Chapter 5
Dynamic Information Space Based on High-Speed Sensor Technology

Masatoshi Ishikawa, Idaku Ishii, Yutaka Sakaguchi, Makoto Shimojo, Hiroyuki Shinoda, Hirotsugu Yamamoto, Takashi Komuro, Hiromasa Oku, Yutaka Nakajima and Yoshihiro Watanabe

Abstract The purpose of this research is to realize a dynamic information space harmonizing human perception system, recognition system, and motor system. Toward this purpose, our key technology is high-speed sensor technology and display technology focusing on vision and haptic sense which performs at the order of kHz. Based on these technologies, our information space can obtain the dynamics of humans and objects in perfect condition and displays information at high speed. As subsystems for our goal, we have newly developed four important elemental

M. Ishikawa (✉) · H. Shinoda · Y. Watanabe
The University of Tokyo, Tokyo, Japan
e-mail: Masatoshi_Ishikawa@ipc.i.u-tokyo.ac.jp

H. Shinoda
e-mail: Hiroyuki_Shinoda@k.u-tokyo.ac.jp

Y. Watanabe
e-mail: Yoshihiro_Watanabe@ipc.i.u-tokyo.ac.jp

I. Ishii
Hiroshima University, Hiroshima, Japan
e-mail: iishii@robotics.hiroshima-u.ac.jp

Y. Sakaguchi · M. Shimojo · Y. Nakajima
University of Electro-Communications, Tokyo, Japan
e-mail: sakaguchi@is.uec.ac.jp

M. Shimojo
e-mail: shimojo@mce.uec.ac.jp

Y. Nakajima
e-mail: nakajima@hi.is.uec.ac.jp

H. Yamamoto
Utsunomiya University, Tochigi, Japan
e-mail: hirotsugu_yamamoto@cc.utsunomiya-u.ac.jp

T. Komuro
Saitama University, Saitama, Japan
e-mail: komuro@mail.saitama-u.ac.jp

H. Oku
Gunma University, Gunma, Japan
e-mail: h.oku@gunma-u.ac.jp

T. Nishida (ed.), *Human-Harmonized Information Technology, Volume 1*,
DOI 10.1007/978-4-431-55867-5_5

technologies including high-speed 3D vision toward insensible dynamics sensing, high-speed resistor network proximity sensor array for detecting nearby object, non-contact low-latency haptic feedback, and high-speed display of visual information for information sharing and operation in real space. Also, in order to achieve the coordinated interaction between individual humans and this information space, we have conducted the research about the human perceptual and motor functions for coordinated interaction with high-speed information environment. In addition, we have developed various application systems based on the concept of dynamic information space by integrating the subsystems.

Keywords High-speed vision · Proximity sensor · Airborne Ultrasound Tactile Display (AUTD) · High-speed LED display · Human interface · Dynamic information environment

5.1 Introduction

In this research, we aim at realizing a dynamic information space harmonizing human perception system, recognition system, and motor system. Toward this goal, we focus on the acquisition of the human and object dynamics perfectly in a information space and the high-speed display whose performance is dramatically improved. Integrating these functions, the platform of the new information space become possible to be built.

However, the conventional visual information sensing and display at video rate (30 Hz, for example) does not have enough performance for the dynamics involved in the high-speed human/object motion. Similarly, the sensing and display technologies of the haptic information does not have enough speed. Also, haptic technologies have a limit requiring physical contact between device and target. In order to allow the high-speed dynamics in the information space, it is essentially required to achieve the non-contact and unrestricted sensing and display.

Therefore, we have developed four new technologies including high-speed 3D vision toward insensible dynamics sensing, high-speed resistor network proximity sensor array for detecting nearby object, noncontact low-latency haptic feedback, and high-speed display of visual information for information sharing and operation in real space.

On the other hand, information provided by the sensing and display exceeding the speed of human perceptual ability is difficult to be utilized for human. In order to overcome this limit, we have conducted the research about the human perceptual and motor functions for coordinated interaction with high-speed information environment. This allows us to realize the coordinated interaction between individual humans and this information space.

As a final goal, we have developed various types of dynamic information space, which are mainly for the computer-human interfaces, by integrating these technologies. This integrated system allows high-speed user interaction in a contactless manner without any constraints about the moving humans and targets.

5.2 High-Speed 3-D Vision Toward Insensible Dynamics Sensing

High-speed vision systems that can capture and process real-time imagery at hundreds or thousands of frames per second (FPS) are an important step toward the realization of harmonized dynamic information environments. These systems are powerful sensing tools for detecting "insensible dynamics," which the human eyes can only barely sense. To completely capture rapid human motion in three-dimensional (3-D) space with only minor occlusions, multiple depth images in different views must be simultaneously captured and processed at a high frame rate. In this section, we introduce structured-light high-speed 3-D vision systems that can capture and process depth images containing 512 × 512 pixels in real time, at 500 FPS on high-frame-rate (HFR) camera-projector systems.

Our structured-light high-speed 3-D vision system contains a 3-D module consisting of a high frame rate (HFR) camera head, an HFR projector, an IDP Express board [1], and a personal computer (PC) equipped with a GPU board. The 3-D module uses a Digital Light Processing (DLP) development kit projector for HFR projection, which is based on digital micromirror device (DMD) technology (Texas Instruments Inc., US), and a monochrome camera head (Photron Ltd., Japan). A system overview in which the DLP LightCommander 5500 is used as a projector is shown in Fig. 5.1. On a level surface 550 mm below the camera, depth information over a 484 × 484 mm square is captured as a 512 × 512 image.

The DLP LightCommander 5500 can project hundreds of 1024 × 768 binary patterns at 1000 FPS or greater. The IDP Express was designed to implement various image-processing algorithms, and to record images and features at a high frame rate onto PC memory. The camera head captures 8-bit gray-level 512 × 512 images at 2000 FPS. The IDP Express board has two camera inputs, along with a

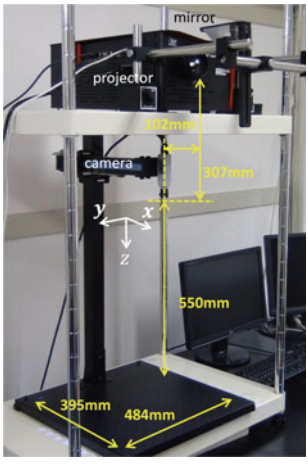

Fig. 5.1 System overview

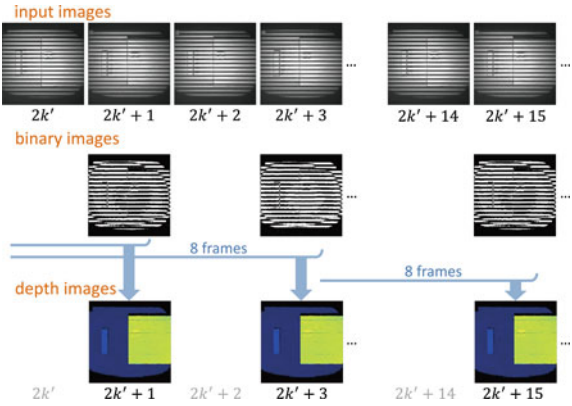

Fig. 5.2 Pipelining-output of depth images

frame-straddling function and trigger I/Os for external synchronization. Two 512 × 512 images and their processed results can be mapped onto PC memory at 2000 FPS, via a Peripheral Component Interconnect Express (PCI-e) bus. The Tesla C1060 is a computer processor board based on the NVIDIA Tesla T10 GPU. It is capable of a processing performance of 933 Gflops/s, using 240 processor cores operating at 1.296 GHz and a bandwidth of 102 GB/s for its internal 4 GB memory. We use a PC with the following specifications: ASUSTeK P6T7 WS main board, Intel Core i7 3.20 GHz CPU, 3 GB RAM, two 16-lane PCI-e 2.0 buses. To compute depth images at a high frame rate with minimal synchronization errors, a motion-compensated structured-light algorithm [2], which is based on Inokuchi's method [3], was implemented; binary light patterns coded with an 8-bit gray code are projected at 1000 FPS, and the projected light patterns are captured at 1000 FPS. Depth image processing was accelerated using parallel processing with 512 blocks of 1 × 512 pixels on the GPU board; the total time is 1.81 ms, and depth image processing of 512 × 512 images can be conducted in real time at 500 FPS (Fig. 5.2).

Figure 5.3 shows the 3-D measurement results for a human hand moving periodically; (a) shows the experimental scenes captured using a standard camera, (b) shows the x, y, and z coordinates of the right hand, (c) depicts the depth images measured by our motion-compensated structured-light method with an 8-bit gray code using a 1000 FPS video (denoted by MCGC1K), (d) shows the 8-bit gray-code structured-light method without motion compensation [4] using a 1000 FPS video (GC1K), and (e) depicts 30 FPS video (GC30). The hand was moved horizontally in a circular orbit at a certain distance from the desk plane at a frequency of once per second. On the desk plane, a computer keyboard, books, and many 3-D objects were placed as background objects, and the left hand was kept stationary. In Fig. 5.3b, the centroid position of the right hand, which was computed by subtracting the background from the MCGC1K depth images, was periodically changed at a rate of once per second. Compared with the MCGC1K and GC1K depth images, the GC30 depth images were incorrect; this occurred because synchronization errors generated by 3-D measurements of moving objects (using different frames) increase when the frame interval is increased.

