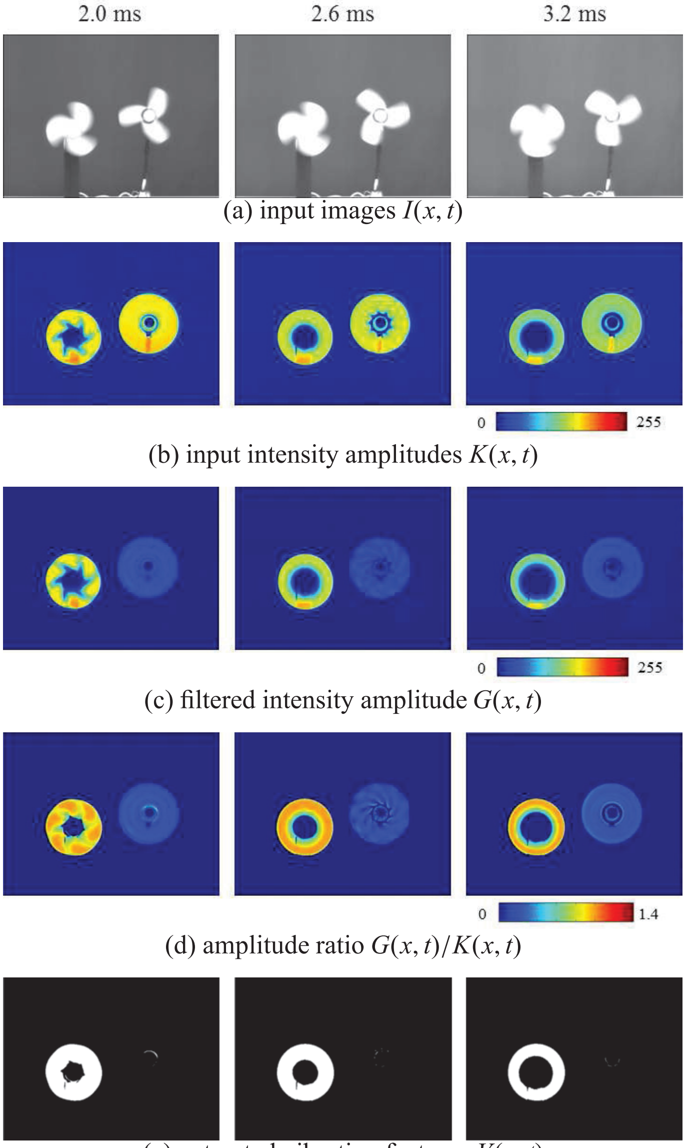
Figure 4.17: Overview of high-frame-rate video shoot.

Figure 4.17 shows the overall view of the video shoot. Two 13 cm diameter fans with three blades were set in front of a black background at a distance of 2 m from the camera. Our target fan was rotating at 43 rps, set to the left of the camera view, and the other fan rotating at 33 rps, was set to the right of the camera view, acting as a obstacle motion. The lens aperture and focal length were set to F1.4 and 25 mm, respectively, and the measurement area was 533×402 mm for 648×488 pixels at a distance of 2 m in front of the camera lens, where one pixel corresponds to 0.68 mm^2. The tap coefficient a_s, b_s of the pixel-level digital filters were set such that they operated as band-pass filters, with a center frequency of $f_0 = 130$ Hz. The half width was 10 Hz to extract the vibration regions of the left rotating fan. The parameters were set to $p = 4$, $\Delta T_f = 12$ ms, and $T_0 = /f_0 = 3$ ms. The thresholds θ_1 and θ_2 for the vibration region extraction were set to 30 and 0.5, respectively.

The exposure time was gradually adjusted from 1.0 ms to 3.2 ms with an interval of 0.2 ms. Figure 4.18(a) shows four input images of 648×488 pixels of different exposure times with different motion blur intensity. Figure 4.18(b) and (c) illustrate amplitude distributions of input images and pixel-wise filtered images at a moment, respectively. Both display a variation with the change of motion blur. Figure 4.18(d) shows the amplitude ratio distributions of pixel-wise filtered images to input images. These are roughly uniform with the change of motion blur. Figure 4.18(e) shows the vibration regions that were extracted by thresholding the amplitude ratios in our proposed algorithm whose size corresponds to the fan's size in the captured images. The averaged values of the input, filtered image amplitude, and corresponding ratio on the extracted vibration regions are illustrated in Figure 4.19. Here, we can observe that although the two amplitudes were varied with the change of motion blur intensity, their ratios remained in the range of 85% to 95%, which can be utilized as a reliable feature for vibration region extraction.

4.4 Experiment for a Flying Multicopter

We analyzed the robustness of our vibration source tracking method with a flying multicopter in two non-controlled outdoor scenarios where additional distraction moving objects and unstructured backgrounds were presented; (a) trees-and-building background; and (b) walking-persons background. In the experiments, we examined that the simultaneous effect of the multiple appearance variations tested in the previous section robustly functions in real scenarios with cluttered and moving backgrounds. The multicopter used in the experiments was an RC EYE One Xtreme (CEI Conrad Electronic Intl. (HK) Ltd., Hong Kong, China.) with four 138-mm dual-blade propellers. The multicopter had dimensions of 225 × 225 × 80 cm, excluding propellers. The flapping frequency of each propeller varied within the range 80–100 Hz according to the flight operation commands. Color 512 × 512 videos of a flying multicopter were captured offline at 1000 fps for 15 s in each scenario with the recording time being limited by the memory size of the high-speed camera. The body and propellers of the multicopter were painted red to extract its location in images for evaluation, whereas our algorithm was processed for gray-level images. In the experiments, the tap coefficients a_s, b_s of the pixel-level digital filters were set to operate as band-pass filters with a center frequency of $f_0 = 80$ Hz (twice the flapping frequency of the dual-blade propellers) and half-width of 20 Hz. The other parameters were set to $p = 4$, $\Delta T_f = 44$ m·s, and $T_0 = 1/f_0 = 6$ m·s. The thresholds θ_1 and θ_2 were set to 20 and 0.5, respectively.



(a) input images $I(x,t)$

(b) input intensity amplitudes $K(x,t)$

(c) filtered intensity amplitude $G(x,t)$

(d) amplitude ratio $G(x,t)/K(x,t)$

(e) extracted vibration features $V(x,t)$

Figure 4.18: (a) input images, intensity amplitude of (b) input and (c) pixel-wise filtered images, (d) amplitude ratios, (e) extracted vibration features.

4.4.1 Trees-and-Building Background

We analyzed the 1000-fps video when the multicopter moves against an unstructured background. The multicopter flew in the right and left directions with vertical elevation twice in 15 s in front of trees and a building, which were located at a distance of approximately 8 m in front of the camera. The focal length, focus distance, and aperture of the lens were set to 12 mm, 8 m, and F2.8, respectively. The measurement areas of 512

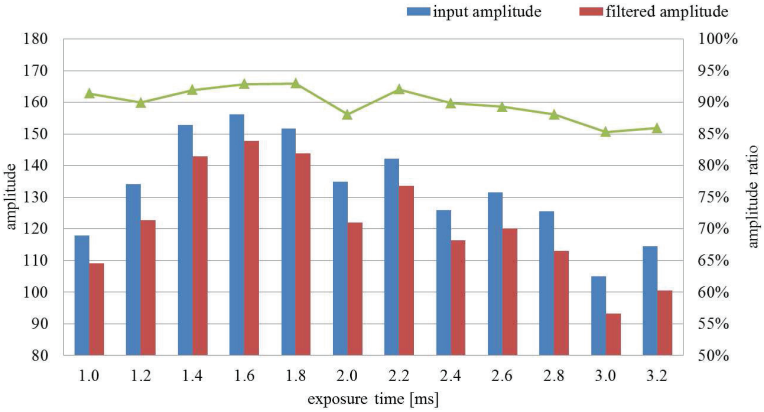


Figure 4.19: average of input, filtered pixels, and corresponding ratios on the extracted pixels.

× 512 pixels were 5.3 × 5.3 m, where one pixel corresponds to 10.3 mm^2 at a distance of 8 m.

Figure 4.20a–d shows the input images and the moving average distributions of the amplitude of the input images and pixel-wise filtered images, as well as the ratio distribution of the two amplitudes'. The images were taken at intervals of 3 s for t = 0–15 s. Figure 4.20e,f show the vibration regions extracted by our algorithm, and magnified images of 32 × 32 pixels around the averaged positions of the extracted pixels, respectively. These averaged positions (blue "+" s) were plotted over the input images as well as those of the red-color regions (red "+" s) in Figure 4.20g; they corresponded to the locations of the red multicopter in images. For comparison, the tracking results of the other appearance-based single-object tracking methods, which were prepared in Open CV Tracking API in Open CV 3.0 [147], were illustrated as color-lined rectangular regions; (1) KCF [142]; (2) TLD [143]; (3) Median Flow [144]; (4) Boosting [145]; and (5) MIL [146]. The color input images at 1000 fps were processed for all the single-object tracking methods, and the object to be tracked was initially defined as the 32 × 24 subimage in the 32 × 32 ROI region at t = 0 s as illustrated in Figure 4.20f.

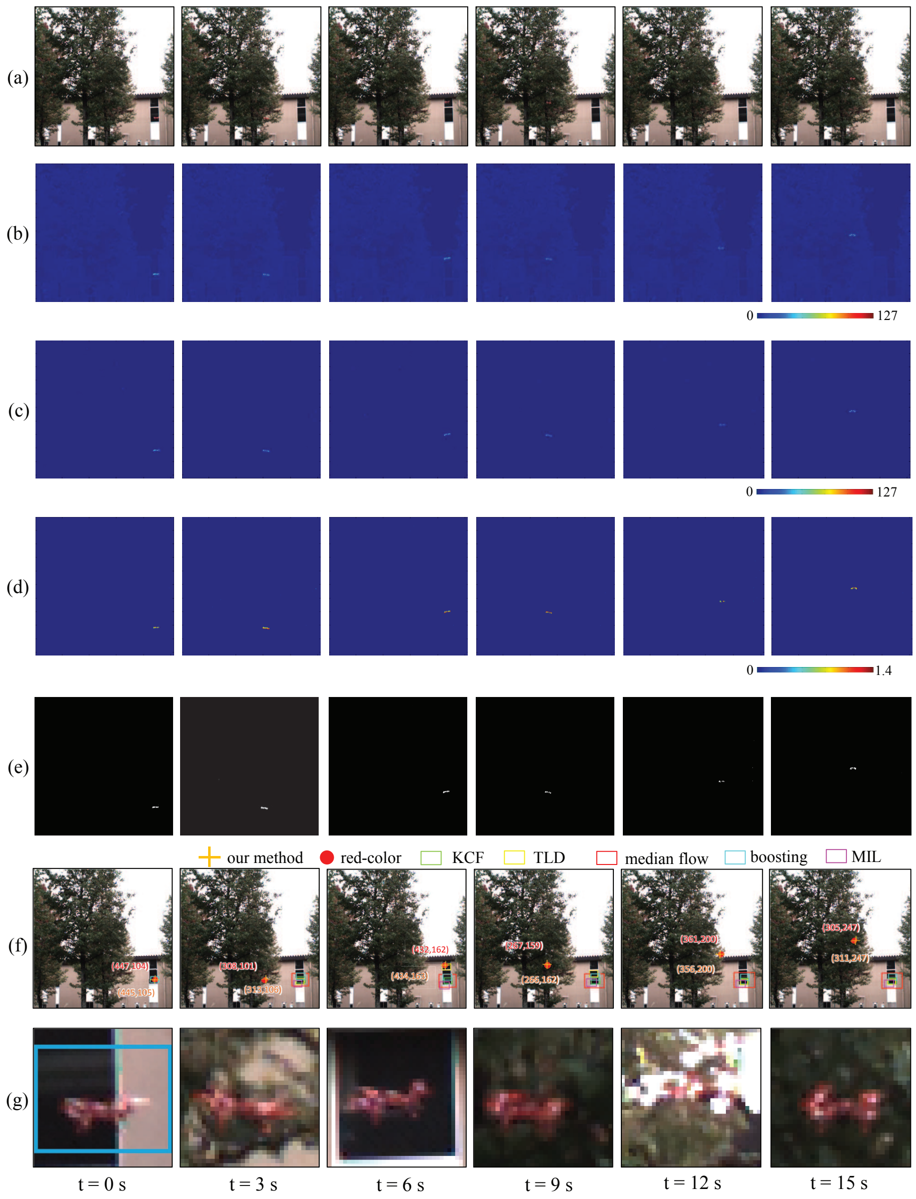
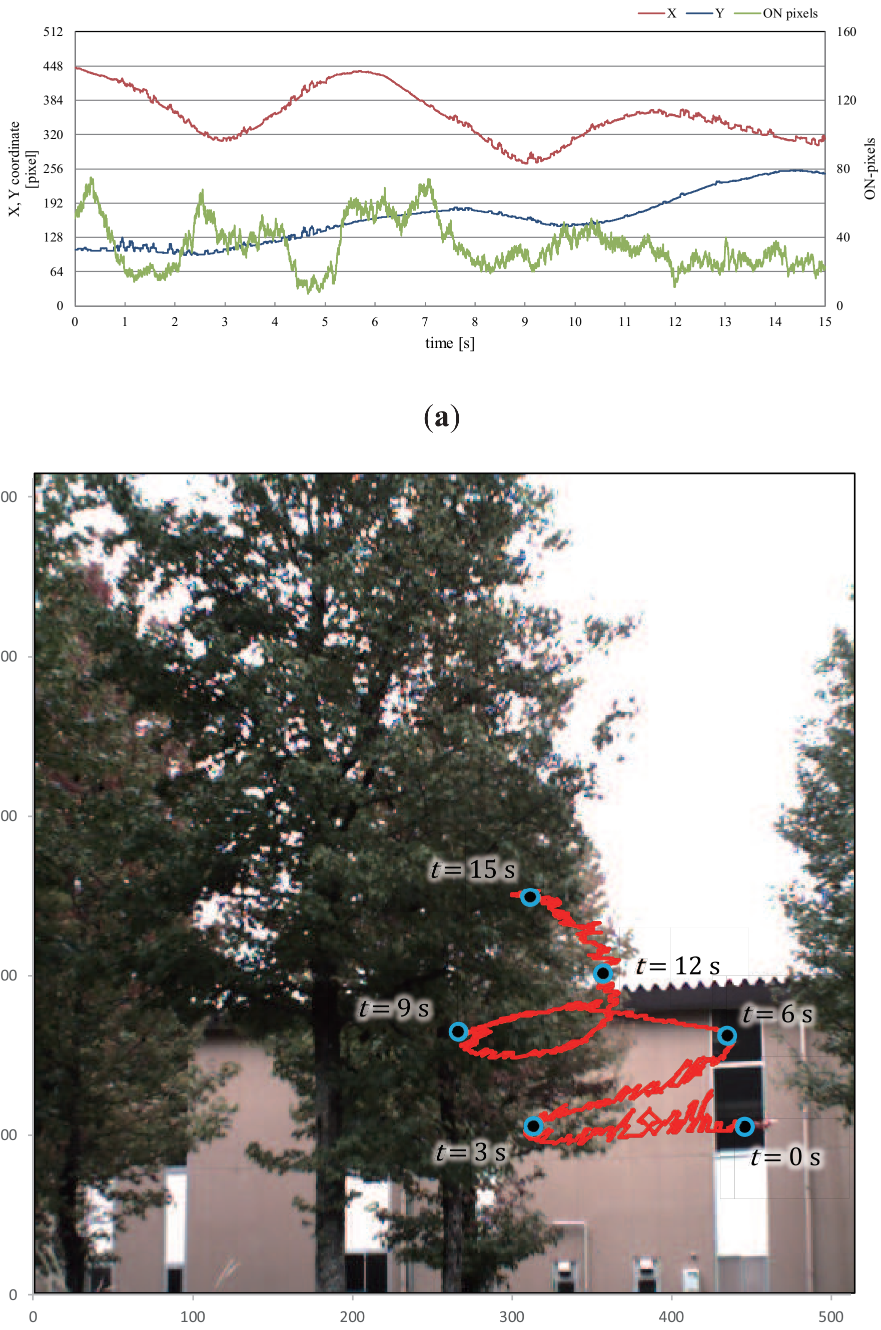


Figure 4.20: (a) Input images; amplitude of (b) input; and (c) pixel-wise filtered images; (d) amplitude ratios; (e) extracted vibration features; (f) tracked positions; (g) magnified images.

It can be seen that certain pixels around the propellers of the multicopter were robustly extracted as vibration features by our algorithm when the background scene just directly behind the multicopter was varying with its flight trajectory (trees at t = 3, 9, 12, and 15 s, and building at t = 0 and 6 s). When t = 0, 3, 6, 9, 12, and 15 s, the averaged positions of the red-color regions in the images, which indicated to the actual locations of the multicopter, were (447,104), (308,101), (432,162), (267,159), (361,200), and (305,247), respectively, whereas those of the extracted pixels were (445,105), (313,106), (434,163), (266,162), (356,200), and (312,247), respectively. Due to the partial occlusion of the propellers by the multicopter itself, the averaged positions of the extracted pixels slightly deviated from the actual locations of the multicopter, however, they almost corresponded with the actual locations of the multicopter and the ROI regions illustrated in Figure 4.20f wholly or partially involved the regions of the multicopter. In Figure 4.20g, it can be seen that the tracking windows largely deviated from the target multicopter and mistracked cluttered background scenes in all the single-object tracking methods. This is because the object to be tracked was determined with a subimage in the low-resolution 32 × 24 region, and there were many unstructured patterns with similar appearance-based features in the background scenes.