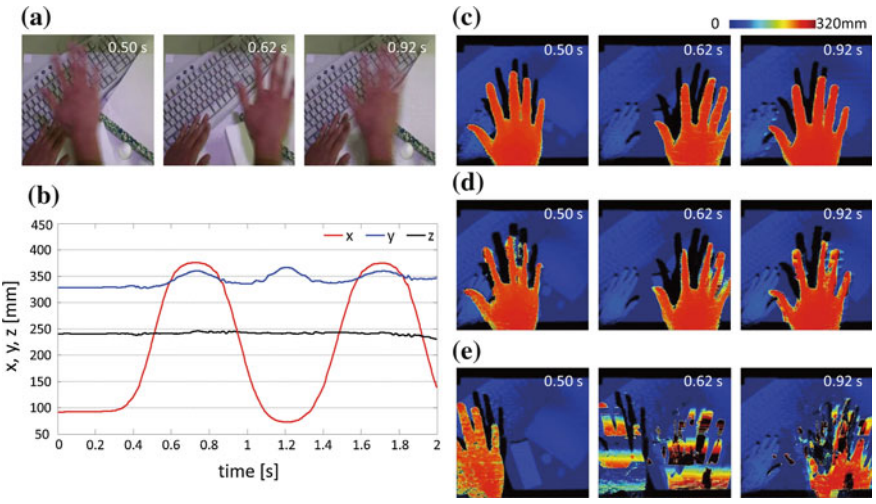

Fig. 5.3 3-D measurement of a moving human hand. **a** Experimental scenes, **b** xyz coordinate values, **c** depth images (MCGC1K), **d** depth images (CGC1K), **e** depth images (GC30)

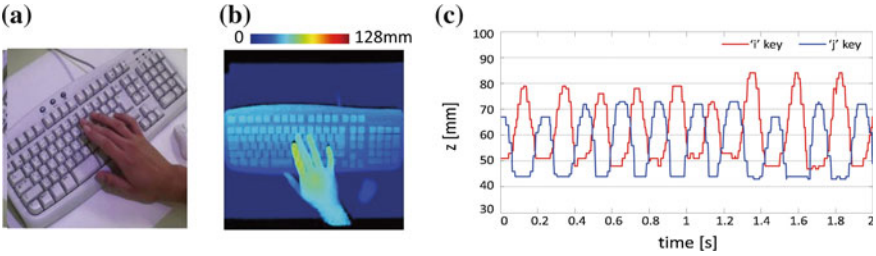

Fig. 5.4 3-D measurement of finger-tapping on a computer keyboard. **a** Experimental scenes, **b** depth image, **c** temporal images of tapped keys

Synchronization errors were significantly minimized by introducing an HFR camera-projector system. Compared with the GC1K depth images, the MCGC1K depth images show that the 3-D shape of the human hand was accurately measured with minimal synchronization errors when the hand was moving. Figure 5.4 shows (a) the experimental scene, (b) the MCGC1K depth image, and (c) the depth information of the "*j*" key and "*i*" key, when the forefinger of the right hand taps the "*j*" key and its middle finger taps the "*i*" key alternatively at a 5 Hz frequency. Our 3-D vision system can detect high-speed finger motions, and the timing of each key tapping, to detect the input content.

By adding an RGB projector to expand the HFR camera-projector system, pixel-wise projection mapping can be conducted onto time-varied 3-D scenes. Infrared (IR) light patterns projected from an IR HFR projector are simultaneously captured and processed for depth image calculation, and the RGB light patterns are interactively generated and projected from the RGB projector onto the 3-D scene. The IR and RGB

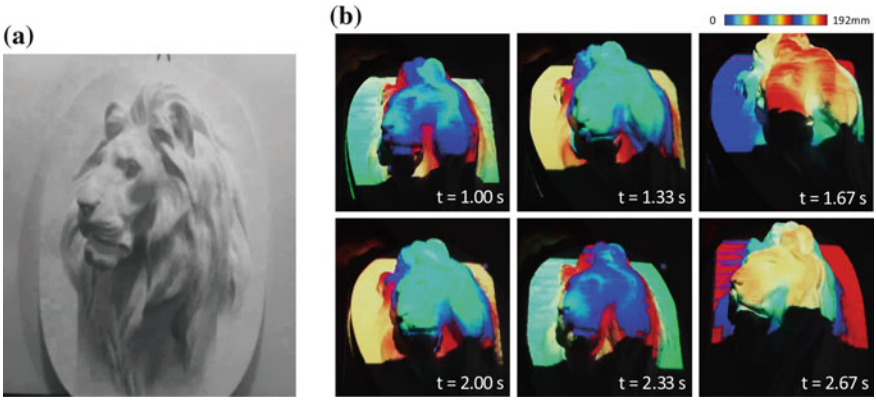

Fig. 5.5 Pixel-wise projection mapping results for a moving lion relief. **a** Plaster lion relief, **b** color-mapped scenes

projectors have the same projection fields. A camera head with an IR wavelength filter can capture only IR light patterns for the 3-D structured-light measurement when the RGB light patterns are projected for enhanced tasks. Figure 5.5 shows the experimental scenes captured using a standard camera, when depth-based color mapping with a cyclic jet color map was conducted onto a 10 cm-deep plaster lion relief (for sensitive and distinct depth visualization). The relief was moved with periodic up-and-down motions and slight rotations by a human hand. It can be seen that the white-surface relief was enhanced by pixel-wise projection mapping with a cyclic jet color map, which can directly visualize its detailed height information for human eyes. Such projection mapping techniques based on high-speed 3-D vision will extend augmented reality (AR)-based applications for dynamic human computer interactions.

To conduct complete 3-D information acquisition with minimal occlusion using multiple camera-projector modules, time division multiplex 3-D structured-light measurement was implemented on an HFR camera-projector system. The timings of light pattern projection and image capture are straddled using a short time delay; this enables each camera-projector system to simultaneously obtain 3-D information in its view field without crosstalk between light patterns projected from different camera-projector systems. Figure 5.6 shows a prototype system for time division multiplex 3-D structured-light measurements; it is composed of two opposing HFR camera-projector modules, an IDP Express board, and a PC equipped with a Tesla C1060 GPU board. The camera-projector module consists of an HFR projector of 854×480 pixels (DLP LightCrafter 4500; Texas Instruments Inc., US) and an HFR color camera head of 512×512 pixels (Photron Ltd., Japan). In the present implementation, both camera-projector modules are operated with 0.5 ms-exposure image capture and 0.7 ms-duration light pattern projection at 500 FPS, whereas the two modules are straddled with a 1-ms time delay. The state of each module is alternatively switched at 1 ms intervals, to reduce interference from the light-patterns pro-

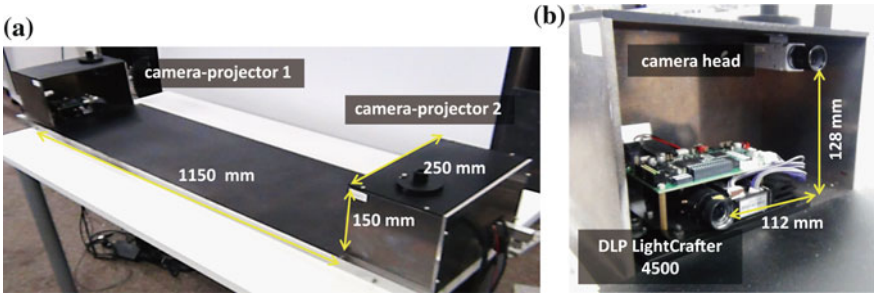

Fig. 5.6 Time division multiplex HFR camera-projector system. **a** Experimental setting, **b** camera-projector module

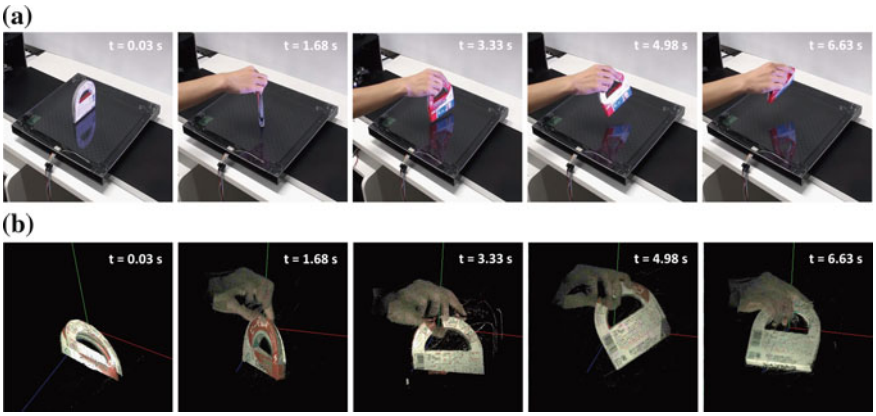

Fig. 5.7 Experimental results when a color-patterned object is moved by a human hand. **a** Experimental scenes, **b** texture-mapped 3-D scenec

jected by the other module. Accelerated by parallel processing on the GPU board, the 512 × 512 pixel depth and color RGB images are computed in 1.43 ms for each camera input; both the left- and right-side-view depth images can be simultaneously processed at 250 FPS on our time division multiplex HFR camera-projector system. Figure 5.7 shows (a) the experimental scenes and (b) the synthesized 3-D scenes when a color-patterned tape box was moved by a human hand. The 3-D scenes that are texture-mapped with color images are synthesized using the left- and right-side-view depth images, and displayed using the OpenGL environment. It can be seen that both sides of the 3-D shapes of the color-patterned tape box and human hand were displayed with minimal occlusion in real time when the human hand was rapidly moving.