(a)

(b)

Figure 4.21: *xy* trajectory of extracted vibration region in “trees-and-building background” experiment. (a) *x*- and *y*-coordinates and number of pixels; (b) *xy* trajectory.

Figure 4.21a illustrates graphs that show changes in the *x*- and *y*-coordinate values of the averaged positions of the extracted pixels and the number of the extracted pixels for 15 s, and the xy trajectory for 15 s was plotted over the input image of 512 × 512 pixels captured at $t = 0$ in Figure 4.21b. Whereas the number of the extracted pixels was not so large and varied in the range of 7 to 75, we have confirmed that the xy trajectory of the averaged positions of the extracted pixels were robustly extracted in correspondence with the left-and-right motion and elevation of the flying multicopter when the background scene directly behind the multicopter was frequently switched to trees in the center and a building in the right side. Here we can observe certain fluctuations in the xy trajectory due to the partial occlusion of the propellers. This is because our method only extracted the regions of the propellers, by excluding the body of the multicopter, and the average positions of the extracted pixels were discretely changed within the region of the multicopter when one propeller was unobservable with occlusion.

4.4.2 Walking-Persons Background

We analyzed the 1000-fps video when the multicopter moves against a background with moving obstacles; the multicopter flew repeatedly in the right and left directions at different heights in front of many persons with quick arm movements, who were walking at a distance of approximately 6 m in front of the camera. The focal length, focus distance, and aperture of the lens were set to 12 mm, 8 m, and F2.8, respectively. The measurement areas of 512 × 512 pixels were 4.7 × 4.7 m, where one pixel corresponds to 9.2 mm^2 at a distance of 6 m.

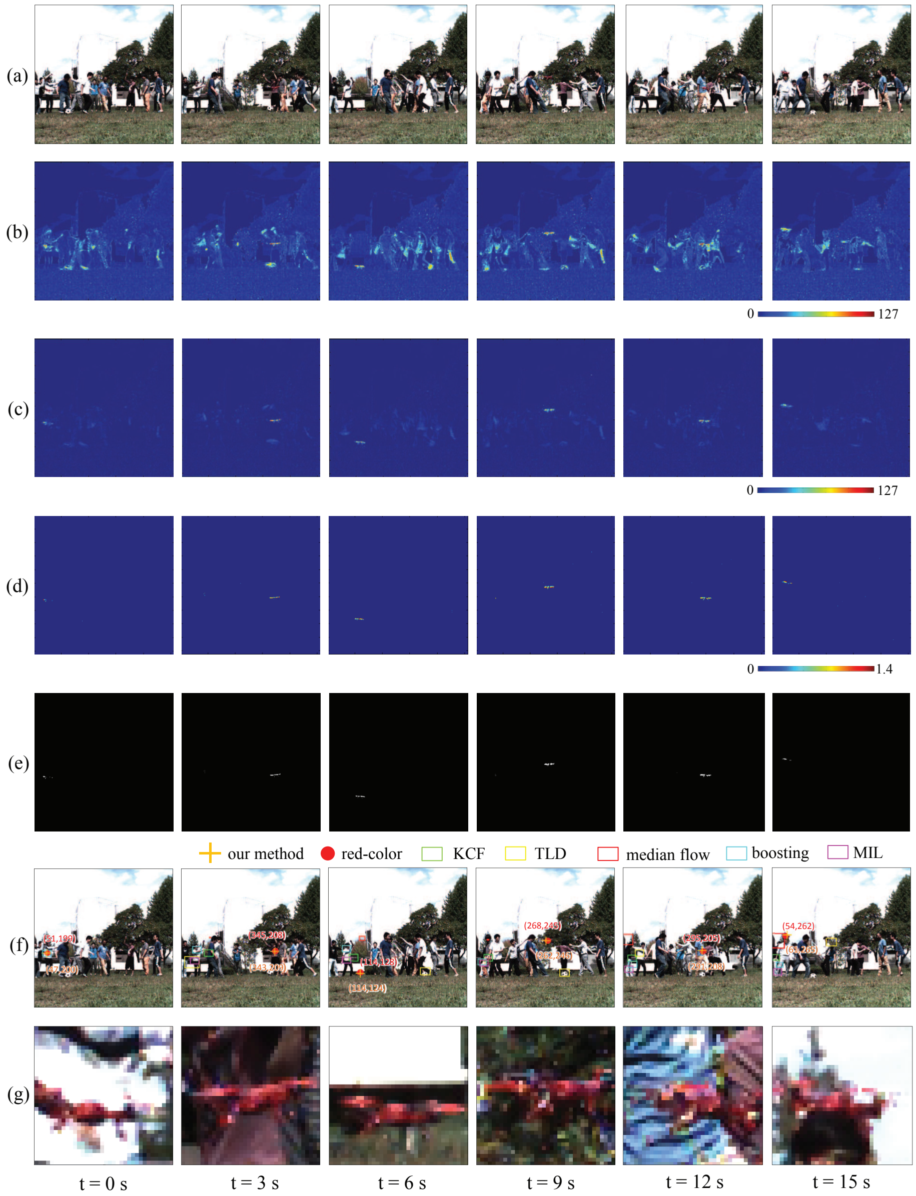


Figure 4.22: (a) Input images; amplitude of (b) input; and (c) pixel-wise filtered images; (d) amplitude ratios; (e) extracted vibration features; (f) tracked positions; (g) magnified images.

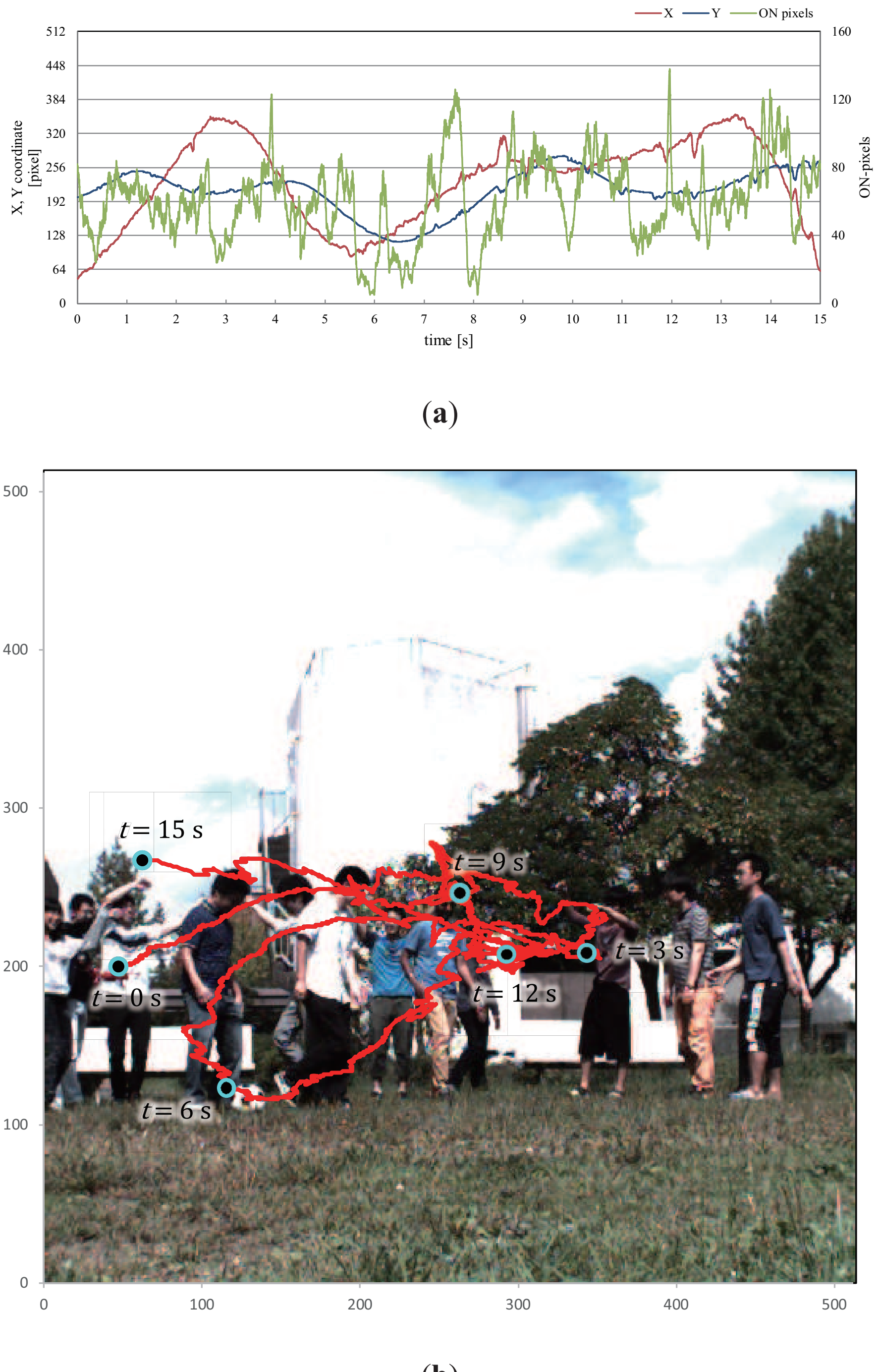


Figure 4.23: xy trajectory of extracted vibration region in "trees-and-building background" experiment. (a) x- and y-coordinates and number of pixels; (b) xy trajectory.

Figure 4.22a–d shows the input images, the moving average distributions of the amplitude of the input images, pixel-wise filtered images, and the ratio distribution of the two amplitudes for $t = 0$–15 s. Figure 4.22e,f show the vibration regions extracted by our algorithm, and magnified images of 32 × 32 pixels around the extracted pixels, respectively. Figure 4.22g shows the averaged positions of the extracted pixels, those of the red-color regions, and the tracking results of the single object tracking methods used in the previous subsection, in which the object to be tracked was initially defined as a 32 × 24 subimage at $t = 0$ s as illustrated in Figure 4.22f. When the multicopter flew repeatedly in the right and left directions at different heights in front of many walking persons, our algorithm extracted certain pixels around the propellers of the multicopter as vibration features without being disturbed by their quick movements. When $t = 0$, 3, 6, 9, 12, and 15 s, the averaged positions of the red-color regions in the images were (51,199), (345,208), (114,128), (263,245), (295,205), and (54,262), respectively, and those of the extracted pixels, (47,200), (343,209), (114,124), (262,246), (291,208), and (61,265), respectively, had slight deviations from them due to the partial occlusion of the propellers, however, the ROI regions illustrated in Figure 4.22f involved the regions of the multicopter at all times. Figure 4.22g shows that the tracking windows with the single-object tracking methods, which were used in the previous subsection, largely deviated from the target mulitcopter, and these appearance-based tracking methods are almost unable to track in this scenario.

Figure 4.23a,b illustrate graphs that show changes in the x- and y-coordinate values of the averaged positions of the extracted pixels and the number of extracted pixels for 15 s, and the xy trajectory for 15 s was plotted over the input image at $t = 0$. Corresponding to the left-and-right motion of the flying multicopter at different heights, the xy trajectory of the averaged positions of the extracted pixels were robustly extracted without any disturbance by the moving background, including the fluctuation due to the partial occlusion of the propellers, whereas the number of extracted pixels largely varied in the range of 5 to 138.

4.5 Concluding Remarks

In this chapter, we analyzed the detectability of a vibration source localization method based on pixel-level digital filters applied to HFR video for rotating fans and a flying multicopter with rotating propellers under various imaging conditions, whose rotational frequencies were distinctly distant from those of the background scenes. The robustness of the method under brightness changes, defocus blur, apparent scale and pose variations, rotational frequency change, complex background and motion blur was demonstrated using several 2000 fps or 300 fps videos of rotating fans captured by adjusting the lens parameters, the shooting angle, and the rotation of the fan or by moving the fan and background pattern. The robustness of images that were simultaneously affected by multiple appearance changes was also demonstrated using a flying multicopter in various outside scenarios.

Chapter 5

Real-Time Vibration Source Tracking Using High-Speed Vision

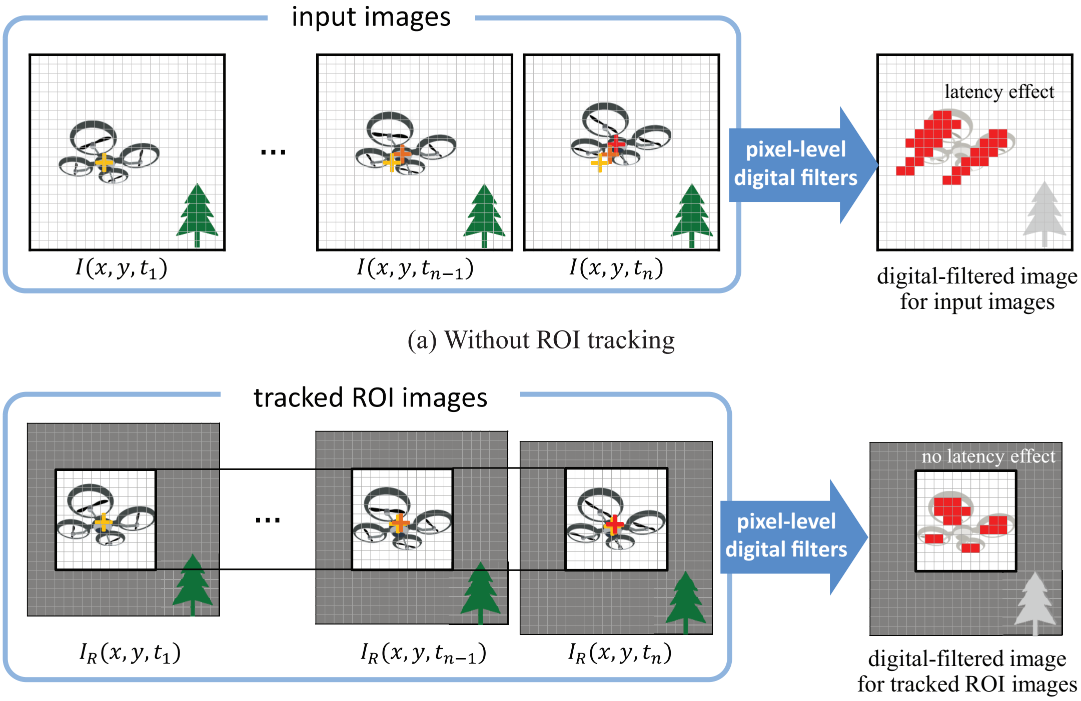
5.1 Introduction

In this chapter, by applying pixel-level digital filters to clipped region-of-interest (ROI) images, in which the center position of a vibrating object is tracked at a fixed position, we reduce the latency effect on a digital filter, which may degrade the localization accuracy in vibration source tracking. Pixel-level digital filters for 128×128 ROI images, which are tracked from 512×512 input images, are implemented on a 1000-fps vision platform that can measure vibration distributions at 100 Hz or higher. Our tracking system allows a vibrating object to be tracked in real time at the center of the camera view by controlling a pan-tilt active vision system.

We present several experimental tracking results using objects vibrating at high frequencies, which cannot be observed by standard video cameras or the naked human eye, including a flying quadcopter with rotating propellers, and demonstrate its performance in vibration source localization with sub-degree-level angular directivity, which is more acute than a few or more degrees of directivity in acoustic-based source localization.

5.2 Latency Effect in Digital Filter

In order to stably extract vibration-based features, it is important to consider the latency effect on a digital filter, on the basis of image data accumulated at the previous



(a) Without ROI tracking

(b) With ROI tracking

Figure 5.1: Pixel-level digital filters for tracked ROI images.

frames; the digital filter requires multiple images captured at frames during several times of the cycle time of the vibration source in order to detect repetitive brightness changes vibrating at its frequency. Thus, vibration-based features are not always suitable for target tracking, because the digital filter latency effect may adversely affect the performance of target tracking with visual feedback control. To reduce this latency effect in digital filters, we introduce the following two ideas in this paper: (a) a digital filter for tracked ROI images, and (b) single-cycle time features.