To enlarge the view fields in 3-D structured-light measurement, which would facilitate more accurate 3-D image acquisitions, the number of camera-projector modules can be increased by connecting many 3-D structured-light measurement systems with short time delays via a TCP-IP network. This extension can be conducted without decreasing the acquisition rate of depth images; however, the exposure time used

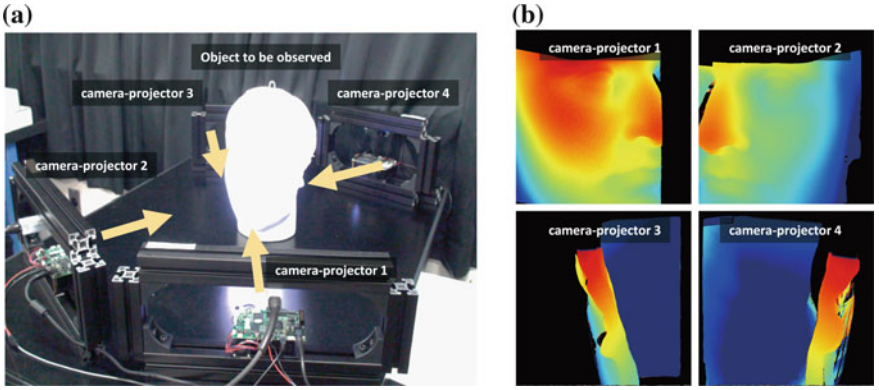

Fig. 5.8 3-D structured-light measurement using four HFR camera-projector modules. **a**Experimental setting, **b** measured depth images

for image capturing and the duration time used for light pattern projection should both be reduced, in inverse proportion to the number of camera-projector modules. Figure 5.8 shows (a) the experimental setting and (b) the measured depth images from different view angles, when a planar head sculpture was observed using four HFR camera-projector modules. Synchronizing the four camera-projector modules with 0.5, 1.0, and 1.5 ms time delays, the four different-view-angle depth images were simultaneously computed at 250 FPS with no crosstalk from light patterns projected from different angles.

Such a structured-light 3-D vision system using HFR camera-projector modules allows the simultaneous detection and localization of dynamic behaviors. It enables complete 3-D information acquisition with minimal occlusion for human-computer interactions; standard cameras operating at dozens of FPS are generally unable to perform such acquisitions. High-speed 3-D vision technology will play a role as one of the most important dynamic sensing technologies in next-generation dynamic information environments.

5.3 High-Speed Resistor Network Proximity Sensor Array for Detecting Nearby Object

This research focuses on the development of high-speed proximity sensor array that excels at close-range sensing from contact to several tens of centimeters. The sensor uses a photo-reflector comprising a light-emitting diode (LED) and a photo-transistor as its detection element. Photo-transistors capture reflected infrared light from the LED, and the position of a proximal object is inferred according to the photo-current distribution. A primary feature of this sensor is that object position inference can be performed using analog computation by a resistor network circuit [5]. Serially

reading the response levels of individual photosensitive elements, as in the case of a general-purpose CCD image, requires complex wiring due to the number of elements and the size of the unit. Moreover, data acquisition and processing times are long. On the other hand, when using proposed sensor, high-speed response (less than 1 ms) and simple wiring requirements are retained, even when the number of elements is increased to accommodate the form of the installation surface. This is an important consideration for application to a complex shape, such as a grip designed to fit the human hand.

5.3.1 Sensor Design and Detection Principles

We refer to the developed azimuth and elevation proximity detector as a dome-shape sensor. The dome-shape sensor is formed from two sensor ring (Fig. 5.9) [6]. Each of the two ring can obtain one-dimensional positional coordinates. Azimuth is determined by orthogonally aligning each sensor's coordinate axis. Each ring can also measure the distance to target objects. Elevation is determined by placing rings in two layers in the height direction, and performing distance detection with each.

This section describes the one-dimensional resistor network proximity sensor array (RNPS), which is the most fundamental part of the dome-shape sensor. Figure 5.9 shows the overview and the circuit structure. The one-dimensional RNSP consists of a resistor network. The resistor network is formed from $n-1$ internal resistors $r_i\,(i = 1 \sim n-1)$ between the photo-reflectors, and an external resistors R_0 that connect the terminals on both ends V_{S1}, V_{S2} and a negative supply $V^{[-]}$. Light from an infrared LED reflects off of objects proximal to the sensor, and is collected in the photo-transistor. This causes photocurrent distribution that corresponds to the distribution of reflected light at the phototransistor surface. When a photocurrent I_i occurs at the ith element, the currents flowing between terminals V_{S1} and V_{S2} are described as follows:

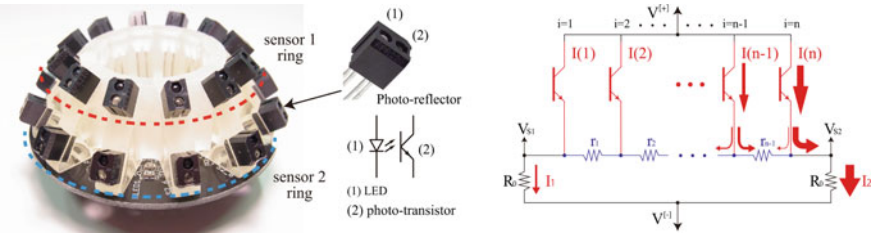

Fig. 5.9 *Left* The dome-shape sensor is formed from two ring of one-dimensional resistor network proximity sensor array (RNPS). The rings are placed in two layers in the height direction. *Right* Circuit diagram of the one-dimensional RNPS: When a photo current occurs at the element, the currents diverges and flows to the negative supply $V^{[-]}$ [6]

$$I_{i1} = \frac{\sum_{k=i}^{n-1} r_k + R_0}{\sum_{k=1}^{n-1} r_k + 2R_0} I_i, \quad I_{i2} = \frac{\sum_{k=1}^{i-1} r_k + R_0}{\sum_{k=1}^{n-1} r_k + 2R_0} I_i \tag{5.1}$$

The difference between the currents flowing to terminals V_{S1} and V_{S2} is therefore

$$I_{i1} - I_{i2} = \frac{\sum_{k=i}^{n-1} r_k - \sum_{k=0}^{i-1} r_k}{\sum_{k=1}^{n-1} r_k + 2R_0} I_i \tag{5.2}$$

The numerator in Eq. (5.2) is the product of the resistance at each element from the center of the serially connected resistor network with the current flowing through it, and represents the one-dimensional moment of the current at the center. When a photocurrent arises in the each element, the photocurrents will therefore flow together, and the following equations will hold:

$$\sum_{i=1}^{n-1} (I_{i1} - I_{i2}) = \sum_{i=1}^{n-1} I_i \frac{\sum_{k=i}^{n-1} r_k - \sum_{k=1}^{i-1} r_k}{\sum_{k=1}^{n-1} r_k + 2R_0} = \frac{V_{S1} - V_{S2}}{R_0} \tag{5.3}$$

$$\sum_{i=1}^{n-1} I_i = I_{all} = \frac{V_{S1} + V_{S2} - 2V^{[-]}}{R_0} \tag{5.4}$$

Removing the one-dimensional moment calculated by Eq. (5.3) from the overall current I_{all} calculated by Eq. (5.4) allows identification of the photocurrent distribution's central point. To improve ease-of-use, Eq. (5.5) is used on the one-dimensional RNPS output to normalize the center point to the range $[-1, 1]$.

$$x_c = -\left(1 + \frac{2R_0}{\sum_{i=1}^{n-1} r_i}\right)\left(\frac{V_{S1} - V_{S2}}{V_{S1} + V_{S2} - 2V^{[-]}}\right) \tag{5.5}$$

When the amount of reflected light gathered by the phototransistor changes according to the distance between the sensor and the proximal object, the total current through the resistor network also changes. This allows estimation of the approximate distance to the object by using Eq. (5.4) to obtain the total current through the resistor network.

5.3.1.1 Azimuth Detection

The azimuthal orientation of the object is detected by deriving cosine and sine values from the one-dimensional RNPS output. Figure 5.10a shows the circular arrangement of the sensor elements of the one-dimensional RNPS. The internal resistors in the

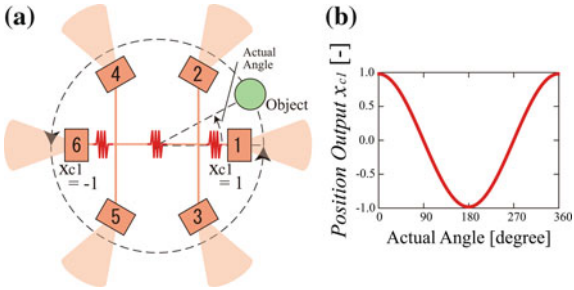

Fig. 5.10 Structure and output of the circular one-dimensional RNPS: The positioning output forms a cosine wave for an object at the outer sensor periphery by the circular arrangement of the sensor elements and the internal resistors set [6]. **a** Structure of sensor 1, **b** output of sensor 1

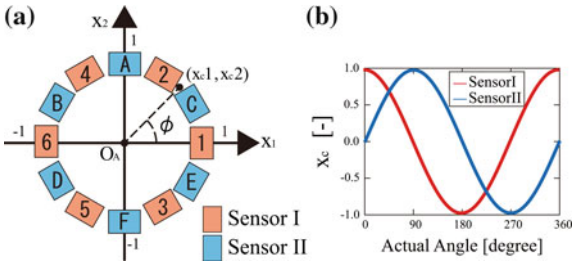

Fig. 5.11 Detection principle of azimuth: The orthogonally orienting two circular sensor allows derivation of the cosine and sine value of the object's azimuth from these positioning output [6]. **a** Structure of sensor, **b** output of sensor

proximity sensors are set so that positioning output forms a cosine wave when an object goes around the sensor, as shown in Fig. 5.10b. In the case where six elements are used (Fig. 5.10), the ratios of internal sensors will be 1:2:1 in proportion to real-space positions. Two such circular one-dimensional RNPS are used, and are respectively called sensor I and sensor II. As Fig. 5.11 shows, orthogonally orienting these sensors creates a 90° phase shift between them, allowing derivation of the sine value of the object's azimuth from the positioning output. The positioning output (x_{c1}, x_{c2}) obtained by the respective sensors and Eq. (5.6) can therefore uniquely determine the object's azimuth ϕ.