In general, the latency effect in digital filters becomes larger as a vibrating object moves more quickly; it is determined by the translation displacement during several vibration cycle times. When the translational speed of the vibrating object is not negligible, several pixels are incorrectly detected as ghost vibration region; periodic changes in image intensities can be observed during several vibration cycle times even when the object does not still locate at these pixels in the current frame, because the object coordinate system moves against the image coordinate system. To cancel such translational displacements in original input images, we use ROI images in which a vibrating object is always tracked

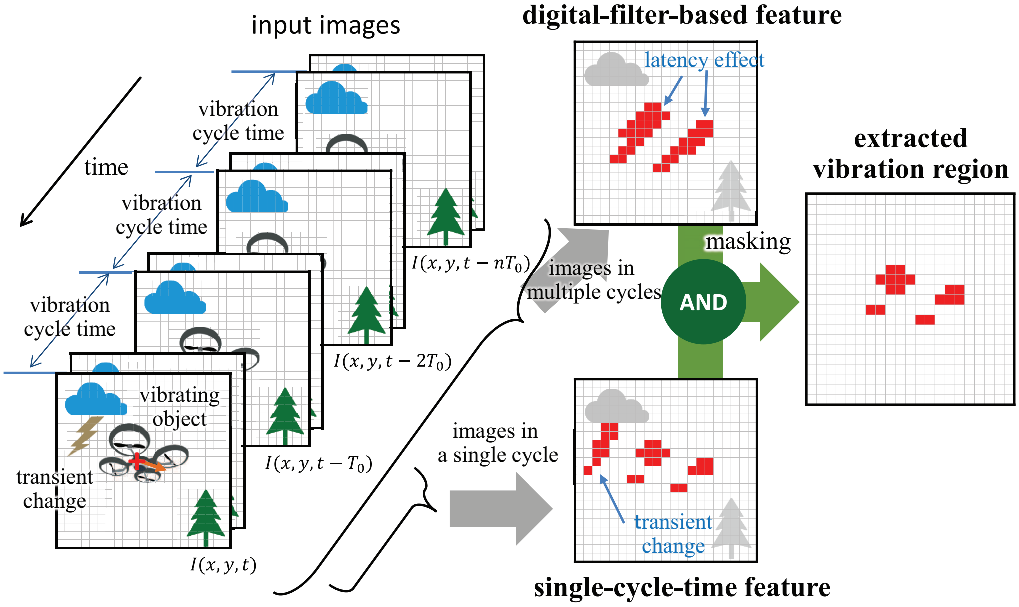


Figure 5.2: Masking process with single-cycle time feature for lowering latency effect in digital filter.

as the image center for the pixel-level digital filter. The effectiveness of the ROI tracking process in a digital filter is illustrated in Fig. 5.1. The latency effect in the digital filter is significantly reduced, because the vibrating object is virtually fixed without translational movement in the ROI images, and its vibration in image brightness is always observed at pixels around the image centers; the ROI-image coordinate system corresponds to the object coordinate system.

To determine the ROI location for tracking a vibrating object, simultaneous calculation of the image features that indicate the object's position in images, such as moment feature-based image centroids, is required. The latency effect still remains when computing vibration-based features for vibration source localization with digital filtered images. To prevent this latency effect spreading over several cycle times of vibration, single-cycle time features were introduced in this study; these features concern only the temporal brightness change in a single cycle time of vibration, and the latencies in computing them may be small within a single cycle time. However, they cannot judge whether there are repetitive brightness changes according to vibration source, or one-time transient brightness changes when the brightness edges of moving scenes pass over pixels during a single cycle time of vibration. Digital filter-based and single-cycle time features have their merits and demerits in terms of latency and accuracy in the target tracking of a vibrating object. In order to suppress the latency effect in digital filters, we introduce an additional process for masking image regions extracted using pixel-level digital filters with single-cycle time features. Fig. 5.2 shows the differences between digital filter-based features and single-cycle time features, and the masking process for decreasing latency effect.

5.3 Tracking Processes

We introduce a vibration-based feature corresponding to the position of a vibrating object for target tracking, which is calculated by using pixel-level digital filters via a masking process by single-cycle time feature extraction for tracked ROI images, in which the vibrating object is always located in their center. We assume that the properties of a vibrating object are initially given, such as its center frequency f_0, and the size and pixel interval of the ROI image are always fixed in the ROI tracking process. The proposed algorithm involves two processes: (a) searching, and (b) ROI tracking. The searching process is executed for detecting a vibrating object in the entire image region; the tracking process is executed for selecting ROI images by assuming HFR vision, in which the translation displacement between frames is small and the tracked ROI image in the current frame is located only in the neighborhood of that in the previous frame. Here, the input image of $N \times N$ pixels and the tracked ROI image of $N' \times N'$ pixels are captured at time t as $I(x, y, t)$ and $I_R(x', y', t)$, respectively; (x, y) and (x', y') are their pixel coordinates.

5.3.1 ROI Tracking Process

5.3.1.1 Selection of ROI image

The ROI image $I_R(x', y', t)$ at time t is selected from the input image $I(x, y, t)$ as

$$I_R(x', y', t) = I(x' + o_x(t), y' + o_y(t), t), \tag{5.1}$$

where the start pixel coordinate $(o_x(t), o_y(t))$ is determined by the location of vibration source $(c_x(t-\delta t), c_y(t-\delta t))$, which is estimated at the previous frame $t-\delta t$ in the searching process or ROI tracking process as

$$(o_x(t), o_y(t)) = \left(c_x(t-\delta t) - \frac{N'}{2}, c_y(t-\delta t) - \frac{N'}{2}\right). \tag{5.2}$$

δt is the frame cycle time of the vision system.

5.3.1.2 Pixel-level digital filters

The ROI image $I_R(x', y', t)$ is filtered at every pixel with a band-pass filter, the center frequency of which is f_0. In this study, the following infinite impulse response (IIR) filter is adopted as a band-pass filter:

$$g(x', y', t) = \sum_{s=0}^{p-1} b_s I_R(x', y', t-s\delta t) - \sum_{s=1}^{p-1} a_s g(x', y', t-s\delta t), \tag{5.3}$$

where p is the filter order and a_s, b_s are the tap coefficients. These parameters determine the center frequency and bandwidth of the band-pass filter.

5.3.1.3 Moving averages of filtered image intensities

The average of the absolute value of the filtered image in a certain interval ΔT_f and that of the brightness of the ROI image are calculated at every pixel as

$$G(x', y', t) = \frac{1}{\Delta T_f} \int_{t-\Delta T_f}^{t} |g(x', y', t)| dt, \tag{5.4}$$

$$K(x', y', t) = \frac{1}{\Delta T_f} \int_{t-\Delta T_f}^{t} I_R(x', y', t) dt, \tag{5.5}$$

where ΔT_f is set to several times of the cycle time $1/f_0$ to extract the vibration region.

5.3.1.4 Vibration region extraction

The vibration region $V(x', y', t)$ is extracted by thresholding the ratio of $G(x', y', t)$ to $K(x', y', t)$ with a threshold θ_2 as

$$V(x', y', t) = \begin{cases} 1 & \left(K(x', y', t) > \theta_1 \text{ and } \dfrac{G(x', y', t)}{K(x', y', t)} > \theta_2 \right) \\ 0 & \text{(otherwise)} \end{cases}, \tag{5.6}$$

where the pixel (x', y') is judged to be an ambiguous pixel not to be extracted when $K(x', y', t)$ is lower than a threshold θ_1.

5.3.1.5 Single-cycle time feature extraction

By inspecting the difference between the maximum and minimum brightnesses of the ROI images during one recent cycle time of vibration at every pixel, time-varying brightness pixels are extracted as a single-cycle time feature:

$$D(x', y', t) = \begin{cases} 1 & (H_{\max}(x', y', t) - H_{\min}(x', y', t) > \theta_3) \\ 0 & \text{(otherwise)} \end{cases}, \tag{5.7}$$

where $H_{\max}(x', y', t)$ and $H_{\min}(x', y', t)$ are defined with a cycle time of vibration $T_0 = 1/f_0$:

$$H_{\max}(x', y', t) = \max\{I_R(x', y', t'), t - T_0 \leq t' \leq t\}, \tag{5.8}$$

$$H_{\min}(x', y', t) = \min\{I_R(x', y', t'), t - T_0 \leq t' \leq t\}. \tag{5.9}$$

The single-cycle time feature is not so sensitive to a slight deviation between the real and assumed cycle times, because the maximum and minimum brightness values do not vary significantly during the cycle time when the frequency of the vibrating object is slightly different from its assumed frequency.

5.3.1.6 Masking by single-cycle time feature

The vibration region $V(x', y', t)$ is masked with the single-cycle time feature $D(x', y', t)$ to reduce the latency effect on a digital filter:

$$F(x', y', t) = V(x', y', t) \cap D(x', y', t), \tag{5.10}$$

where "∩" means logical AND operation.

5.3.1.7 Localization of vibration source

To localize a vibrating object in the ROI image, the image centroid $(c_x(t), c_y(t))$ of $F(x', y', t)$ is calculated by using its 0th and 1st moment features, $M_0(t)$ and $M_x(t)$, $M_y(t)$:

$$(c_x(t), c_y(t)) = \left(\frac{M_x(t)}{M_0(t)} + o_x(t), \frac{M_x(t)}{M_y(t)} + o_y(t) \right), \tag{5.11}$$

$$M_0(t) = \sum_{x',y'} F(x', y', t), \tag{5.12}$$

$$M_x(t) = \sum_{x',y'} x' F(x', y', t), \quad M_y(t) = \sum_{x',y'} y' F(x', y', t). \tag{5.13}$$

When the number of one-pixels in $F(x', y', t)$, which corresponds to $M_0(t)$, is smaller than a threshold θ_4, we consider the vibration source localization failed or unstable and change to the searching process for the entire image region; otherwise, we go back to 1) and continue the ROI tracking process at the next frame.

5.3.2 Searching Process

In our proposed algorithm, we switch to a searching process for detecting a vibrating object in the entire image region when the algorithm starts initially or vibration source localization has failed in the ROI tracking process. In the searching process, gray-level images of $M \times M$ pixels, $I_S(x'', y'', t)$, which are obtained from the input image of $N \times N$ pixels at intervals of m pixels ($N = mM$), are processed,

$$I_S(x'', y'', t) = I(mx'', my'', t), \tag{5.14}$$

where (x'', y'') is the pixel coordinate of $I_S(x'', y'', t)$.

The searching process is similar to subprocesses 2), 3), and 4) in the ROI tracking process; $V_S(x'', y'', t)$ is extracted as the vibration region in $I_S(x'', y'', t)$ with the same parameters

as those used in the ROI tracking process. To localize a vibrating object in the entire image region, the image centroid $(c_x(t), c_y(t))$ in the xy coordinate is calculated using the moment features of $V_S(x'', y'', t)$:

$$(c_x(t), c_y(t)) = \left(m\frac{M'_x(t)}{M'_0(t)}, m\frac{M'_x(t)}{M'_y(t)} \right), \tag{5.15}$$

$$M'_0(t) = \sum_{x'',y''} V_S(x'', y'', t), \tag{5.16}$$

$$M'_x(t) = \sum_{x'',y''} x'' V_S(x'', y'', t), \ M'_y(t) = \sum_{x'',y''} y'' V_S(x'', y'', t). \tag{5.17}$$

The searching process is iteratively executed until the number of ON-pixels in $V_S(x'', y'', t)$ is greater than a threshold θ_5; otherwise, we revert to the ROI tracking process again at the next frame. The centroid of $V_S(x'', y'', t)$ is given as the initial parameter of the ROI tracking process.

5.4 System Configuration

To show the effectiveness of vibration source localization using pixel-level digital filters, we developed a prototype system that can simultaneously track a vibrating object in the center of a camera view by implementing the vibration source localization algorithm described in Section **??** in a high-speed target tracking system, whereby real-time image processing at 1000 fps or more can be executed. The prototype system consists of a high-speed vision platform, IDP Express [10], and a pan-tilt active vision system. Fig. 5.3 shows the configuration of the target tracking system for vibrating objects.

IDP Express consists of a compact camera head, the dimensions and weight of which are 23×23×77 mm and 145 g, respectively, when no lens is mounted, a dedicated FPGA board (IDP Express board), and a personal computer (PC). The camera head

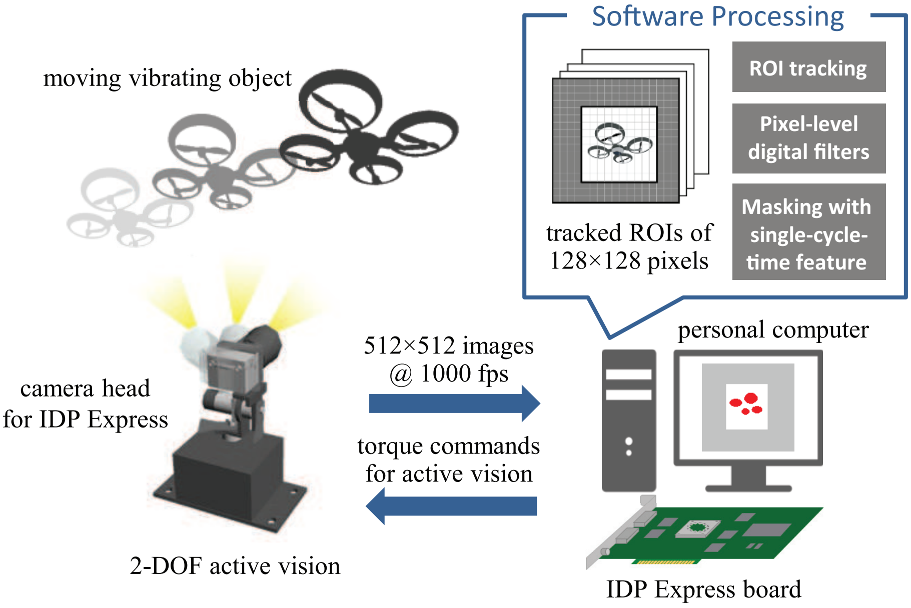

Software Processing
ROI tracking
Pixel-level digital filters
Masking with single-cycle-time feature
tracked ROIs of 128×128 pixels
moving vibrating object
camera head for IDP Express
512×512 images @ 1000 fps
torque commands for active vision
personal computer
2-DOF active vision
IDP Express board

Figure 5.3: System configuration of target tracking system.

can capture and transfer 8-bit gray-level 512×512 images to the IDP Express board at 2000 fps. The IDP Express board has an FPGA for hardware implementation of the user-specific algorithms and can transfer 512×512 input images and their processed results to the standard PC memory at 2000 fps. We used a PC with the following specifications: ASUSTeK SABERTOOTH X79 mainboard, Intel Core i7-4820K @ 3.70 GHz CPU, 8 GB memory, and two 16-lane PCI-e 3.0 buses. On the PC, various API functions that facilitate camera head control and memory-mapped data access can be used on Windows 7 (64-bit) for application software development.

The 2 degrees of freedom (DOF) active vision system is moved by pan and tilt motors, which are compact and high-speed motors. The size of the active vision system is 12×12×7 cm without a camera head. To track a moving object, the camera head mounted on the active vision system is controlled with visual feedback so that the center of its camera view coincides with the image centroid of the vibration region, which is obtained using the vibration source localization algorithm.

The vibration source localization algorithm is implemented by software on the IDP Express and executed in real time on the PC at 1000 fps for 8-bit gray-level images of 128×128 pixels ($N' = M = 128$), which are selected from the input images of 512×512

Table 5.1: Execution times of ROI tracking process

	time [ms]
1) Image acquisition and ROI image selection	0.02
2) Pixel-level digital filters	0.47
3) Moving averages of filtered image intensities	0.09
4) Vibration region extraction	0.06
5) Single-cycle time feature extraction	0.01
6) Masking by single-cycle time feature	0.01
7) Localization of vibration source	0.01
Total 1)–7)	0.66

pixels ($N = 512$). The processing time of the algorithm depends on the filter order p, the interval for moving averages ΔT_f, and the duration time for single-cycle time feature extraction T_0. Table 5.1 summarizes the execution times of the ROI tracking process. The total execution time of the searching process is 0.83 ms, corresponding to subprocesses 1)–4) and 7) in the ROI tracking process; 128×128 images were used in both the searching and ROI tracking processes in this study. Thus, the total execution times in both modes are less than 1 ms, and we confirmed that the pan and tilt motors of the active vision system are controlled by using the image centroid at a visual feedback rate of 1000 fps.

5.5 Experiments

5.5.1 Vibration Source Localization at a Fixed Camera Position

To verify the performance of our system at a fixed camera position without tracking control of the active vision system, we present the experimental results for a rotating fan moved by a linear slider. Fig. 5.4 shows the experimental setup. In the experiment, a 13-cm-diameter fan with three blades was installed on a linear slider 125 cm in front of the camera head. To verify the position of its rotation center, a red marker was attached at the center. A textured pattern of 25×25 cm size, which comprised shades of green, was set behind the fan as the moving background; it was moved together with the fan by the linear slider; the moving background was used to verify that the ROI tracking process can perfectly suppress the noises of the moving background and reduce the latency effect in digital filter. A textured wallpaper comprising shades of green was set behind the

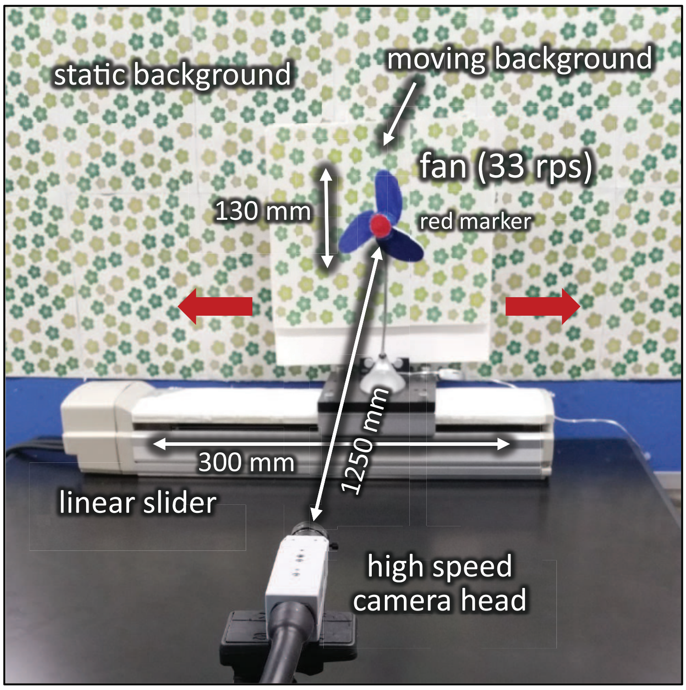


Figure 5.4: Experimental setup for fixed camera position experiment.

above-mentioned textured pattern and fan as the static background; the static background was used to check whether the searching process correctly finds a vibrating object to be localized in 128×128 images, which are obtained from the entire 512×512 input image at intervals of 4 pixels. The measurement area was 1200×1200 mm for 512×512 pixels at a distance of 125 cm in front of the camera head, where one pixel corresponds to 2.0 mm. The tap coefficients a_s, b_s of the pixel-level digital filters were set so that they operated as the band-bass Butterworth filters, the center frequency of which was $f_0 = 100$ Hz and half width was 10 Hz. The parameters were set to $p = 4$, $\Delta T_f = 40$ ms, and $T_0 = 1/f_0 =$ 10 ms. The thresholds θ_1 and θ_2 for vibration region extraction were set to 48 and 0.3, respectively. The threshold θ_3 for single-cycle time feature extraction was set to 40. The thresholds θ_4 and θ_5 for vibration source detection were set to 256 and 16, respectively.