$$\phi = \begin{cases} \frac{\pi}{2}\mathrm{sgn}(x_{c2}) & x_{c1} = 0 \\ \arctan\left(\frac{x_{c2}}{x_{c1}}\right) & x_{c1} > 0 \\ \pi\,\mathrm{sgn}(x_{c2}) + \arctan\left(\frac{x_{c2}}{x_{c1}}\right) & x_{c1} < 0 \end{cases} \tag{5.6}$$

5.3.1.2 Elevation Detection

The two circular one-dimensional RNPS provide two distance outputs I_{all}, which are used to detect elevation without compromising the azimuth detection characteristics. The two circular RNPS are placed as shown in Fig. 5.12a. Because changes in elevation will change the distance to each of the sensors, the photocurrent flowing

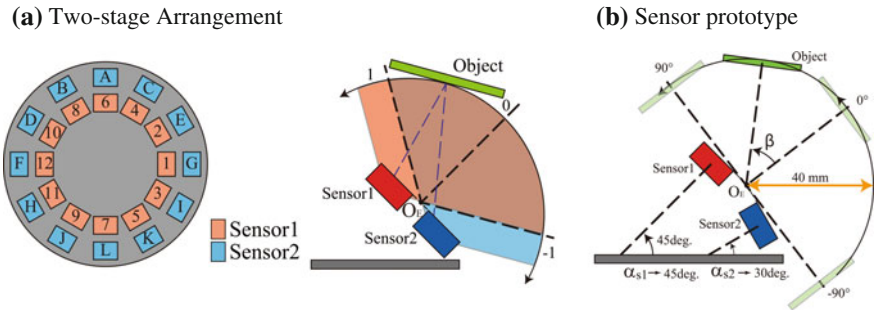

Fig. 5.12 Detection principle of elevation: **a** The two circular sensors are placed in two layers in the height direction. The changes in elevation will change the distance to each of the sensors. **b** Overview of simulation operation for determining element tilts: The angle at which the normalized elevation output is 0 is taken as 0°, and a 90 mm square object was moved through the space between −90° and 90° with respect to the origin of the elevation angle O_E [6]

through sensors 1 and 2 will also change. Specifically, taking the distance output of the two circular RNPS as I_{all1}, I_{all2} allows use of Eq. (5.7) to find the elevation output θ, normalized to the range [−1, 1].

$$\theta = \frac{I_{all1} - I_{all2}}{I_{all1} + I_{all2}} \tag{5.7}$$

The normalized elevation output will be +1 if only sensor 1 is responding, −1 if only sensor 2 is responding, and continuously varying in the range $-1 < \theta < 1$ in the case where both sensors are responding. In this method, the range at which elevation can be detected depends on the tilt and positioning of the two circular RNPS. This allows the detection range to be set according to the intended purpose of the dome-shape sensor.

5.3.2 Prototyping and Experiments

5.3.2.1 Sensor Prototype

When designing a dome-shape sensor, it is possible to adjust element tilt and positioning according to the intended application. Here we developed equipment for the purpose of tracking a human hand. We identified the following design requirements:

- The elevation detection range must include the space to the sides and above the dome-shape sensor.
- The sensor must be shaped to fit a human hand, based on a circular form with approximately 65 mm diameter.

- Detected objects would primarily be human palms, and thus an approximate planar square of $90 \times 90\,\text{mm}^2$.

We varied the tilt of the elements in sensors 1 and 2 (α_{s1}, α_{s2}, respectively), and determined the element positioning best suited to our basic circular form. Because it would be difficult to create and test prototype sensors with a variety of tilt combinations, we used a RNPS detection simulator developed in our lab.

As a result of the simulations, Fig. 5.12b shows the side view of the dome-shape sensor prototype. Twelve elements were evenly spaced along the dome sensor's periphery.

5.3.2.2 Detection Characteristics

(1) Azimuth Detection Characteristics

In this section, we examine the azimuth detection characteristic by an experiment with the prototyped dome-shape sensor. The distance between the sensor and the object was 30–60 mm.

Figure 5.13a shows the results of the experiment.

The figure indicates that the azimuth output of the sensor increased monotonically with increasing azimuth of the object.

(2) Elevation Detection Characteristics

Figure 5.13b indicates that the absolute value of elevation output near 0° or 90° increases with increasing distance.

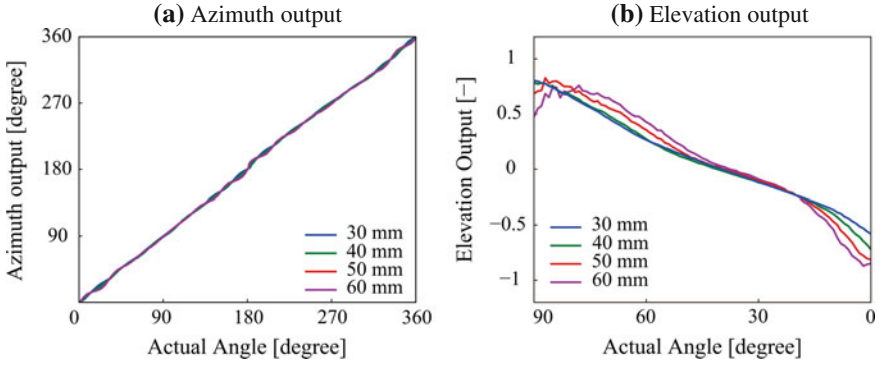

Fig. 5.13 **a** Experimental results with changing the distance of 30–60 mm. Detection is possible over the entire 360° range. Detection error was within approximately ±5°. **b** Elevation output of the experimental results with changing the distance of 30–60 mm

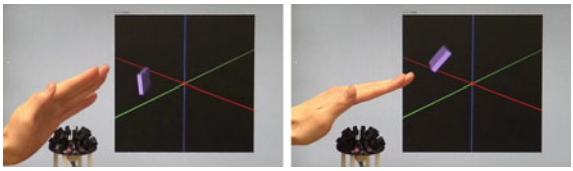

Fig. 5.14 Human-machine interface usage of the dome-shape sensor: The sensor with the rapid response features can be applied to non-contact human interface applications. This photo shows a detected palm position [6]

The elevation output difference due to distance change decreases with decreasing distance between the sensor and the object. Thus, incorporating the dome-shape sensor into the grip will allow tracking of hands by using the result in proximity as target.

5.3.3 Conclusion

The present study developed a high-speed proximity sensor array for simultaneous detection of azimuth and elevation. The features of the proposed dome-shape sensor include rapid responsiveness and simpler wiring, while maintaining a 360° sensing range and detection from sensor sides to top. The rapid response features of the method can be applied to noncontact human interface applications (Fig. 5.14). In future research, we plan to create more practical dome-shape sensors with wider detection range on the same principle, and investigate potential applications.

5.4 Noncontact Low-Latency Haptic Feedback

One of the key components of the Dynamic Information Space is a haptic display without constraining users' behavior. Haptics would be an indispensable modality in the computational support of human dynamic motion though it is still in an early stage of the applications. In order to indicate the motion direction, notify the motion timing, and comprehend the circumstance in fast human motions, haptics would be a promising modality to transmit such information with acceptable delays as realtime feedbacks. A problem in haptic feedbacks had been physical stimulation to human skins. If the device is a bulky mechanical system installed on the ground or a table, the workspace is limited to a narrow area around the device. It is also difficult to receive passive stimulation in free motions using such mechanical systems. Therefore, the recent technological main stream is to develop small and lightweight haptic devices wearable or installed in a mobile device. Based on the recent technological achievement, haptics is beginning to find wide ranging applications as computer interface.

However, the problem that still remains is the area to receive haptic stimulation is limited and localized in practical scenes. It is not straightforward to stimulate various body parts to support the motions. And it is also a problem that such devices should be worn in advance before haptic feedbacks are necessary, which cannot be supposed in some applications.

With this background, we developed a noncontact tactile display that produces tactile sensation using ultrasound traveling in the air [7]. Airborne Ultrasound Tactile Display (AUTD) was first proposed by the authors and demonstrated with a small ultrasound phased array [8, 9]. In this project, we extended this to a large aperture phased array to widen the workspace. In the following subsection, we describe the system design to realize such large aperture and effective tactile display. The total system combined with sensors and visual displays will be described in the later section.

5.4.1 Remote Vibrotactile Sensation Produced with Airborne Ultrasound

There are some options to generate tactile sensations in noncontact manners. Air flow produced by a propeller or jet nozzle would be a feasible method to stimulate the human skin remotely. Very strong infrared would induce temperature elevation when it is radiated on the skin. The features of ultrasound stimulation are:

1. A small spot comparable to the sound wavelength can be selectively stimulated. The distance from the sound source can be as large as the aperture of the phased array keeping the convergence,
2. Low frequency ultrasound ~40 kHz can propagate a long distance (with −1 dB/m at 40 kHz),
3. Ultrasound amplitude can be modulated at a frequency as high as 1 kHz. The force produced by the ultrasound radiation pressure on the skin can be controlled with 1 kHz bandwidth.