First, we checked the detectable frequency in the searching process in the experiment for a rotating fan at a fixed location with no slider motion. The rotation frequency of this fan time-varied in a range of 17 to 44 rps. Fig. 5.5 shows the 512×512 input images with different rotation speeds, and the extracted vibration region of 128×128 pixels in the searching process. Fig. 5.6 shows the number of extracted pixels in the searching

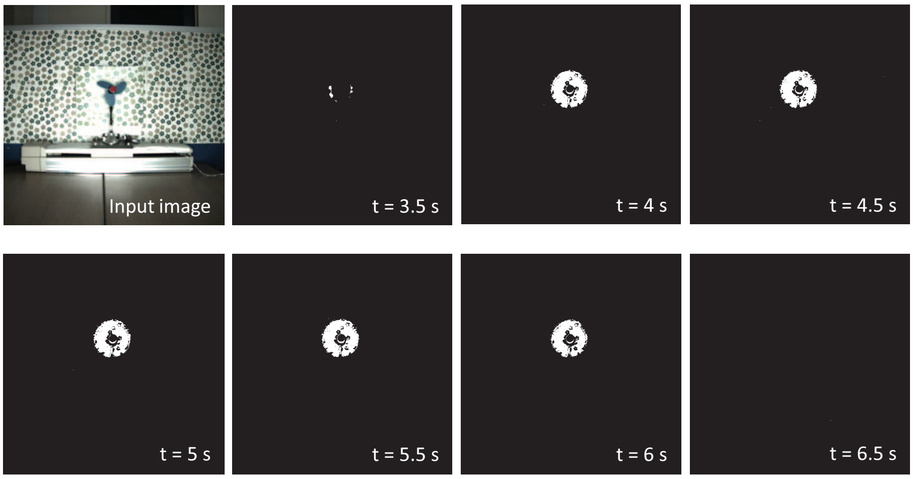


Figure 5.5: Input images of 512×512 pixels and extracted vibration regions of 128×128 pixels in the searching process for different rotation speeds.

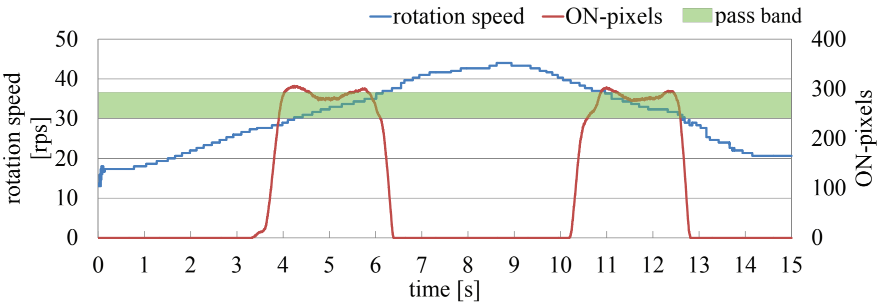
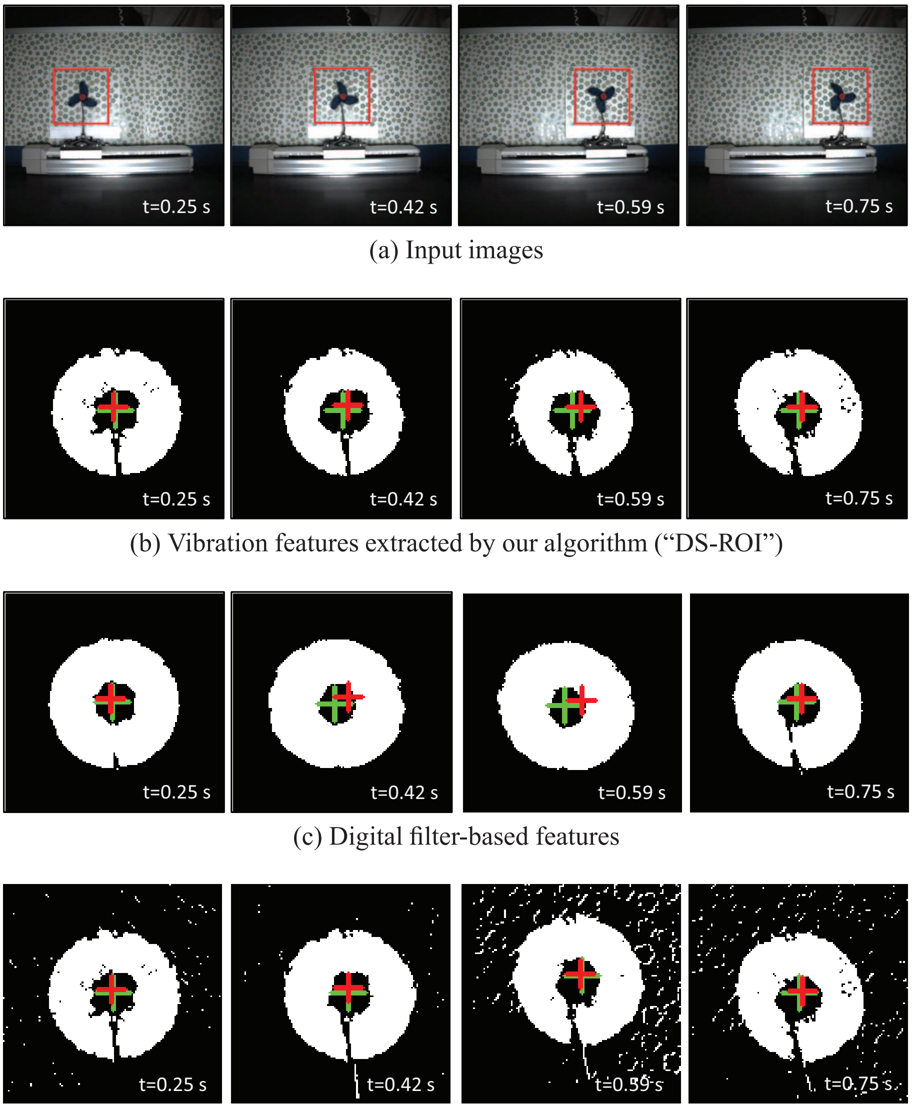


Figure 5.6: Number of extracted pixels in the searching process for different rotation speeds.

process for different rotation speeds. When the rotation speed was in a range of 29 rps to 37 rps, our method could distinctly extract the pixels around the fan as vibration pixels. The brightness values at the extracted pixels varied in a frequency range of 87 to 111 Hz. This corresponds to the parameters of the band-pass filters used in the experiment, whose center frequency was 100 Hz and half width was 10 Hz.

Next, we obtained the verification results when the linear slider moved alternately in the right and left directions with an amplitude of 30 cm and at a cycle time of approximately 1 s; the fan rotated at a constant speed of 33 rps. Fig. 5.7(a) shows the input images of 512×512 pixels, taken at intervals of 0.17 s. We confirmed that the searching process worked correctly and switched to the ROI tracking process to track the rotating fan. The start time of observation in the ROI tracking process was $t = 0$; the experimental



(a) Input images

(b) Vibration features extracted by our algorithm ("DS-ROI")

(c) Digital filter-based features

(d) Single-cycle time features

Figure 5.7: Input images of 512×512 pixels and extracted vibration features in the tracked ROI regions of 128×128 pixels at a fixed camera position.

data were discussed only for the ROI tracking process here. Fig. 5.8 shows the changes in the image intensities for t = 0.2~0.6 s at points A(256, 100), B(256, 155), and C(256, 186) when the fan was passing over the vertical line of x = 256. Corresponding to three times of the rotational frequency of the fan, periodic changes at a frequency of 100 Hz were observed at point C locating around the fan; small changes were observed at point A on the static background and low-frequency changes were observed at point B on the moving background. Using our algorithm, the image region around a vibrating object can be

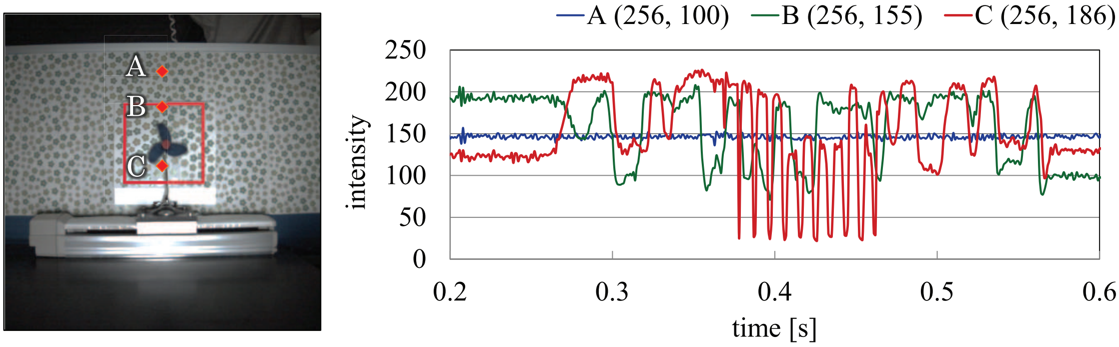
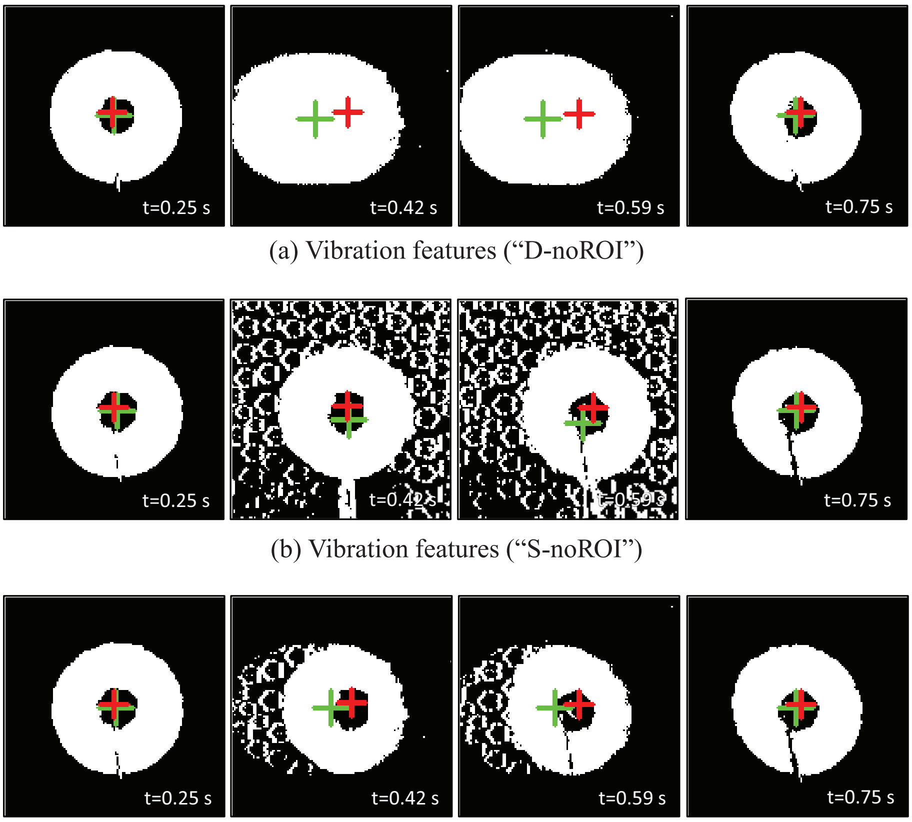


Figure 5.8: Temporal changes in image intensities at a fixed camera position.

extracted by detecting these changes in image intensities with pixel-level digital filters.

Fig. 5.7(b) shows the digital filter-based features masked with the single-cycle time features in the ROI regions of 128×128 pixels, which were tracked by our algorithm ("DS-ROI"). In Fig. 5.7(a), the red-line squares indicate the tracked ROI regions. The red "+"s are the marker positions at the center of the fan and the green "+"s are the *xy* centroids of the image features in the tracked ROI regions. It can be seen that the *xy* centroids virtually coincide with the marker positions. Fig. 5.7(c) and (d) shows the digital filter-based features and the single-cycle time features in the red-line square ROI regions, respectively, which were calculated in real-time "DS-ROI" measurement. Owing to the latency effect in the digital filter, Fig 5.7(c) indicates that the difference between the marker position and the centroid of the digital filter-based features became larger when the fan was moved quickly. In Fig. 5.7(d), it can be seen that the single-cycle time features contained several background errors, corresponding to the moving background, which may contribute to large deviations from the actual position of the fan.

To verify the effectiveness of the ROI tracking process in reducing the latency effect in the digital filter, Fig. 5.9 shows the vibration features extracted without ROI tracking, which were clipped in the red-line square ROI regions of 128×128 pixels, as illustrated in Fig. 5.7(a): digital filter-based features without ROI tracking ("D-noROI," (a)), single-cycle time features without ROI tracking ("S-noROI," (b)), and digital filter-based features masked with single-cycle time features without ROI tracking ("DS-noROI," (c)). The vibration features in Fig. 5.9 were calculated offline using the input image sequences



(a) Vibration features (“D-noROI”)

(b) Vibration features (“S-noROI”)

(c) Vibration features (“DS-noROI”)

Figure 5.9: Vibration features in the image region of 128×128 pixels, which were calculated for the fixed whole image region with no ROI tracking.

stored during the real-time “DS-ROI” measurement.

Fig. 5.10 shows the *xy* coordinate values of the marker and the image centroids of the vibration features extracted by “DS-ROI,” “D-noROI,” “S-noROI,” and “DS-noROI” measurements for 2.5 s. Fig. 5.11 shows the deviations of these image centroids from the marker position. It can be seen that the latency effect in the digital filter when the digital filter-based features in Fig. 5.7(c) were calculated for the tracked ROI regions was reduced, as compared with those in Fig. 5.9(a) calculated for the fixed whole image regions. Fig. 5.9(c) indicates that this latency effect with no ROI tracking still remains largely in the digital filter-based features masked with single-cycle time features, because the dilated areas of the digital filter-based features involved several background errors when the fan was moving quickly, which were miss-detected in the single-cycle time features. Figs. 5.10 and 5.11 show that the image centroids in “DS-ROI” measurement most

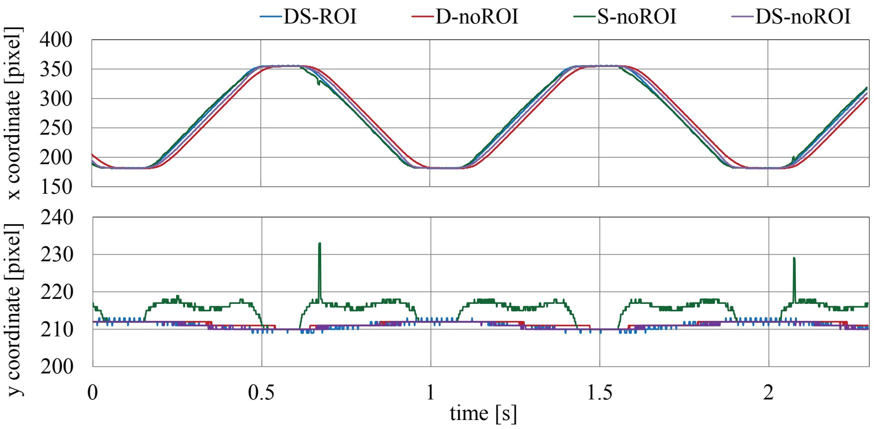


Figure 5.10: Image centroids of vibration features.