Though the maximum pressure would be limited to 100 mN practically and a large aperture ultrasound transducer array is necessary, the above features are desirable for non-constrained high speed tactile feedback. A problem in principle is that time delay by the sound velocity is inevitable. For 340 m/s sound velocity, the stimulation delay is as large as 3 ms/m. But for many applications, this delay would be acceptable.

The key physical phenomena called radiation pressure is a nonlinear acoustic phenomenon, to convert the alternating sound pressure amplitude into static pressure on the reflection surface. The details are described below.

5.4.1.1 Acoustic Radiation Pressure

Physics showed sound waves are accompanied with radiation pressure proportional to the energy density of the sound [10]. Since ultrasound with a short wavelength can be localized in a small area, it can create a concentrated radiation pressure on a skin. The radiation pressure applied to the surface reflecting an ultrasound is given as

$$P = \frac{\alpha p^2}{\rho c^2} \tag{5.8}$$

where p [Pa] denotes the effective value of sound pressure, c [m/s] the sound velocity in the medium, ρ[kg/m^3] the density of the medium, and α a coefficient determined by the reflectance. When ultrasound propagates in the air and blocked off by liquid or solid, almost all of the ultrasound is reflected on the boundary and in this case α becomes 2 in vertical incidence. Following this equation, we can control temporal profiles of radiation pressure by amplitude modulation of ultrasound pressure. The attenuation of ultrasound propagation in the air depends on its frequency. The attenuation rate of 40 kHz ultrasound in air is about 1 dB/m, while the frequency is much higher than the highest frequency that human can feel as vibrotactile stimulation ($\sim$1 kHz) [11].

5.4.1.2 Controlling Radiation Pressure

The Airborne Ultrasound Tactile Display (AUTD) in this project is an ultrasound transducer array with 10 mm period [8, 9]. By setting a proper phase shift on each transducer, the three-dimensional position of the focal point can be controlled. Let $\mathbf{r}_i$ be the position of the ith transducer among N transducers and $\mathbf{r}_F$ be the desirable focal position. The phase shift on i th transducer is calculated so that it compensates the phase delay through distance:

$$\theta_i = k|\mathbf{r}_i - \mathbf{r}_F|, \tag{5.9}$$

where k denotes the wavenumber. In the prototype, a 1.5-cm-diameter spot of radiation pressure can be formed around the array. The spot center can be moved in accuracy of 1 mm horizontally. Its position can be switched at a refreshing rate of over 2 kHz. The whole phase calculation can be done within 50 μs in the system, which is faster than 2 kHz.

The ultrasound amplitude is controlled with pulse width modulation. The pulse frequency is fixed to the transducer's resonant frequency $f = 40$ kHz and the amplitude is controlled with the pulse width d [s] as

$$p = p_0 \sin(\pi d / T) \tag{5.10}$$

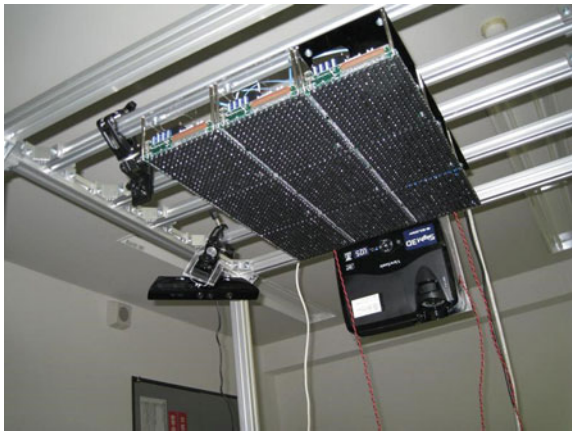

Fig. 5.15 Airborne ultrasound tactile display with 3 by 3 units [7]. The size of a unit is $19 \times 15\,\mathrm{cm}^2$

where $T = 1/f$, and p_0 denotes the maximal amplitude for $d/T = 1/2$ [9]. The pulse width $d(t)$ is calculated so that the radiation pressure P becomes the target value. It should be noticed that the radiation pressure P always has a positive value (Fig. 5.15).

5.4.2 Multi-unit Phased Array Scheme

In this project, we developed a freely extendable phased array system [7]. Extending the phased array aperture size enlarges the workspace where focusing is ensured. We design an AUTD unit of $19 \times 15\,\mathrm{cm}^2$ having a serial input port and three output ports as shown in Fig. 5.16. Each unit has FPGAs that calculate each signal phase of the transducer on it. The serial signal receiver is programmed in the Master FPGA. Master FPGA receives the focal point position (x, y, z) and the amplitude p, and transmit the information to the next unit. Since the signal delay between the neighbor units is sufficiently small ($\leq$50 ns), multiple units can be connected freely. Since it has three output ports, $1 + 3 + 3^2 + 3^3 + 3^4 = 121$ units can be connected within 200ns delay.

Each AUTD unit must know its position in the global coordinates and its own posture in advance. This information is at first sent to the Master FPGA of each AUTD unit. Since all the focal information sent is in the global coordinates, the Master FPGAs convert it to their own local coordinates, which the Slave FPGAs receive. Slave FPGAs calculate the phase delays and generate all driving signals on transducers according to the focal position and the amplitude.

The duty cycle of 40 kHz pulse $d(t)$ can be quantized up to 640 levels. As the result, the radiation pressure $P(t)$ is quantized up to 320 levels. The current system can switch the duty cycle fast enough to generate vibrotactile sensation with the temporal resolution of 0.5 ms (2 kHz).

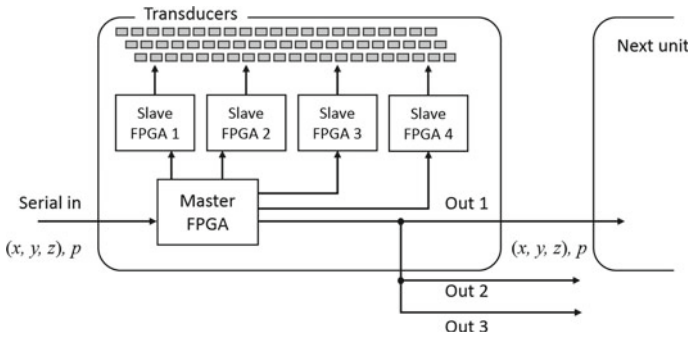

Fig. 5.16 System architecture of an AUTD unit

5.4.3 Experiments

We fabricated a phased array composed of 3×3 units as shown in Fig. 5.15 and measured the spatial distribution of generated ultrasound amplitude, in order to confirm ultrasound focusing. The 9-unit AUTD was mounted on an aluminum cabinet so that the transducer surface faced the ground, parallel to the ground. A standard microphone (B&K Type 4138) with a pre-amplifier (B&K Type 2670) was mounted on the 3D stage, with the xy-plane parallel to the AUTD. The xy-coordinates corresponded to the lattice of transducers and the z-axis was vertical to the AUTD surface. The recorded voltage was amplified with a power amplifier (B&K Type 5935). We used the reference sound source that provides 94 dB sine wave of 1 kHz for the mapping of the recorded voltage to sound pressure. The frequency characteristic of the whole recording system was almost flat from 0 to 40 kHz. The total array size was $576 \times 454.2\,\mathrm{mm}^2$.

The focal point was set to $(0, 0, -600\,\mathrm{mm})$. We measured sound pressure near the focal point and estimated the produced radiation pressure from Eq. (5.8). The parameters were set as $c = 340\,\mathrm{m/s}$, $\rho = 1.18\,\mathrm{kg/m}^3$ and $\alpha = 2$. The left figure of Fig. 5.17 depicts the calculated radiation pressure distribution in the plane at $z = -600\,\mathrm{mm}$. High radiation pressure can be seen localized within the diameter of $10 \sim 15\,\mathrm{mm}$. The right figure of Fig. 5.17 shows a 1D distribution across the focal point along $(y, z) = (0, -600)$. Focusing is clearly observed.

5.4.4 Conclusion

A non-contact vibrotactile display with a wide workspace was developed and its performance was examined. The developed AUTD is extendable by connecting multiple ultrasound transducer units and we constructed an array of $576 \times 454.2\,\mathrm{mm}^2$ aperture. The system succeeded in producing highly localized vibrotactile sensations

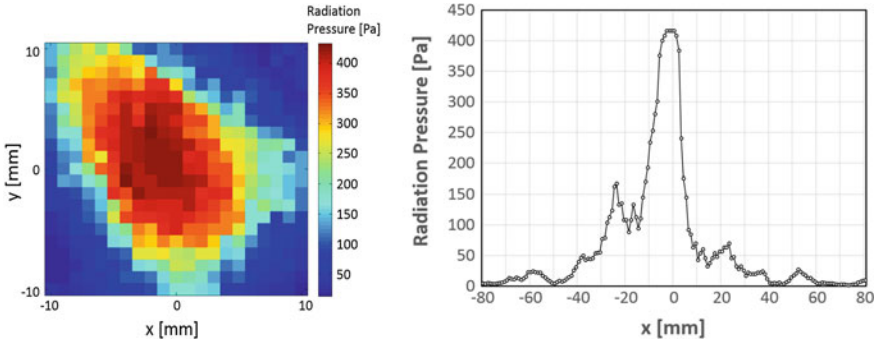

Fig. 5.17 Experimental results [7]. Distribution of acoustic radiation pressure measured at 60 cm from a 3 × 3 unit AUTD

on human skin 600 mm apart from the device. The focal intensity was experimentally demonstrated to be 74 mN. Temporal profiles of vibrotactile sensations were programmable with a sampling rate of 2 kHz and 320-level quantization.