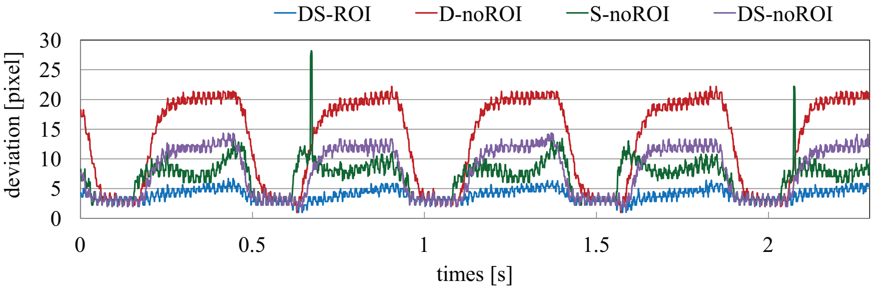


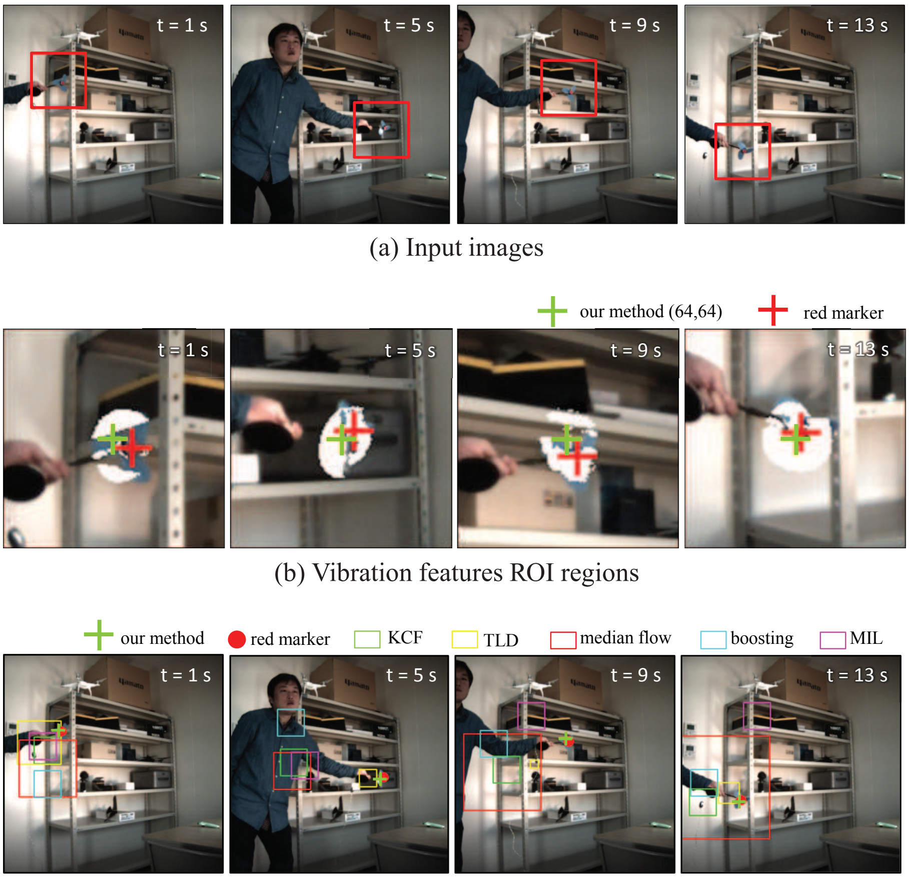
Figure 5.11: Deviation of image centroids of vibration features.

closely matched the marker position of the fan; the image centroids in “D-noROI” measurement widely deviated from the marker position because of the latency effect in the digital filter when the fan moved faster, and those in “S-noROI” measurement had certain deviations in the direction of y, which were caused by the moving background’s pattern. A maximum deviation of 20 pixels was observed in “DS-noROI” measurement and the mean of the deviation in “DS-ROI” measurement was 4.2 pixels, which corresponded to 8.4 mm at a distance of 125 cm from the camera head. Therefore, by introducing the ROI tracking process, the latency effect in a digital filter can be mitigated without compromising the filtering properties for detection of invisible high-speed vibration. Vibration source localization with directive errors of 0.39 degrees was realized in “DS-ROI” measurement.

Next, we verified the robustness of the proposed tracking method using a 33-rps-rotation fan in a non-controlled indoor scenario, by comparing the obtained tracking results with several state-of-the-art vision-based tracking methods. The fan to be tracked was the same as that used in the previous experiment, and all parameters of the proposed algorithm were set to the same values as those provided in the previous experiment. The fan was observed at a fixed camera position without tracking control of the active vision system, and it was alternately moved in the right and left directions with a manual vertical shake at 1 Hz, at a distance of approximately 2.0 m from the camera head in a laboratory scene. The measurement area was 160×160 mm for 512×512 pixels at a distance of 2.0 m from the camera head, where one pixel corresponds to 3.0 mm.

Fig. 5.12 shows (a) the input images of 512×512 images, (b) the vibration features extracted using the proposed algorithm, which were plotted in white over the ROI images of 128×128 pixels with their centroids and the marker positions, and (c) the tracked results obtained using the proposed and several single-object tracking methods, which were taken at intervals of 4 s. We used the following single-object tracking methods: (1) KCF [142], (2) TLD [143], (3) Median Flow [144], (4) Boosting [145], and (5) MIL [146], which were prepared in Open CV tracking API in Open CV 3.0 [147], and illustrated as color-line rectangular regions. The color input images of 512×512 pixels at 1000 fps, which were captured during execution of the proposed method in real time, were processed offline. The object to be tracked was initially defined as a 32×32 sub-image region around the fan at $t = 0$; which was the start time of observation. Fig. 5.13 shows the centroid of the vibration features, the marker position, and the tracked positions obtained using single-object tracking methods for 13 s.

It can be observed that certain pixels around the rotating fan were extracted as the vibration features when the background scene behind the fan was varied with its motion. The ROI regions of 128×128 pixels were correctly tracked for the moving fan, as shown in Fig. 5.12(b). In Fig. 5.12(c) and Fig. 5.13, the centroid of the vibration features matched with the marker position in the proposed method, whereas the tracked windows were considerably deviated from the target fan and background scenes were tracked incorrectly for all single-object tracking methods. This was because there were unstructured patterns with similar appearance-based features in the background scenes. Table 5.2 summarizes the means of the deviations from the marker position in all the methods; they were com-



(a) Input images

(b) Vibration features ROI regions

(c) Tracked positions

Figure 5.12: Input images of 512×512 pixels, masked vibration features in the tracked ROI regions of 128×128 pixels, and tracked results for single-object tracking methods in indoor scene experiment.

puted as the averaged deviations for 13 s. The proposed method achieved vibration source localization with directive errors of 0.7 degrees in the uncluttered indoor scene, in which appearance-based single-tracking methods did not work correctly.

5.5.2 Target Tracking Experiment

Next, we present the experimental results for real-time tracking when the camera head was mechanically controlled by the 2-DOF active vision system such that the vibrating object was located at the center of the camera view. Fig. 5.14 shows the experimental setup. We tracked a fan rotating at 33 rps as a vibrating object, which was the same as that described in Subsection 5.5.1. The fan was alternately moved in the right and left di-

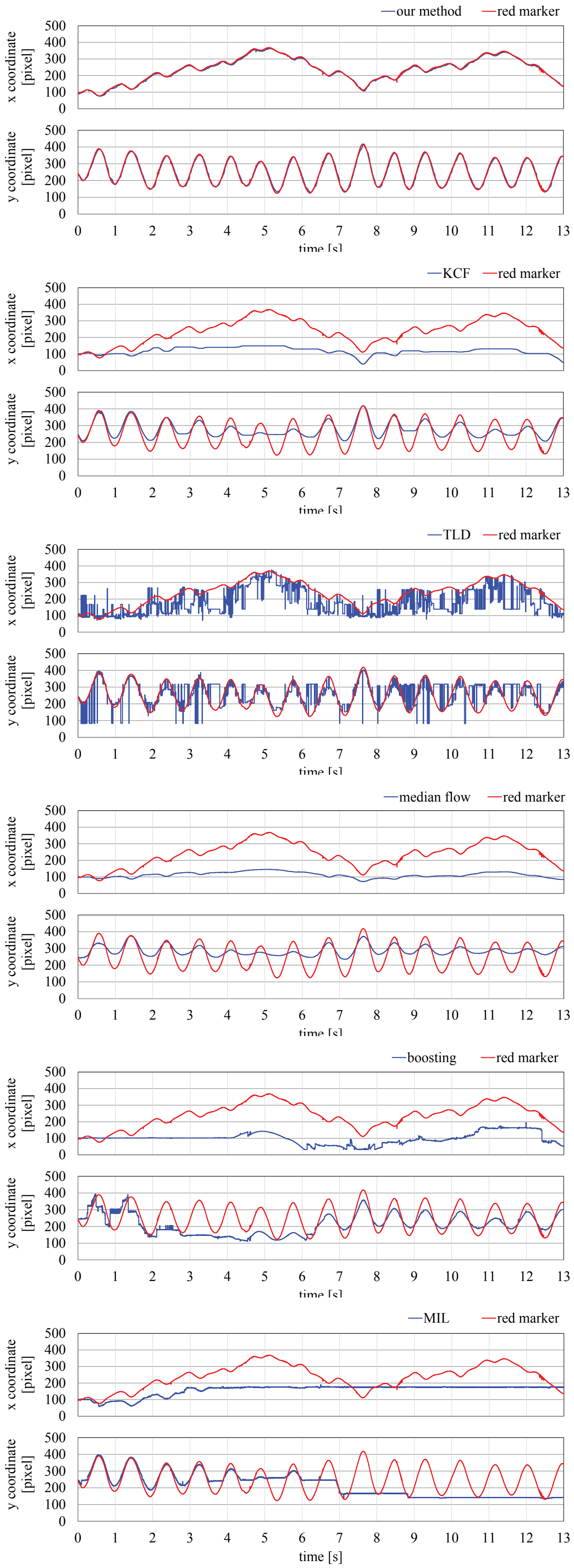


Figure 5.13: Image centroids of vibration features and tracked position for single-object tracking methods in indoor scene experiment.

Table 5.2: Averaged deviations of image centroids of vibration features and tracked positions for single-object tracking methods.

vibration features	KCF	TLD	median flow	boosting	MIL
8.6	127.7	85.6	139.9	157.3	119.5

(unit: pixel)

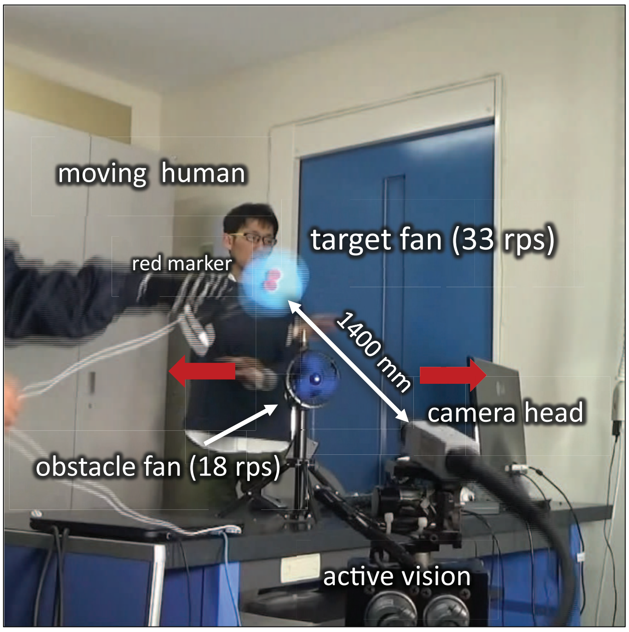
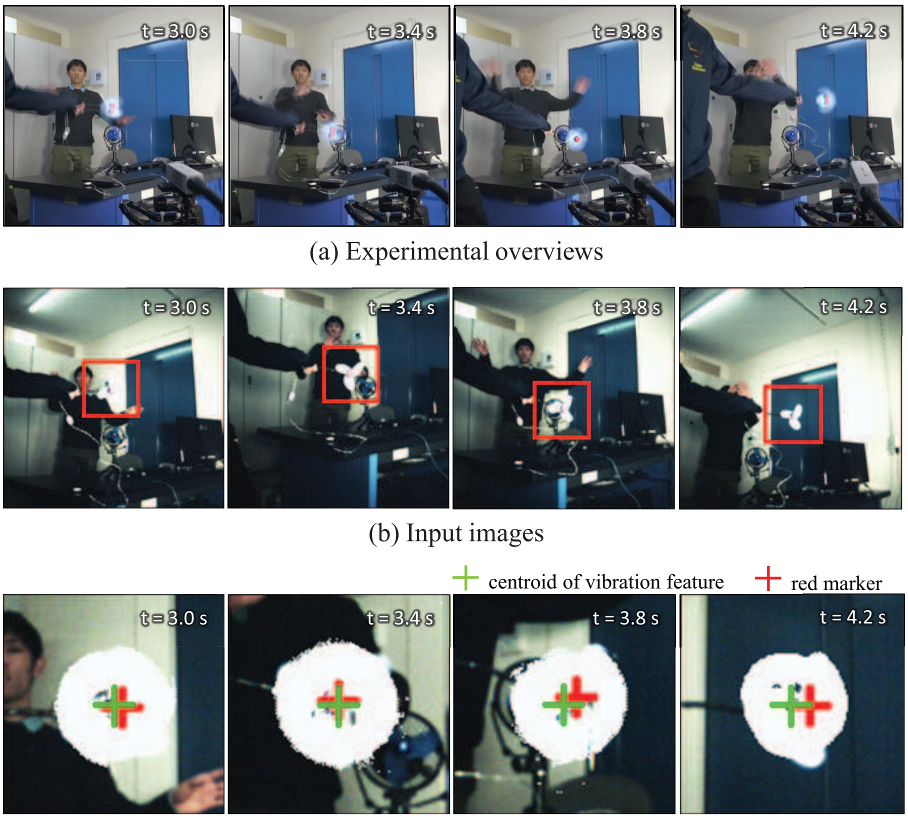


Figure 5.14: Experimental setup for mechanical tracking experiment.

rections, with a manual vertical shake at 2 Hz, at a distance of approximately 1.4 m from the active vision system in a laboratory scene. The measurement area was 500×500 mm for 512×512 pixels at a distance of 1.4 m in front of the camera head, where one pixel corresponds to 1.5 mm. As obstacles to tracking, a human with moving arms and a fan rotating at 18 rps were set in the background. The static laboratory scene may act as dynamically changing disturbances in the camera view, according to the egomotion of the camera head. All the parameters of our algorithm were set to the same values as those provided in Subsection 5.5.1. The control target for the centroid of the vibration features was set to the center of an input image of 512×512 pixels, (256, 256).

Fig. 5.15 shows (a) the experimental overviews, which were monitored by the second video camera at a fixed position, (b) the input images of 512×512 pixels, in which the



(a) Experimental overviews

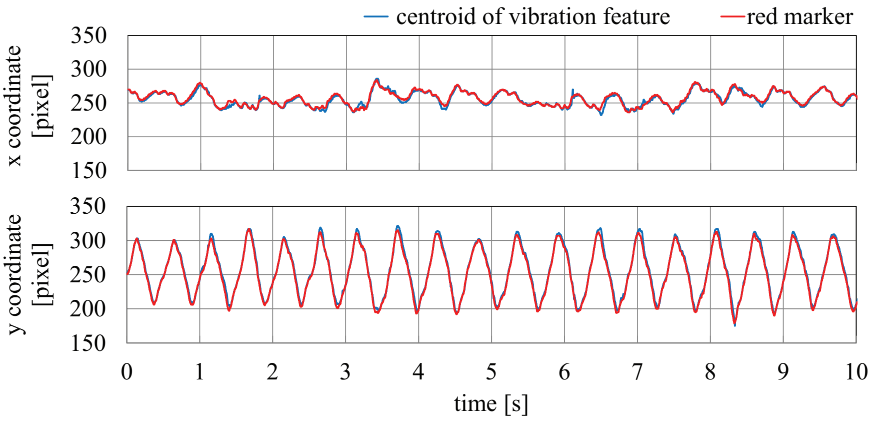
(b) Input images

(c) Vibration features in ROI regions

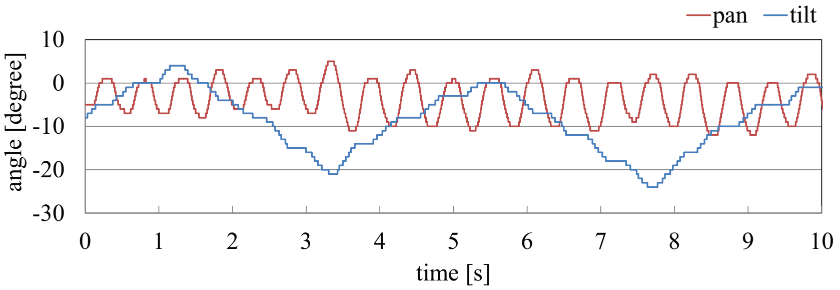
Figure 5.15: Experimental overviews, input images of 512×512 pixels, and masked vibration features in the tracked ROI regions of 128×128 pixels in mechanical tracking experiment.

red-line squares indicate the tracked ROI regions, and (c) the vibration features extracted by our algorithm, which were white-plotted over the ROI images of 128×128 pixels with their centroids and the marker positions. These images were taken at intervals of 0.4 s for t = 3.0~4.2 s; t = 0 was the start time of observation. Fig. 5.16 shows (a) the centroid of the vibration features and the marker position, and (b) the pan and tilt angles of the active vision system for 10 s. Fig. 5.17 shows the deviations of the extracted centroids from the marker position. Fig. 5.18 shows the temporal changes in the image intensities at A (64, 20), B (64, 85) and C (64, 108) in the ROI image of 128×128 pixels for t = 2.6~2.9 s, when the 33-rps fan passed over the 18-rps fan and the moving human in the background.