5.5 High-Speed Display of Visual Information for Information Sharing and Operation in Real Space

Recent requirements for information displays include not only resolution but also frame rate and latency because interactive interfaces must show the responses without noticeable delay for users. Light emitting diode (LED) is a prospective light source for high-speed operations. Instantaneous operation of visual information in real space is considered to be one of the essentials for human-harmonized information technology. In this section, high-frame-rate and low-latency LED displays are described in Sects. 5.5.1 and 5.5.2. In Sect. 5.5.3, we describe aerial imaging technique that enables floating LED screen with wide viewing angle.

5.5.1 High-Frame-Rate LED Display

In current LED display systems, a large LED screen is constructed by tiling LED units. The control signals for LED units are given by an LED video processor, which distributes an input image into image data to tiled LED units. In order to transmit high-frame-rate (HFR) images via a current digital video interface, we introduced a spatiotemporal coding, as shown in Fig. 5.18. Four sub-fields pixel values are spatially arranged into quad pixels in a frame. The coded image signal is transmitted into the

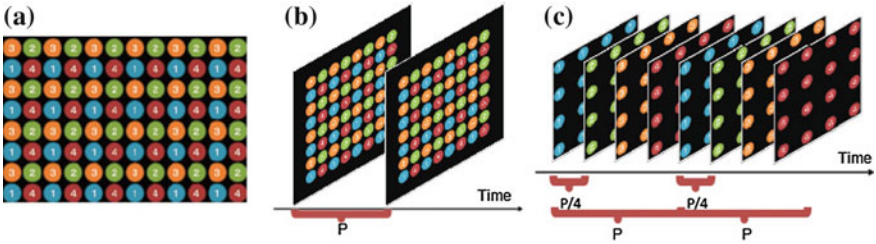

Fig. 5.18 **a** Composition of a coded image. **b** Spatiotemporally coded image signal is input into a video processor at an interval P (=8.3 ms for 120 Hz). **c** The spatiotemporal codes are decoded by the video processor. Sub-frames are re-freshed at the quadrupled rate (480 fps) of the input DVI signal

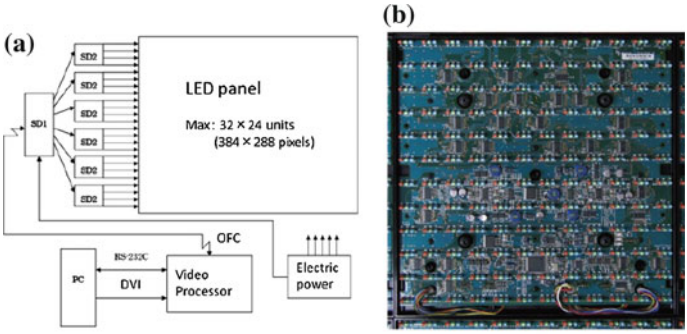

Fig. 5.19 **a** Composition of 480-fps LED display system. SD denotes signal distributor. OFC is an optical fiber cable. **b** A photograph of an LED unit

LED video processor at the refresh rate of the graphic card. When using 120 Hz DVI signal, the interval is 8.3 ms. The LED video processor decodes the spatial codes into temporal displayed images. The decoded image data are fed to each LED unit at the quadrupled frame rate of the input DVI signal. Thus, we have developed the HRF LED panel displays full-color (24-bits) images at 480 fps. Its maximum luminance is 5000 cd/m^2 [12] (Fig. 5.19).

The developed HFR LED panel was utilized for a kind of steganography, of which objective is to provide decoding fun of a hidden secret with a waving hand [13, 14], as shown in Fig. 5.20. The proposed method of displaying information is a kind of steganography technique with a novel way of decoding the hidden message. Such steganography can be enabled by LED's high speed of response time and high brightness that is enough to make afterimage.

In this experiment, a pair of coded images were alternatively shown on the LED panel. A text was embedded in black and white on a gray background image. A secret text was embedded into HFR images and represented at 240 fps. When the video image of the LED display was taken with camera at 60 fps, as shown in Fig. 5.21a, no text was perceived. A high-speed video (1200 fps) reveals flip-flop of the encoded

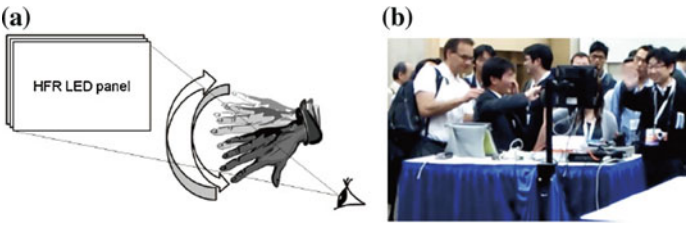

Fig. 5.20 **a** Schematic illustration of hand-waving steganography. By viewing the high-frame-rate LED panel, the viewer perceives the embedded information. **b** A photograph at an exhibition of hand-waving steganography

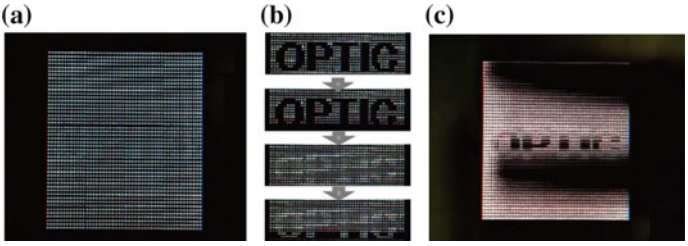

Fig. 5.21 **a** Schematic illustration of hand-waving steganography. By viewing the high-frame-rate LED panel, the viewer perceives the embedded information. **b** A photograph at an exhibition of hand-waving steganography

images, as shown in Fig. 5.21b. An example of viewed image through a waving hand is shown in Fig. 5.21c. Moving fingers blocked a portion of the flip-flop images and the hidden text was decoded. The position of the decoded portion was changed instantaneously because waving the hand was not synchronized to alternating the encoded images on the LED panel. However, after waving the hand in a certain time, the secret text was perceived. Viewers enjoy decoding with waving their hands continuously until they read the text.

5.5.2 Smart LED Tile

For the applications for visualization of sensory outputs, such as visualization of sound, acceleration, and rotation, it is necessary to reduce the total latency. We have originally designed and developed a new LED display module that integrate the sensors and display devices, named smart LED tile (SLT) [15].

Architecture of SLT is shown in Fig. 5.22. SLT integrates a microcontroller, sensors, a wireless module, and battery within the size of an LED panel, as shown in Fig. 5.23. We have designed all the circuit boards of which size is smaller than the LED panel (5 cm $\times$ 5 cm). It is possible to put the smart LED tiles face to face. The wireless network communication is based on ZigBee standard. SLT shows sensed

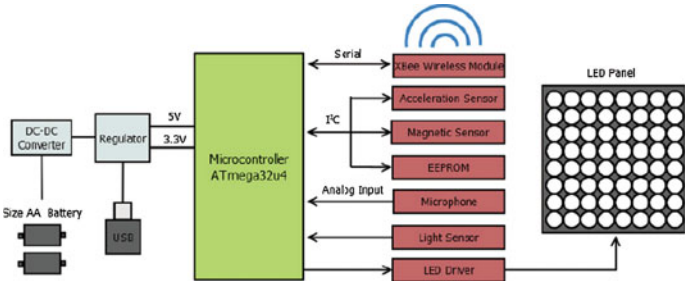

Fig. 5.22 Smart LED architecture that integrates LED panel, LED driver, sensors, a processor, a wireless communication module, and battery

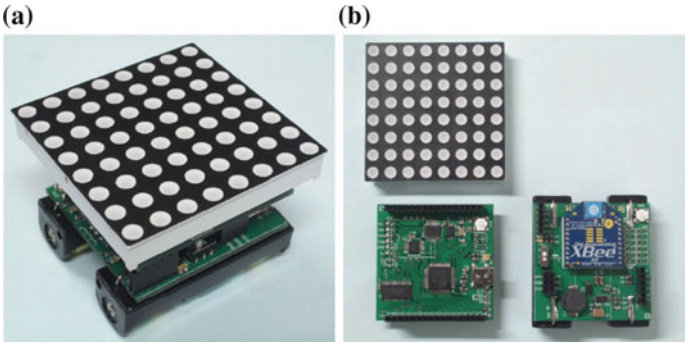

Fig. 5.23 **a** Smart LED tile and **b** its electronic boards and LED panel

information instantly and automously without any assistance with a host PC. SLT builds a wireless sensor network to share sensed information even when the smart tiles are moved. This feature is a solution for addressing and communication problems for a large LED screen with real-time sensor inputs.

Visualization of the acceleration vector of a smart LED tile has been conducted. Acceleration vector is mapped onto the RGB color space. The color is changed depending on the direction and the value of the acceleration. Experimental results are shown in Fig. 5.24. It is confirmed that changes of the acceleration are shown instantly on the LED panel when a smart LED tile is waved in front of a camera.