In Fig. 5.18, the periodic changes at 100 Hz at point A around the 33-rps fan was within the frequency range of the pixel-level band-pass filters, whereas the temporal



(a) Centroid of vibration features



(b) Pan and tilt angles of active vision

Figure 5.16: Centroid of vibration features, and pan and tilt angles of active vision system in mechanical tracking experiment.

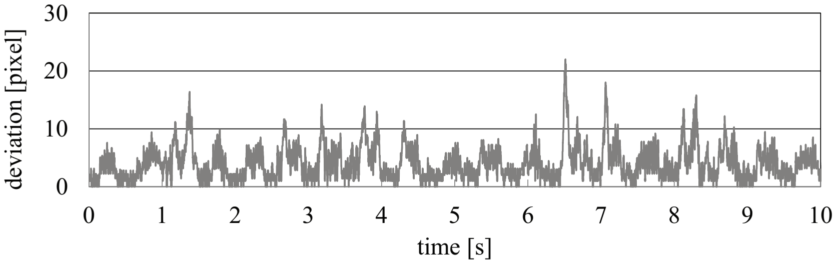


Figure 5.17: Deviation of image centroids of vibration features in mechanical tracking experiment.

changes at points *B* and *C*, which correspond to the 18-rps fan and the moving human, respectively, were much lower than 100 Hz, and had few components in the filter frequency range. Thus, the rotating fan was continuously extracted and tracked around the center of the camera view without miss-tracking any obstacle when the ROI region was passing over the obstacle fan and the moving human, as shown in Fig. 5.15. In Fig. 5.15(b) and

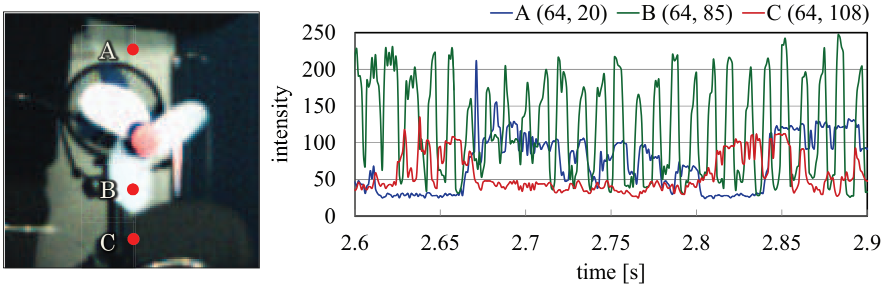
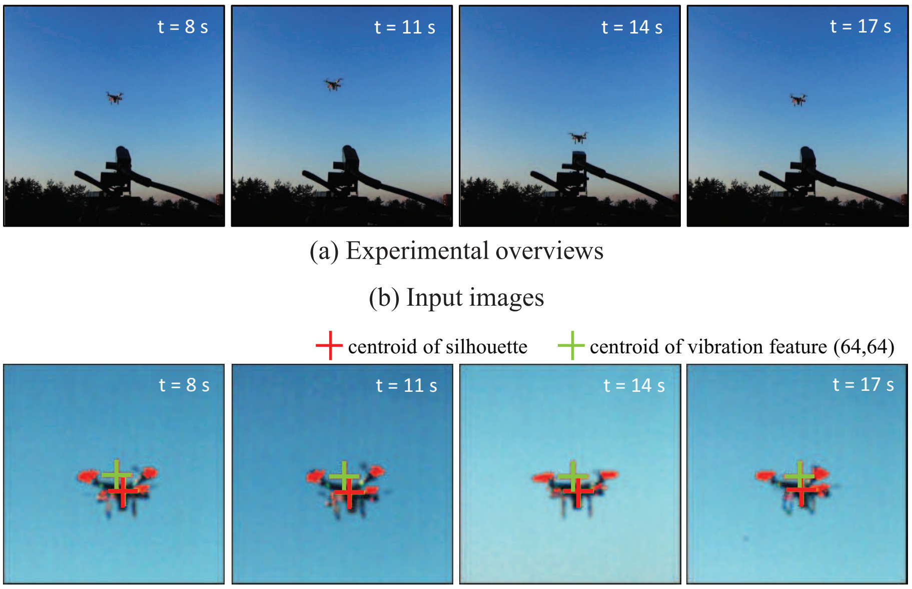


Figure 5.18: Temporal changes in image intensities in mechanical tracking experiment.

Fig. 5.16(a), the centroid of the vibration features deviates slightly from the control target of (256, 256). This tendency indicates that the tracking error from the image center was caused mainly by the mechanical limit of the active vision system because the latency effect on the digital filter in the experiment was insignificant. The vibrating object to be tracked maintained its position at the center of the ROI region, and the deviation of the centroid from the marker position is always small. The mean of the deviation was 4.5 pixels, and it corresponded to 6.8 mm at a distance of 1.4 m from the camera head. Thus, vibration source localization was realized with directive errors of 0.28 degrees in the experiment.

5.5.3 Flying Quadcopter Tracking Experiment

Finally, we present the experimental results when the active vision system tracked a quadcopter flying in an outdoor scene as a vibrating object using our vibration source localization method. The quadcopter used in the experiment was a Phantom 1 (DJI Co. Ltd, China) with four 21-cm dual-blade propellers; its dimensions are 29×29×20 cm. The quadcopter was flying diagonally upward at a distance of approximately 8 m from the active vision system, and the flapping frequencies of the four propellers changed in the range from 85 Hz to 95 Hz corresponding to the flight operation command. The measurement area was 7.25×7.25 m for 512×512 pixels at a distance of 8 m in front of the camera head, on which the f = 6 mm C-mount lens was mounted; one pixel corresponds to 1.41 cm. According to twice the flapping frequencies of the dual-blade propellers, the tap coefficients a_s, b_s of pixel-level digital filters were set such that they operated as the



(a) Experimental overviews

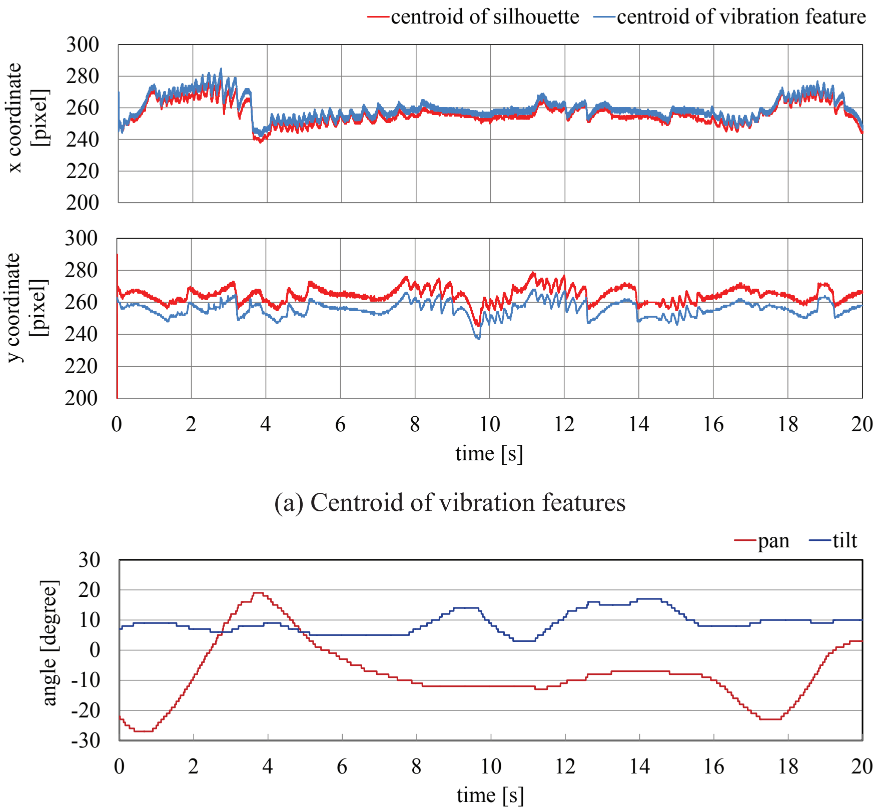
(b) Input images

(c) Vibration features in ROI regions

Figure 5.19: Experimental overviews, input images of 512×512 pixels, and masked vibration features in the tracked ROI regions of 128×128 pixels in flying quadcopter tracking experiment.

band-bass filters, the center frequency of which was $f_0 = 180$ Hz and half width was 20 Hz. The parameters were set to $p = 4$, $\Delta T_f = 22$ ms, and $T_0 = 1/f_0 = 6$ ms. The thresholds $\theta_1 \sim \theta_5$ were the same as those provided in Subsection 5.5.1.

Fig. 5.19 shows (a) the experimental overviews, monitored using a video camera at a fixed position, (b) the input images of 512×512 pixels with the tracked ROI regions enclosed by the red-lines, and (c) the extracted vibration features, which were plotted in white over the ROI images of 128×128 pixels with their centroids and the ground-truth positions of the quadcopter. We used the xy centroid of the quadcopter's silhouette in the tracked ROI images as its ground-truth position. The silhouette was obtained as a binary image by thresholding the dark-brightness pixels since there was no obstacle and background in all the tracked ROI images. The images were taken at intervals of 3 s for $t = 8 \sim 17$ s; $t = 0$ was the start time of observation. Fig. 5.20 shows (a) the centroid of the vibration features, and (b) the pan and tilt angles of the active vision system for 20 s. Fig. 5.21 shows the deviations of the extracted centroids from the centroids of the silhouettes and the corrected position for the four propellers, where many pixels were extracted



(a) Centroid of vibration features

(b) Pan and tilt angles of active vision

Figure 5.20: Centroid of vibration features and pan and tilt angles of active vision system in flying quadcopter tracking experiment.

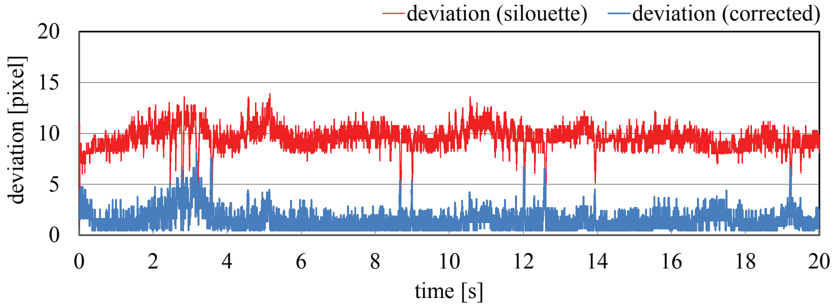


Figure 5.21: Deviation of image centroids of vibration features in flying quadcopter tracking experiment.

as vibration region. The corrected position was computed by adding an offset value to the centroid of the silhouette, because there was a certain displacement between the center position of the four propellers and that of the quadcopter's silhouette. Assuming the displacement was constant in the images, the offset value was set as (3.4, 8.8); it was the

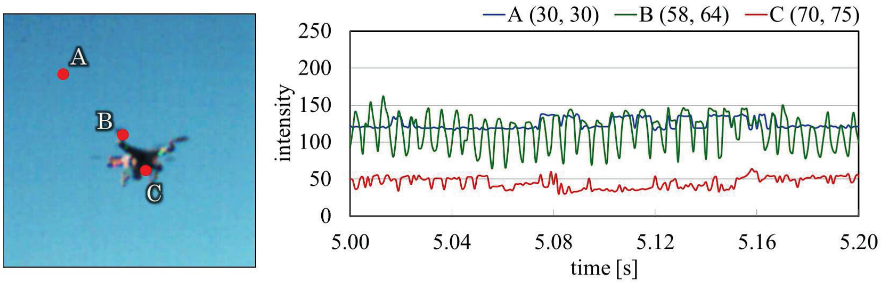


Figure 5.22: Temporal changes in image intensities in flying quadcopter tracking experiment.

averaged displacement between the extracted centroids and the centroids of the silhouettes for $t = 0 \sim 20$ s. Fig. 5.22 shows the temporal changes in the image intensities at $A(30, 30)$, $B(58, 64)$ and $C(70, 75)$ in the ROI images of 128×128 pixels for $t = 5.0{\sim}5.2$ s. In the experiment, the target quadcopter was moved at its full speed in horizontal and vertical directionsat a distance of 6~8 m from the camera head.

In Fig. 5.22, the periodic changes at approximately 180 Hz at point A around a dual-blade propeller was within the frequency range of the pixel-level band-pass filters, whereas the temporal changes at points B and C, which correspond to the body of the quadcopter and the background sky, respectively, had few components in the filter frequency range. As shown in Fig. 5.19(c), the regions around the four propellers were extracted as vibration features by the pixel-level band-pass filters under lower resolution or partially occluded conditions such that the size of the propeller was 13 pixels or less in the camera view. Because of the mechanical delay in the tracking control of the active vision system, a slight deviation of the centroid of the vibration features from the image center can be observed in Fig. 5.19(b) and Fig. 5.20(a). Centroid of the quadcopter's vibration region was consistantly tracked at the center of the ROI regions, as shown in Fig. 5.19(c), without any significant latency effect on the digital filter, and the deviation from the center position of the quadcopter's silhouette was constantly observed for $t = 0 \sim 20$ s, as shown in Fig. 5.21. The means of the deviations from the centroids of the silhouettes and the corrected position for the four propellers were 9.5 pixel and 1.5 pixel, respectively. At a distance of 8 m from the camera head, they corresponded to 13.8 cm and 2.1 cm, respectively. The former represents the offset displacement between the quadcopter's silhouette and its four propellers, and the latter indicates the averaged localization error in tracking a flying quadcopter. Thus, the flying quadcopter was tracked in images with directive errors of 0.15 degrees in the experiment.

5.6 Concluding Remarks

vision-based vibration source localization algorithm based on vibration-based image features, which are obtained by detecting periodic changes in image intensities at pixels around objects vibrating at audio-level frequency by using pixel-level band-pass digital filters. This approach can significantly reduce the latency effect on digital filters by introducing the ROI tracking process into a digital filter. Our algorithm was implemented on a high-speed vision platform with a 2-DOF active vision system, and its effectiveness in pixel-level vibration source localization was demonstrated by conducting several tracking experiments for vibrating objects with visual feedback control at 1000 fps; a fan rotating at 33 rps was tracked in images with directive errors of 0.28 degrees, and a flying quadcopter with rotating propellers was tracked with directive errors of 0.15 degrees. The experimental results show that the proposed HFR-vision-based method with pixel-level digital filters enables vibration source localization with sub-degree-level angular directivity in real time.

Chapter 6

Conclusion

This study, firstly we concentrated on the primitive vibration source localization with pixel-level band-pass filters for temporal brightness changes, and it did not directly concern the geometric motion of a target object; the frequency range of temporal brightness changes at pixels around the target object may not be matched with that of its geometric motion when the target object has a periodic surface pattern. To realize a more universal vibration feature detector, which is invariant to any spatial appearance of the target object, it becomes more effective to apply our pixel-level band-pass filters to geometric motion fields estimated by optical flow [148, 149]. This is one of well-known image processing algorithms, instead of using the image brightness. Besides, by combining our proposed dynamics-based vibration feature with appearance-based recognition methods, the accuracy and robustness in vibration source localization will be remarkably improved when the target frequency range overlaps with that of background scenes. Thus, in future work, we intend to improve these points toward more universal vibration source localization under more extreme conditions and accelerate the computational speed for real-time processing of HFR video.

Then, we proposed a real-time tracking of a non-large object vibrating at a known frequency, whose apparent size on the image sensor is less than the ROI region, and the time-variation in vibration frequency and apparent size was not considered. In the future, we plan to improve the algorithm for more universal vibration source localization with automated tuning of parameters for digital filters, ROI scale adjustment, and other additional processes as well as the performance of the system by using GPUs for enlarging the spatial resolution of the ROI region and accelerating the visual sampling rate for higher frequency vibration detection; and expand our system for practical applications, such as simultaneous surveillance of flying drones with accurate localization.

Bibliography

[1] T.M. Bernard, B.Y. Zavidovique, and F.J. Devos, “A programmable artificial retina,” *IEEE J. of Solid-State Circuits*, Vol. 28, No. 7, pp. 789–797, 1993.

[2] J.E. Eklund, C. Svensson, and A. Astrom, “VLSI implementation of a focal plane image processor - A realization of the near-sensor image processing concept,” *IEEE Trans. on VLSI Systems*, Vol. 4, No. 3, pp. 322–335, 1996.

[3] T. Komuro, I. Ishii, and M. Ishikawa, “Vision chip architecture using general-purpose processing elements for lms vision system,” *Proc. of IEEE Int. Workshop on Computer Architecture for Machine Perception*, pp. 276–279, 1997.

[4] M. Ishikawa, K. Ogawa, T. Komuro, and I. Ishii, “A cmos vision chip with simd processing element array for lms image processing,” *Proc. IEEE Int. Solid-State Circuits Conf.*, pp. 206–207, 1999.

[5] T. Komuro, S. Kagami, and M. Ishikawa, “A Dynamically Reconfigurable SIMD Processor for a Vision Chip,” *IEEE J. of Solid-State Circuits*, Vol. 39, No. 1, pp. 265–268, 2004.

[6] I. Ishii, K. Yamamoto, and M. Kubozono, “Higher order autocorrelation vision chip,” *IEEE Trans. on Electron Devices*, Vol. 53, No. 8, pp. 1797–1804, 2006.

[7] S. Hirai, M. Zakoji, A. Masubuchi, and T. Tsuboi, “Realtime FPGA-based vision system,” *J. of Robotics and Mechatronics*, Vol. 17, No. 4, pp. 401–409, 2005.

[8] Y. Watanabe, T. Komuro, and M. Ishikawa, “955-fps real-time shape measurement of a moving/deforming object using high-speed vision for numerous-point analysis,” *Proc. IEEE Int. Conf. Robot, and Autom.*, pp. 3192–3197, 2007.

[9] I. Ishii, T. Taniguchi, R. Sukenobe, and K. Yamamoto, "Development of high-speed and real-time vision platform, H^3 Vision," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, pp. 3671–3678, 2009.

[10] I. Ishii, T. Tatebe, Q. Gu, Y. Moriue, T. Takaki, and K. Tajima, "2000 fps real-time vision system with high-frame-rate video recording," *Proc. IEEE Int. Conf. Robot, and Autom.*, pp. 1536–1541, 2010.

[11] I. Ishii, Y. Nakabo, and M. Ishikawa, "Target tracking algorithm for 1ms visual feedback system using massively parallel processing," *Proc. IEEE Int. Conf. Robot, and Autom.*, pp. 2309–2314, 1996.

[12] Y. Nakabo, M. Ishikawa, H. Toyoda, and S. Mizuno, "1 ms column parallel vision system and its application of high speed target tracking," *Proc. IEEE Int. Conf. Robot, and Autom.*, pp. 650–655, 2000.

[13] Y. Nakamura, K. Kishi, and H. Kawakami, "Heartbeat synchronization for robotic cardiac surgery," *Proc. IEEE Int. Conf. Robot, and Autom.*, pp. 2014–2019, 2001.

[14] Y. Nakabo, I. Ishii, and M. Ishikawa, "3D tracking using two high-speed vision systems," *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, pp. 360–365, 2002.

[15] D. Shiokata, A. Namiki, and M. Ishikawa, "Robot Dribbling Using a High-Speed Multifingered Hand and a High-Speed Vision System," *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, pp. 3945–3950, 2005.

[16] S. Mizusawa, A. Namiki, and M. Ishikawa, "Tool Manipulation by a Multifingered Hand Using a High-speed Vision," *Proc. of the 8th SICE Syst. Integration Division Annual Conf.*, pp. 55–66, 2007.

[17] Y. Nie, I. Ishii, K. Yamamoto, K. Orito, and H. Matsuda, "Real-time scratching behavior quantification system for laboratory mice using high-speed vision," *J. of Real-Time Image Processing*, Vol. 4, No. 2, pp. 181–190, 2009.

[18] Y.D. Wang, I. Ishii, T. Takaki, and Kenji Tajima, “An Intelligent High-Frame-Rate Video Logging System for Abnormal Behavior Analysis,” *J. of Robotics and Mechatronics*, Vol. 23, No. 1, pp. 53–65, 2011.

[19] H. Yang, T. Takaki, and I. Ishii, “Simultaneous dynamics-based visual inspection using modal parameter estimation,” *J. of Robotics and Mechatronics*, Vol. 23, No. 1, pp. 180–195, 2011.

[20] H. Yang, T. Takaki, and I. Ishii, “A Structural Damage Quantification Method for HFR-Video-Based Modal Testing,” *J. of Syst. Design and Dynamics*, Vol. 5, No. 4, pp. 624–641, 2011.

[21] K. Okumura, H. Oku, M. Ishikawa, “High-speed gaze controller for millisecond-order pan/tilt camera,” *Proc. IEEE Int. Conf. Robot, and Autom.*, pp. 6186–6191, 2011.

[22] I. Ishii, S. Takemoto, T. Takaki, M. Takamoto, K. Imon, K. Hirakawa, “Real-time laryngoscopic measurements of vocal fold vibrations,” *Proc. IEEE Int. Conf. Robot, and Autom.*, pp. 6186–6191, 2011.

[23] H. Gao, Q. Gu, T. Aoyama, T. Takaki, and I. Ishii, “A self-projected light-section method for fast three-dimensional shape inspection,” *Int. J. of Optomechatroics*, Vol. 6, No. 4, pp. 289–303, 2012.

[24] I. Ishii, T. Ichida, Q. Gu, T. Takaki, “500-fps face tracking system,” *J. of Real-Time Image Processing*, Vol. 8, No. 4, pp. 379–388, 2013.

[25] Q. Gu, T. Takaki, and I. Ishii, “Fast FPGA-based Multi-object Feature Extraction,” *IEEE Trans. on Circ. Syst. Video Tech.*, Vol. 23, No. 1, pp. 30–45, 2013.

[26] A. Namiki, S. Matsushita, T. Ozeki, K. Nonami, “Hierarchical processing architecture for an air-hockey robot system,” *Proc. IEEE Int. Conf. Robot, and Autom.*, pp. 1187–1192, 2013.

[27] J. Chen, T. Tamamoto, T. Aoyama, T. Takaki, I. Ishii, “Simultaneous projection mapping using high-frame-rate depth vision,” *Proc. IEEE Int. Conf. Robot, and Autom.*, pp. 4506–4511, 2014.

[28] Q. Gu, T. Aoyama, T. Takaki, and I. Ishii, “Simultaneous vision-based shape and motion analysis of cell fast-flowing in a microchannel,” *IEEE Transactions on Automation Science and Engineering*, Vol. 12, No. 1, pp. 204–215, 2015.

[29] Y. Wu, J. Lim, M. Yang, “Object tracking benchmark,” *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 37, No. 9, pp. 1834–1848, 2015.

[30] X. Li, W. Hu, C. Shen, Z. Zhang, A. Dick, and A. Hengel, “A survey of appearance models in visual object tracking,” *ACM Trans. Intell. Syst. Technol.*, Vol. 4, No. 4, pp. 1–48, 2013.

[31] T. Senst, V. Eiselein, C. Shen, and T. Sikora, “Robust local optical flow for feature tracking,” *IEEE Trans. Circuits Syst. Video Technol.*, Vol. 22, No. 9, pp. 1377–1387, 2012.

[32] D. Doyle, A. Jennings, and J. Black, “Optical flow background estimation for real-time pan/tilt camera object tracking,” *Measurement*, Vol. 48, pp. 195–207, 2014.

[33] D. Guo, A. Van de Ven, and X. Zhou, “Red blood cell tracking using optical flow methods,” *IEEE J. Biomed. Health Inform.*, Vol. 18, No. 3, pp. 991–998, 2014.

[34] O. Zoidi, A. Tefas, and I. Pitas, “Visual object tracking based on local steering kernels and color histogramss,” *IEEE Trans. Circuits Syst. Video Technol.*, Vol. 23, No. 5, pp. 870–882, 2013.

[35] D. Kim, H. kim, and S. Ko, “Spatial color histogram based center voting method for subsequent object tracking and segmentation,” *Image Vis. Comput.*, Vol. 29, No. 12, pp. 850–860, 2011.

[36] P. Liang, E. Blasch, and H. Ling, “Encoding color information for visual tracking: algorithms and benchmark,” *IEEE Trans. Image Process.*, Vol. 24, No. 12, pp. 5630–5644, 2015.

[37] F. Bousetouane, L. Dib, and H. Snoussi, “Improved mean shift integrating texture and color features for robust real time object tracking,” *Vis. Comput.*, Vol. 29, No. 3, pp. 155–170, 2013.

[38] J. Ning, L. Zhang, D. Zhang, and C. Wu, “Robust object tracking using joint color-texture histogram,” *Int. J. Pattern Recognit. Artif. Intell.*, Vol. 23, No. 7, pp. 1245–1263, 2009.

[39] J. Wang, and Y. Yagi, “Integrating color and shape-texture features for adaptive real-time object tracking,” *IEEE Trans. Image Process.*, Vol. 17, No. 2, pp. 235–240, 2008.

[40] H. Zhou, Y. Yuan, and C. Shi, “Object tracking using SIFT features and mean shift,” *Comput. Vis. Image Underst.*, Vol. 113, No. 3, pp. 345–352, 2009.

[41] W. Zhao, and C. Ngo, “Object tracking using SIFT features and mean shift,” *IEEE Trans. Image Process.*, Vol. 22, No. 3, pp. 980–991, 2013.

[42] S. Zhang, C. Bauckhage, and A. Cremers, “Informed Haar-Like Features Improve Pedestrian Detection,” *Proc. IEEE Conf. Comp. Vis. Patt. Recog.*, pp. 947–954, 2014.

[43] S. Pavani, Y. Yuan, D. Delgado-Gomez, and A. Frangi, “Gaussian weak classifiers based on co-occurring Haar-like features for face detection,” *Pattern Anal. Appl.*, Vol. 17, No. 2, pp. 431–439, 2014.

[44] N. Dalal, and B. Triggs, “Histograms of Oriented Gradients for Human Detection,” *Proc. IEEE Conf. Comp. Vis. Patt. Recog.*, pp. 886–893, 2005.

[45] B. Wu, Y. Yuan, C. Kao, C. Jen, Y. Li, Y. Chen, and J. Juang, “A Relative-Discriminative-Histogram-of-Oriented-Gradients-Based Particle Filter Approach to

Vehicle Occlusion Handling and Tracking," *IEEE Trans. Ind. Electron.*, Vol. 61, No. 8, pp. 4228–4237, 2014.

[46] P. Chen, C, Huang, C. Lien, and Y. Tsai, "An efficient hardware implementation of HOG feature extraction for human detection," *IEEE Trans. Intell. Transp. Syst.*, Vol. 15, No. 2, pp. 656–662, 2014.

[47] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 24, No. 7, pp. 971–987, 2002.

[48] B. Yang, and S. Chen, "A comparative study on local binary pattern (LBP) based face recognition: LBP histogram versus LBP image," *Neurocomputing*, Vol. 120, pp. 365–379, 2013.

[49] A. Satpathy, X. Jiang, and W. Eng, "LBP-based edge-texture features for object recognition," *IEEE Trans. Image Process.*, Vol. 23, No. 5, pp. 1953–1964, 2014.

[50] A. Satpathy, and M. Trivedi, "Looking at vehicles on the road: A survey of vision-based vehicle detection, tracking, and behavior analysis," *IEEE Trans. Intell. Transp. Syst.*, Vol. 14, No. 4 pp. 1773–1795, 2013.

[51] R. Chavez-Garcia, and O. Aycard, "Multiple sensor fusion and classification for moving object detection and tracking," *IEEE Trans. Intell. Transp. Syst.*, Vol. 17, No. 2 pp. 525–534, 2016.

[52] D. Llorca, S. Sánchez, M. Ocaña, and M. Sotelo, "Vision-based traffic data collection sensor for automotive applications," *Sensors*, Vol. 10, No. 1 pp. 860–875, 2010.

[53] G. Schuster, and A. Katsaggelos, *Rate-Distortion Based Video Compression: Optimal Video Frame Compression and Object Boundary Encoding*; Kluwer Academic Publishers: Dordrecht, The Netherlands, 2013.

[54] S. Rautaray, and A. Agrawal, "Vision based hand gesture recognition for human computer interaction: A survey," *Artif. Intell. Rev.*, Vol. 43, No. 1 pp. 1–54, 2015.

[55] V. Prisacariu, and I. Reid, “3D hand tracking for human computer interaction,” *Image Vis. Comput.*, Vol. 30, No. 3 pp. 236–250, 2012.

[56] D. Tran, and J. Yuan, “Optimal Spatio-Temporal Path Discovery for Video Event Detection,” *Proc. IEEE Conf. Comp. Vis. Patt. Recog.*, pp. 3321–3328, 2011.

[57] J. Meng, J. Yuan, G. Wang, and Y. Tan, “Object Instance Search in Videos via Spatio-Temporal Trajectory Discovery,” *IEEE Trans. Multimed.*, Vol. 18, No. 1 pp. 116–127, 2016.

[58] M. Jain, J. Van Gemert, H. Jégou, P. Bouthemy and G. Cees, “Action Localization with Tubelets from Motion,” *Proc. IEEE Conf. Comp. Vis. Patt. Recog.*, pp. 740–747, 2014.

[59] G. Yu, and J. Yuan, “Fast Action Proposals for Human Action Detection and Search,” *Proc. IEEE Conf. Comp. Vis. Patt. Recog.*, pp. 1302–1311, 2015.

[60] G. Gkioxari, and J. Malik, “Finding Action Tubes,” *Proc. IEEE Conf. Comp. Vis. Patt. Recog.*, pp. 759–768, 2015.

[61] P. Mettes, J. Van Gemert, S. Cappallo, T. Mensink, and G. Cees, “Bag-of-Fragments: Selecting and Encoding Video Fragments for Event Detection and Recounting,” *Proc. of the 5th ACM on Int. Conf. on Multimedia Retrieval*, pp. 427–434, 2015.

[62] A. González, D. Vázquez, S. Ramos, A. López, and J. Amores, “Spatiotemporal Stacked Sequential Learning for Pedestrian Detection,” *Proc. of the Iberian Conf. Patt. Recog. Image Analysis*, pp. 3–12, 2015.

[63] N. Jiang, H. Su, W. Liu, and Y. Wu, “Discriminative Metric Preservation for Tracking Low-Resolution Targets,” *IEEE Trans. Image Process.*, Vol. 21, No. 3 pp. 1284–1297, 2012.

[64] S. Biswas, G. Aggarwal, P. Flynn, and K, Bowyer, “Pose-robust recognition of low-resolution face images,” *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 35, No. 12 pp. 3037–3049, 2013.

[65] S. Argentieri, P. Danes, and P. Soueres, “A survey on sound source localization in robotics: Binaural to array processing methods,” *Comput. Speech Lang.*, vol. 34, no. 1, pp. 87–112, Nov. 2015.

[66] J. Lanslots, F. Deblauwe, and K. Janssens, “Selecting sound source localization techniques for industrial applications,” *Sound Vibrat.*, vol. 44, no. 6, pp. 6–9, 2010.

[67] A. Fuchs, C. Feldbauer, and M. Stark, “Monaural sound localization,” *Proc. INTERSPEECH*, Aug. 2011, pp. 2521–2524.

[68] G. Jang and T. Lee, “A maximum likelihood approach to single-channel source separation,” *J. Mach. learn. Res.*, vol. 4, pp. 1365–1392, Dec. 2003.

[69] K. Kim and Y. Kim, “Monaural sound localization based on structure-induced acoustic resonance,” *Sensors*, vol. 15, no. 2, pp. 3872–3895, Feb. 2015.

[70] T. Van den Bogaert, T. Klasen, M. Moonen, L. Van Deun, and J. Wouters, “Horizontal localization with bilateral hearing aids: Without is better than with,” *J. Acoust. Soc. Amer.*, vol. 119, no. 1, pp. 515–526, Jan. 2006.

[71] M. Raspaud, H. Viste, and G. Evangelista, “Binaural source localization by joint estimation of ILD and ITD,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 1, pp. 68–77, Jan. 2010.

[72] T. May, S. Van De Par, and A. Kohlrausch, “A probabilistic model for robust localization based on a binaural auditory front-end,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 1, pp. 1–13, Jan. 2011.

[73] U. H. Kim, K. Nakadai, and H. G. Okuno, “Improved sound source localization in horizontal plane for binaural robot audition,” *Appl. Intell.*, vol. 42, no. 1, pp. 63–74, Jan. 2015.

[74] M. Aytekin, E. Grassi, M. Sahota, and C. F. Moss, “The bat head-related transfer function reveals binaural cues for sound localization in azimuth and elevation,” *J. Acoust. Soc. Amer.*, vol. 116, no. 6, pp. 3594–3605, Dec. 2004.

[75] H. Nakashima and T. Mukai, “3D sound source localization system based on learning of binaural hearing,” *Proc. IEEE Int. Conf. Syst., Man, Cybern.*, Oct. 2005, pp. 3534–3539.

[76] X. Zhong, W. Yost, and L. Sun, “Dynamic binaural sound source localization with ITD cues: Human listeners,” *J. Acoust. Soc. Amer.*, vol. 137, no. 4, pp. 2376–2376, Jan. 2015.

[77] S. Vesa, “Binaural sound source distance learning in rooms,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 8, pp. 1498–1507, Nov. 2009.

[78] Y. C. Lu and M. Cooke, “Binaural estimation of sound source distance via the direct-to-reverberant energy ratio for static and moving sources,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1793–1805, Sept. 2010.

[79] E. Georganti, T. May, S. Van De Par, and J. Mourjopoulos, “Sound source distance estimation in rooms based on statistical properties of binaural signals,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 8, pp. 1727–1741, Aug. 2013.

[80] B. Mungamuru and P. Aarabi, “Enhanced sound localization,” *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 34, no.3, pp. 1526–1540, Jun. 2004.