5.5.3 Aerial Imaging by Retro-Reflection (AIRR)

Aerial and transparent display is a prospective technique for digital signage to provide sensation to viewers. We propose aerial imaging by retro-reflection (AIRR) [16, 17]. Its principle is shown in Fig. 5.25a. Light rays that are reflected on the beam splitter impinge the retro-reflective material. After the retro-reflection, the lights travel reversely toward the light source. About a half of the retro-reflected

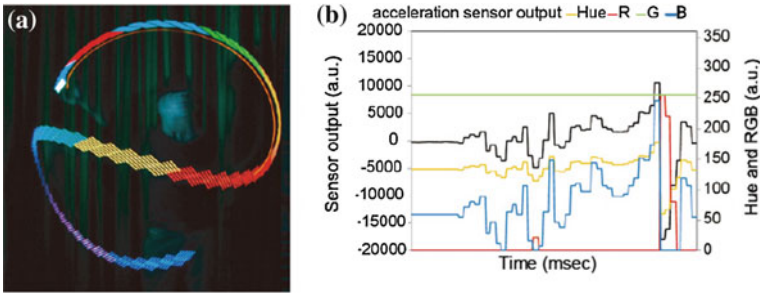

Fig. 5.24 **a** Visualization of acceleration by use of a smart LED tile. **b** Acceleration sensor outputs and converted color shown on the smart LED tile

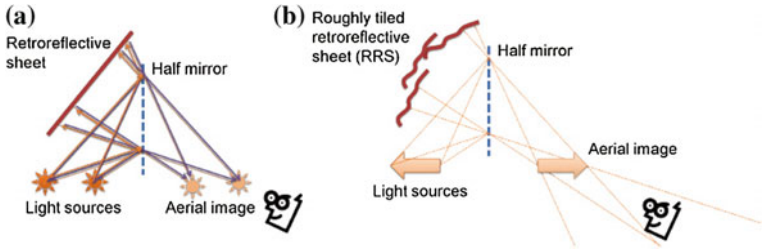

Fig. 5.25 **a** Principle of aerial imaging by retro-reflection (AIRR). **b** Aerial image is independent from the curvature of the retro-reflectors

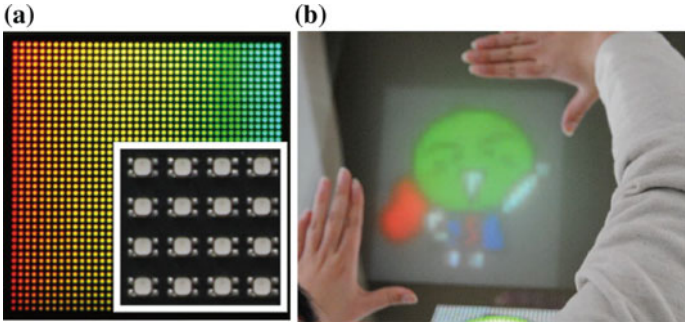

Fig. 5.26 **a** A photograph of 960-fps LED panel. The *inset* is a close-up of the LED panel. **b** Aerial image of the LED panel formed between hands

lights are transmitted through the beam splitter and form aerial image of the light sources. The basic setups were used for inverting pseudoscopic images in holographic display [18]. As shown in Fig. 5.25b, the image position does not depend on the position and curvature of the retro-reflective fabric.

Experimental results are shown in Fig. 5.26. HFR (960 fps) LED panel [19] was used for the light sources. Aerial image is floating between the hands. Note that the black regions between LED lamps (about 3 mm), shown in the inset of Fig. 5.26a, are filled in the aerial image.

5.6 Human Perceptual and Motor Functions for Coordinated Interaction with High-Speed Information Environment

It is essential to understand the nature of human sensori-motor system in order to achieve the coordinated interaction between individual humans and information environment. In this research project, we have dealt with three research topics related to this issue: (1) Effective error feedback timing for visuo-motor adaptation, (2) Prediction of human action based on spatio-temporal structure of human body movement (i.e., motor synergy), and (3) Nature of human visual perception for high-speed stimulus presentation. Below, we describe the first and third topics in detail.

5.6.1 Effective Error Feedback Timing for Visuo-Motor Adaptation

One of the human astonishing abilities is sensori-motor learning/adaptation: Humans can flexibly update internal memory so as to achieve a given task in a changing environment. In shooting a ball to a visual target, for example, people can readily modify the throwing action according to the ball weight. When the visual environment is distorted by a wedge prism or virtual reality devices, moreover, the shooting error gradually decreases and people correctly shoot the target after a few dozen trials. Such flexible ability plays a significant role when humans act in new information environments.

How does our brain acquire the information required for regaining task performance? For prism adaptation in a shooting or reaching task, visual information of the endpoint is essential because the adaptation hardly proceeds if the endpoint cannot be seen. The endpoint error calculated from visual information drives the adaptation. Here, in addition to spatial (or position) information, the timing of information feedback is also significant. Specifically, Kitazawa et al. [20] demonstrated that prism adaptation was slowed when the visual feedback was delayed: If visual presentation of the endpoint was delayed for more than 50 ms, the magnitude of adaptation diminished significantly. This effect has been replicated by recent studies [21, 22]. These findings suggest that human brain may accept error signals most effectively when they are synchronized with the end of reaching movements. In other words, temporal association between the motor action and its sensory consequence is an essential factor in visuo-motor learning [22].

In a shooting task, a ball reaches a target some time after it leaves the shooter's hand, meaning that the timing of the task-end (i.e., ball impact) is dissociated from that of motor execution (i.e., body movement). Here, an interesting question arises whether error feedback should be linked to *task-end* or *movement-end* (Fig. 5.27a, b). Considering that the motor adaptation is the mechanism for maintaining task per-

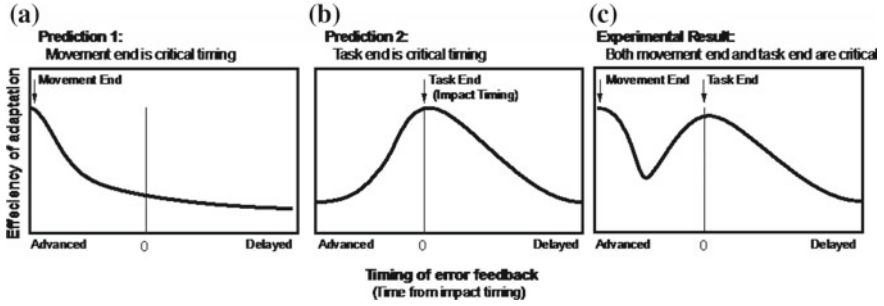

Fig. 5.27 Effect of error feedback timing on visuo-motor adaptation

formance, the timing of task-end, rather than movement-end, may be critical for accepting feedback information.

To answer the above question, we conducted a behavioral experiment using a virtual shooting task: Subjects were asked to control their wrist movements to shoot a target on a screen by moving a cursor as accurately as possible. A visual shift was implemented by displacing the ball trajectory on the screen. The time from the throwing action to impact was fixed to 600 ms, independent of the wrist movement speed. The timing of visual feedback of the impact location was manipulated as an experimental condition. The amount of visual feedback delay/advance was chosen from nine values: −500, −300, −200, −100, 0, 100, 200, 300, and 500 ms, where negative values mean that the task feedback was brought in advance of the ball impact and positive values mean the delayed feedback. We also prepared a condition that the impact timing was indicated by an additional timing cue so as to examine the effect of certainty of the impact timing. The magnitude of adaptation was measured by the amount of aftereffect estimated from the adaptation curve.

First of all, the visual shift introduced in our virtual shooting environment resulted in a learning curve similar to those reported for prism adaptation experiments. This confirmed that our experimental environment was meaningful.

The result shows that the amount of aftereffect varied depending on the timing of feedback. The aftereffect was large under the −500 ms condition (that is, when feedback was given just after movement-end) and decreased under the −300 ms condition. It then increased again and peaked broadly across the 0–500 ms feedback delay range (i.e., around task-end). When the feedback was delayed 1000 ms, the aftereffect decreased again. The effect of feedback delay was statistically significant. A similar pattern of results was obtained when the impact timing was indicated by an additional timing cue.

These results demonstrate that the efficiency of visuo-motor adaptation varied depending on the timing of error feedback, and increased around movement-end and task-end. This tendency was consistently observed irrespective of the time between movement-end and task-end, and regardless of whether a timing cue signaled task-end. Therefore, the present result indicates that the timing of error feedback

significantly affected the efficiency of visuo-motor adaptation, and that efficiency was enhanced both around movement-end and task-end (Fig. 5.27c).

Finally, we would like to discuss the mechanism underlying the acceptance of error information in visuo-motor adaptation. At least three ways are possible in which the brain determines the timing of error acceptance: (1) locked to the time of motor command generation; (2) specified by the sensory information accompanying movement-end or task-end; or (3) predicted within the brain. As we discussed in our journal paper in detail [23], however, we consider that no single mechanism determines the timing of error acceptance. Therefore, we speculate that multiple error acceptance mechanisms must be involved in visuo-motor adaptation.

5.6.2 Human Perception for High-Speed Visual Presentation

We investigated the nature of human visual perception for flicker stimuli using a high-speed video projector. The point is that we used flicker stimulus with a wide range of frequencies which could not be presented with conventional video displays (including CRT and LC displays). In another study, we examined the perception of a high-speed moving object. A high-speed projector improves the temporal resolution of stimulus presentation, which brings temporally dense visual information: Moving stimuli can be presented more continuously and smoothly compared with the conventional display devices.

It is broadly accepted that people could not detect the temporal change in luminance of a flickering stimulus if its frequency is higher than 50–60 Hz. This threshold frequency is called “critical fusion frequency (CFF)”. CFF has a quite important role in visual device. For example, the refresh rate of visual displays has been determined to be higher than CFF. People cannot distinguish two above-CFF flickering stimuli with equal subjective luminance if they are presented simultaneously at different locations. Recently, however, it has been reported that when the stimuli are presented sequentially at the same position, a transient ”twinkle” can be perceived around the moment of their changeover [24]. We name this phenomenon ”transient twinkle perception (TTP),” and examined its nature by a psychophysical experiment (Fig. 5.28a, b).