[81] C. H. Knapp and G. C. Carter, “The generalized correlation method for estimation of time delay,” *IEEE Trans. Acoust., Speech, Signal Process*, vol. 24, no. 4, pp. 320–327, Aug. 1976.

[82] J. Benesty, “Adaptive eigenvalue decomposition algorithm for passive acoustic source localization,” *J. Acoust. Soc. Amer.*, vol. 107, no. 1, pp. 384–391, Mar. 2000.

[83] J. Chen, J. Benesty, and Y. Huang, “Time delay estimation in room acoustic environments: An overview,” *EURASIP J. Appl. Signal Process.*, vol. 2006, pp. 1–19, 2006.

[84] A. Brutti, M. Omologo, and P. Svaizer, “Comparison between different sound source localization techniques based on a real data collection,” *Proc. Hands-Free Speech Commun. and Microphone Arrays*, May 2008, pp. 69–72.

[85] R. O. Schmidt, “Multiple emitter location and signal parameter estimation,” *IEEE Trans. Antennas Propag.*, vol. 34, no. 3, pp. 276–280, Mar 1986.

[86] K. Nakamura, K. Nakadai, F. Asano, Y. Hasegawa, and H. Tsujino, “Intelligent sound source localization for dynamic environments,” *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Oct. 2009, pp. 664–669.

[87] K. Okutani, T. Yoshida, K. Nakamura, and K. Nakadai, “Outdoor auditory scene analysis using a moving microphone array embedded in a quadrocopter,” *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Oct. 2012, pp. 3288–3293.

[88] L. J. Griffiths and C. W. Jim, “An alternative approach to linearly constrained adaptive beamforming,” *IEEE Trans. Antennas Propag.*, vol. 30, no. 1, pp. 27–34, Jan. 1982.

[89] B. D. Van Veen and K. M. Buckley, “Beamforming: A versatile approach to spatial filtering,” *IEEE ASSP Mag.*, vol. 5, no. 2, pp. 4–24, Apr. 1988.

[90] J. M. Valin, F. Michaud, and J. Rouat, “Robust localization and tracking of simultaneous moving sound sources using beamforming and particle filtering,” *Robot. Auton. Syst.*, vol. 55, no. 3, pp. 216–228, Mar. 2007.

[91] T. J. Tsai, A. Stolcke, and M. Slaney, “A study of multimodal addressee detection in human-human-computer interaction,” *IEEE Trans. Multimedia*, vol. 17, no. 9, pp. 1550–1561, Sept. 2015.

[92] V. P. Minotto, C. R. Jung, and B. Lee, “Simultaneous-speaker voice activity detection and localization using mid-fusion of SVM and HMMs,” *IEEE Trans. Multimedia*, vol. 16, no. 4, pp. 1032–1044, Jun. 2014.

[93] M. F. Fallon and S. J. Godsill, “Acoustic source localization and tracking of a time-varying number of speakers,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 4, pp. 1409–1415, May 2012.

[94] C. Zhang, D. Florêncio, D. E. Ba, and Z. Zhang, “Maximum likelihood sound source localization and beamforming for directional microphone arrays in distributed meetings,” *IEEE Trans. Multimedia*, vol. 10, no. 3, pp. 538–548, Apr. 2008.

[95] F. Ribeiro, D. Florencio, D. Ba, and C. Zhang, “Geometrically constrained room modeling with compact microphone arrays,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 5, pp. 1449–1460, Jul. 2012.

[96] T. Hori, S. Araki, T. Yoshioka, M. Fujimoto, S. Watanabe, T. Oba, A. Ogawa, K. Otsuka, D. Mikami, K. Kinoshita and others, “Low-latency real-time meeting recognition and understanding using distant microphones and omni-directional camera,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 2, pp. 499–513, Feb. 2012.

[97] J. S. Hu, C. Y. Chan, C. K. Wang, M. T. Lee, and C. Y. Kuo, “Simultaneous localization of a mobile robot and multiple sound sources using a microphone array,” *Advan. Robot.* vol. 25, no. 1-2, pp. 135–152, 2011.

[98] E. Martinson and D. Brock, “Auditory perspective taking,” *IEEE Trans. Cybern.*, vol. 43, no. 3, pp. 957–969, Jun. 2013.

[99] F. Keyrouz, “Advanced binaural sound localization in 3-D for humanoid robots,” *IEEE Trans. Instrum. Meas.*, vol. 63, no. 9, pp. 2098–2107, Aug. 2014.

[100] R. Li and D. He, “Rotational machine health monitoring and fault detection using EMD-based acoustic emission feature quantification,” *IEEE Trans. Instrum. Meas.*, vol. 61, no. 4, pp. 990–1001, Apr. 2012.

[101] T. He, D. Xiao, Q. Pan, X. Liu, and Y. Shan, “Analysis on accuracy improvement of rotor-stator rubbing localization based on acoustic emission beamforming method,” *Ultrasonics*, vol. 54, no. 1, pp. 318–329, Jan. 2014.

[102] J. H. Zhang and B. Han, “Analysis of engine front noise using sound intensity techniques,” *Mech. syst. signal process.*, vol. 19, no. 1, pp. 213–221, Jan. 2005.

[103] V. G. C. Cook and A. Ali, “End-of-line inspection for annoying noises in automobiles: Trends and perspectives,” *Appl. Acoust.*, vol. 73, no. 3, pp. 265–275, Mar. 2012.

[104] J. A. Ballesteros, E. Sarradj, M. D. Fernández, T. Geyer, and M. J. Ballesteros, “Noise source identification with Beamforming in the pass-by of a car,” *Appl. Acoust.*, vol. 93, pp. 106–119, Jun. 2015.

[105] D. Yang, Z. Wang, B. Li, and X. Lian, “Development and calibration of acoustic video camera system for moving vehicles,” *J. of Sound and Vibrat.*, vol. 330, no. 11, pp. 2457–2469, May 2011.

[106] B. Barsikow, W. F. King, and E. Pfizenmaier, “Wheel/rail noise generated by a high-speed train investigated with a line array of microphones,” *J. of Sound and Vibrat.*, vol. 118, no. 1, pp. 99–122, Oct. 1987.

[107] P. Castellini, A. Sassaroli, A. Paonessa, A. Peiffer, and A. Roeder, “Average beamforming in reverberant fields: Application on helicopter and airplane cockpits,” *Appl. Acoust.*, vol. 74, no. 1, pp. 198–210, Jan. 2013.

[108] Y. Li, M. Smith, and X. Zhang, “Measurement and control of aircraft landing gear broadband noise,” *Aerosp. Sci. Technol.*, vol. 23, no. 1, pp. 213–223, Dec. 2012.

[109] J. Busset, F. Perrodin, P. Wellig, B. Ott, K. Heutschi, T. Rühl, and T. Nussbaumer, “Detection and tracking of drones using advanced acoustic cameras,” *SPIE Security+ Defence*, pp. 96470F–96470F, 2015.

[110] T. Pham and N. Srour, “TTCP AG-6: acoustic detection and tracking of UAVs,” *Defense and Security*, 2004, pp. 24-30.

Cutler02a

[111] Dedrone. (2015, November 28) Multi-Sensor Drone Warning System [Online]. Available: http://www.dedrone.com/en/dronetracker/drone-detection-hardware.

[112] Droneshield. (2015, November 28) HOW DRONESHIELD WORKS [Online]. Available: https://www.droneshield.com/how-droneshield-works.

[113] K. Nakadai, T. Takahashi, H. G. Okuno, H. Nakajima, Y. Hasegawa, and H. Tsujino, "Design and Implementation of Robot Audition System'HARK'–Open Source Software for Listening to Three Simultaneous Speakers," *Advanced Robotics*, vol. 24, no. 5-6, pp. 739–761, 2010.

[114] F. Grondin, D. Létourneau, F. Ferland, V. Rousseau, and F. Michaud, "The ManyEars open framework," *Autonomous Robots*, vol. 34, no. 3, pp. 217–232, Feb. 2013.

[115] J. Bonnal, S. Argentieri, P. Danès, and J. Manhès, "Speaker localization and speech extraction with the EAR sensor," *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Oct. 2009, pp. 670–675.

[116] A. Jana, *Kinect for windows SDK programming guide*, Birmingham, UK: Packt Publishing Ltd, 2012, pp. 231–235.

[117] T. Mizumoto, K. Nakadai, T. Yoshida, R. Takeda, T. Otsuka, T. Takahashi, and H. G. Okuno, "Design and implementation of selectable sound separation on the Texai telepresence system using HARK," *Proc. IEEE Int. Conf. Robot. Autom.*, May 2011, pp. 2130–2137.

[118] J. M. Valin, F. Michaud, and J. Rouat, "Robust 3D localization and tracking of sound sources using beamforming and particle filtering," *Proc. IEEE ICASSP*, May 2006, vol. 4, pp. 841–844.

[119] V. Lunati, J. Manhes, and P. Danes, "A Versatile System-on-a-Programmable-Chip for Array Processing and Binaural Robot Audition," *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Oct. 2012, pp. 998–1003.

[120] L. A. Seewald, L. Gonzaga, M. R. Veronez, V. P. Minotto, and C. R. Jung, “Combining SRP-PHAT and two Kinects for 3D Sound Source Localization,” *Expert Systems with Applications*, vol. 41, no. 16, pp. 7106–7113, 2014.

[121] G. Berkovic and E. Shafir, “Optical methods for distance and displacement measurements,” *Advan. Opt. Photon.*, vol. 4, no. 4, pp. 441-471, 2012.

[122] Z. Ji and M. C. Leu, “Design of optical triangulation devices,” *Opt. Laser Technol.*, vol. 21, no. 5, pp. 339–341, Oct. 1989.

[123] R. G. Dorsch, G. Häusler, and J. M. Herrmann, “Laser triangulation: Fundamental uncertainty in distance measurement,” *Appl. Opt.*, vol. 33, no. 7, pp. 1306–1314, Mar. 1994.

[124] S. Hertega and J. Liljencrantz, “Measurement of human vocal fold vibrations with laser triangulation,” *Opt. Eng.*, vol. 40, no. 9, pp. 2041–2044, Nov. 2001.

[125] J. H. Wu, R. S. Chang, and J. A. Jiang, “A novel pulse measurement system by using laser triangulation and a CMOS image sensor,” *Sensors*, vol. 7, no. 12, pp. 3366–3385, Dec. 2007.

[126] J. H. Wu and R. S. Chang, “No-touch pulse measurement by laser triangulation,” *Proc. Biomedical Optics*, 2005, pp. 383–390.

[127] P. Castellini, M. Martarelli, and E. P. Tomasini, “Laser doppler vibrometry: Development of advanced solutions answering to technology’s needs,” *Mech Syst. and Signal Process.*, vol. 20, no. 6, pp. 1265–1285, Aug. 2006.

[128] M. Radzieński, Ł. Doliński, M. Krawczuk, and M. Palacz, “Damage localisation in a stiffened plate structure using a propagating wave,” *Mech Syst. and Signal Process.*, vol. 39, no. 1, pp. 388–395, Aug.-Sept. 2013.

[129] N. B. Roozen, L. Labelle, M. Rychtáriková, and C. Glorieux, “Determining radiated sound power of building structures by means of Laser Doppler vibrometry,” *J. of Sound and Vibrat.*, vol. 346, pp. 81–99, Jun. 2015.

[130] H. Tabatabai, D. E. Oliver, J. W. Rohrbaugh, and C. Papadopoulos, "Novel applications of laser doppler vibration measurements to medical imaging," *Sens. Imaging*, vol. 14, no. 1-2, pp. 13–28, Jun. 2013.

[131] A. D. Kaplan, J. A. OrSullivan, E. J. Sirevaag, P. Lai, and J. W. Rohrbaugh, "Hidden state models for noncontact measurements of the carotid pulse using a laser doppler vibrometer," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 3, pp. 744–753, Mar. 2012.

[132] E. Caetano, S. Silva, and J. Bateira, "A vision system for vibration monitoring of civil engineering structures," *Exp. Tech.*, vol. 35, no. 4, pp. 74–82, Aug. 2011.

[133] H. G. Maas and U. Hampel, "Photogrammetric techniques in civil engineering material testing and structure monitoring," *Photogramm. Eng. Rem. Sens.*, vol. 72, no. 1, pp. 39–45, Jan. 2006.

[134] J. G. Chen, N. Wadhwa, F. Durand, W. T. Freeman, and O. Buyukozturk, "Developments with motion magnification for structural modal identification through camera video," *Dynam. Civil Struct., Volume 2, Proc. Int. Modal Anal. Conf.*, 2015, pp. 49–57.

[135] J. Lohscheller, U. Eysholdt, H. Toy, and M. Döllinger, "Phonovibrography: Mapping high-speed movies of vocal fold vibrations into 2-D diagrams for visualizing and analyzing the underlying laryngeal dynamics," *IEEE Trans. Med. Imag.*, vol. 27, no. 3, pp. 300–309, Mar. 2008.

[136] D. D. Mehta, D. D. Deliyski, T. F. Quatieri, and R. E. Hillman, "Automated measurement of vocal fold vibratory asymmetry from high-speed videoendoscopy recordings," *J. SPEECH LANG. HEAR. R.*, vol. 54, no. 1, pp. 47–54, Feb. 2011.

[137] A. P. Pinheiro, D. E. Stewart, C. D. Maciel, J. C. Pereira, and S. Oliveira, "Analysis of nonlinear dynamics of vocal folds using high-speed video observation and biomechanical modeling," *Digital Signal Process.*, vol. 22, no. 2, pp. 304–313, Mar. 2012.

[138] X. Dai, X. Shao, Z. Geng, F. Yang, Y. Jiang, and X. He, "Vibration measurement based on electronic speckle pattern interferometry and radial basis function," *Opt. Commun.*, vol. 355, pp. 33–43, Nov. 2015.

[139] R. Worland and B. Boe, "Measurements of coupled drumhead vibrations using electronic speckle-pattern interferometry," *J. Acoust. Soc. Amer.*, vol. 134, no. 5, pp. 4157–4157, Nov. 2013.

[140] T. Eck and S. J. Walsh, "Measurement of vibrational energy flow in a plate with high energy flow boundary crossing using electronic speckle pattern interferometry," *Appl. Acoust.*, vol. 73, no. 9, pp. 936–951, Sept. 2012.

[141] A. A. Dyrseth and S. Skatter, "Vibration analysis of logs with electronic speckle pattern interferometry," *Appl. opt.*, vol. 36, no. 16, pp. 3649–3656, Jun. 1997.

[142] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "High-speed tracking with kernelized correlation filters," *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. 37, no. 3, pp. 583–596, 2015.

[143] Z. Kalal, J. Matas, and K. Mikolajczyk, "Pn learning: Bootstrapping binary classifiers by structural constraints," *Proc. IEEE Conf. Comp. Vis. Patt. Recog.*, 2010, pp. 949–56.

[144] Z. Kalal, K. Mikolajczyk, and J. Matas, "Forward-Backward Error: Automatic Detection of Tracking Failures," *Proc. Int. Conf. Patt. Recog.*, Aug. 2010, pp. 2756–2759.

[145] H. Grabner, M. Grabner, and H. Bischof, "Real-time tracking via on-line boosting," *Proc. British Mach. Vis. Conf.*, Sep. 2006, pp. 6–11

[146] B. Baneko, M. H. Yang, and S. Belongie, "Visual Tracking with Online Multiple Instance Learning," *Proc. IEEE Conf. Comp. Vis. Patt. Recog.*, June 2009, pp. 983–990.

[147] OpenCV. (2016, October 12) OpenCV 3.0 [Online]. Available: http://opencv.org/opencv-3-0.html.

[148] H. Liu, M. T. Hong, M. Herman, T. Camus, and R. Chellappa, “Accuracy vs efficiency trade-offs in optical flow algorithms,” *Comput. Vis. Image Underst.*, vol. 72, no. 3, pp. 271–286, 1998.

[149] J. Barron, D. Fleet, and S. Beauchemin, “Performance of optical flow techniques,” *Int. J. Comput. Vis.*, vol. 12, no. 1, pp. 43–77, 1994.

Acknowledgment

Firstly, I wish to appreciate my advisor, Prof. Idaku Ishii, who provided me an opportunity to join our laboratory and led me into academic circle. The harvest of my overseas study should owe to his patient instruction both in concrete research and professional attitude. His strict attitude of work and endless passion of pursuing innovation have significantly affected my attitude facing trouble both in work and life.

Besides my advisor, I would like to express my gratitude to Dr. Takeshi Takaki, Dr. Tadayoshi Aoyama, and Dr. Qingyi Gu. Their invaluable suggestions helped me overcome the unfamiliarity with fresh experimental environment when I initially joined our laboratory.

I would also like to express my heartfelt gratitude to Ms. Yukari Kaneyuki (educational administrator), Ms. Rumi Horiuchi and Ms. Etsuko Yokoyama (laboratory secretary). During my lonely and tough overseas life in Japan, they were my most reliable staffs in our institution, I received considerate attention both in my study and life from them.

Last but not the least, I would like to express my profound gratitude to my parents. They devoted all their spiritual energy to my education in the past two decades and supported me spiritually throughout writing this thesis and my life in general.

January, 2017
Mingjun Jiang