Significance of this phenomenon is that it suggests that human visual system can deal with the above-CFF visual stimuli. Recent progress in device technology has brought us new display devices with high refresh rates. Investigation of nature of our visual system for high-speed visual presentation is important for examining the merits and demerits of such novel devices.

On the other hand, this phenomenon is suggestive for understanding the computational process of our visual system. What mechanism causes TTP? A simple hypothesis is that perceptual luminance is determined by the temporal moving average of the physical luminance of the stimuli. In the present study, we examined whether this hypothesis can explain the TTP phenomenon using a computational model (boxes in Fig. 5.28a, b).

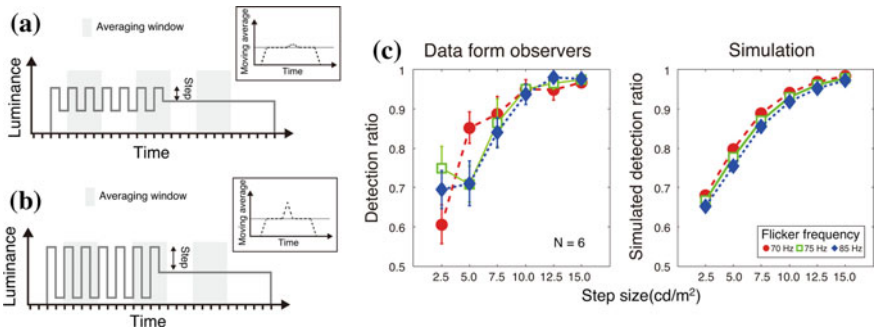

Fig. 5.28 Transient twinkle perception (TTP) [25]

In the psychophysical experiments, we adopted a high-speed video pro-jector (DLP) for presenting high-speed visual stimuli. In the experiment, ring-shaped stimuli having a sinusoidal luminance profile were presented on a uniform background (50 cd/m^2). Subjects were asked to discriminate TTP condition (sequential presentation of stationary and flickering stimuli) from no TTP condition (stationary stimulus only) by temporal 2-AFC. Flicker frequency was set to 70, 75, or 85 Hz in one experiment (using CRT), and to 100, 150, 200, 250, or 300 Hz in the other experiment (using DLP).

Subjects could discriminate TTP from no TTP conditions while the flicker frequency was no more than 200 Hz. In addition, correct rates were decreased as the amplitude of the flicker stimulus was smaller. This means that our visual system can detect the transient luminance change above CFF, but such detection requires a sufficient luminance difference between flickering and stationary stimuli [25].

We built a computational model for explaining these experimental findings. Our model consists of two processing stages, moving average and nonlinear transformation. In order to normalize the output of moving average, we calculated the relative deviation of the moving average from the long-term mean (DMR). Next, we transformed the DMR into the probabilistic value by applying a non-linear function. The resultant probability corresponded to the ratio that the subjects perceived the transient twinkle. Parameter values of the model were optimized to fit the result of the psychophysical experiments.

The behavior of the computational model is quite similar to that of human subjects (Fig. 5.28c). In addition, this model successfully replicated the result of another psychophysical experiment where the temporal deviation of moving average was induced by temporal perturbation to the stimuli. These results suggest that TTP may be basically brought by the short-term luminance averaging mechanism in the visual system, and that the larger luminance difference results in the larger temporal deviation from the long-term luminance mean which causes TTP. In other words, the perception of transient twinkles is determined by whether or not the amount of temporally averaged luminance around the transition exceeds a certain threshold.

As for the perception of a moving object, we have examined several topics; below, we briefly report the result on the misperception of position of a moving object.

It has been known that position of a moving object is often misperceived. Especially, the distance between the endpoints (or the length of motion path) of a reciprocally moving object tends to be perceived shorter than the veridical distance. However, it has been reported that this "perceptual shrinkage" hardly occurs if the object moves in one-way and /or at high speed [26]. In the present study, we tested this using a high-speed video projector which can present moving stimuli with a refresh rate of 500 Hz, much higher than the previous experiments. We measured the perceptual bias at the onset and offset positions of an one-way (12.8°) moving object at a speed faster than 30 deg/s. Different from the previous reports [26], perceptual shrinkage was observed also in the high-speed one-way motion: The onset position was perceived shifted toward the direction of motion while the offset position was perceived shifted backward. The amount of the shift was larger at the onset position than those at the offset position.

Therefore, the present study demonstrates that the high-speed and one-way motion could distort the perceptual position of a moving object, using high-speed visual presentation. This result supports the model that the perceptual position is represented by the spatiotemporal positional averaging and trajectory detection mechanism, even in the high-speed condition, again implying that the simple averaging mechanism may be the fundamental signal processing of our visual system [27].

In summary, we examined the characteristics of visual system for high-speed visual information in the present study. We have obtained several other findings related to visual motion perception, including position perception and trajectory prediction of a moving object. These results were obtained by means of utilizing a high-speed video projector whose refresh rate was much higher than the conventional video devices. In this sense, the high-speed visual technology can be a novel tool for investigating the mechanism of human visual system. Together, it implies that novel visual devices have possibilities to extract human potential abilities and to realize more sophisticated coordination between humans and information environments.

5.7 Development of the Systems for Dynamic Information Space

5.7.1 High-Speed and Non-contact Interfaces for Human Support

With the progress of hardware and image processing technologies, camera-based user interfaces are becoming familiar. For example, a gesture interface in which users can remotely control devices is realized by recognizing human hand motions from images captured by a camera. Augmented reality (AR) technology is also attracting

attention, which fuses real space and virtual space by overlaying computer graphics (CG) on images captured by a camera.

However, user interface systems using a camera/cameras have a problem of input-output delay (latency). Even in conventional input interfaces such as a mouse, delay of image drawing and displaying sometimes affects their operability. When a camera is used, the delay due to the camera's low frame rate, transfer delay, and image processing delay are added, which often results in a system with low responsivity.

On the other hand, we are developing new user interface systems which realizes high responsivity by using a high frame rate camera instead of a standard camera. In this section, an example of such systems is shown.

5.7.1.1 AR Typing Interface for Mobile Devices

Mobile devices equipped with a touch panel, such as smartphones, have become widely used. One of the advantages of a smartphone is that it allows the user to do works anywhere that previously required a PC. However, the screen of a smartphone is small and there are users feeling that the device lack usability because of the small operating area on the surface. This problem is particularly annoying in character entry.

To solve this problem, we have proposed an interface that allows a user to type on a virtual keyboard with his/her multiple fingers in the space behind a mobile device. This interface overlays a virtual keyboard and the user's hand on real images captured by a camera, and recognizes user's hand motions using optical flow information. We named this interface *AR typing interface* as it uses AR technology, which overlays CG on real images from a camera.

We constructed a PC-based experimantal system instead of using a real mobile device to evaluate the usability of the proposed interface. The system consists of a 4.3-in. display, a small high-frame-rate camera, and a PC. The image size and the frame rate of the camera is 320 × 240 pixels and 112 fps, respectively. Figure 5.29 shows an appearance of the system.

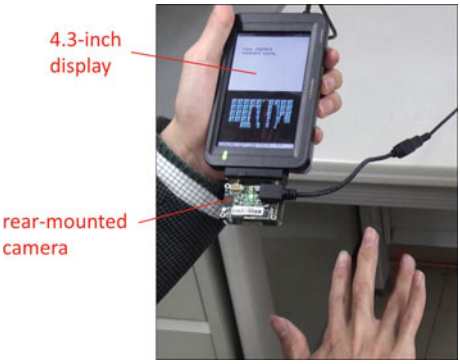

Fig. 5.29 The appearance of the experimental system [28]. A small high-frame-rate camera (112 fps @ 320 × 240 pixels) is attached to the back of a small (4.3 in.) display. A virtual keyboard is overlaid on real images captured by the camera. A user can operate the keyboard with his/her hand in the space behind the device

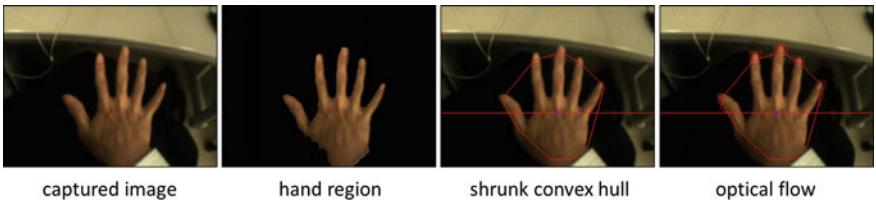

Fig. 5.30 The process to calculate optical flow [28]. A hand region is extracted by using skin color information. Optical flow is calculated in the hand region outside the shrunk convex hull of the hand region and above the centroid of the hand region

The system uses optical flow information to recognize typing action. To reduce computational cost, optical flow is calculated only in the regions around fingertips. The process is shown in Fig. 5.30. First, a hand region is extracted from a camera image by using skin color information. Then, the convex hull of the hand region is extracted and shrunk. Optical flow is calculated in the hand region outside the shrunk convex hull and above the centroid of the hand region.

After calculating optical flow, moving regions are detected by thresholding the flow magnitudes. The regions are separated by labelling and the average of the flow magnitudes in each region is calculated. The region with the maximum average magnitude is regarded as a key pressing region and its centroid is used as the pressed position.

When the whole hand is moving, almost all fingertip regions are regarded as moving regions. In such a case, the system does not perform key pressing action recognition.

Using this simple algorithm, fast typing recognition with a processing time of 8.33 ms (about 120 fps) is realized.

By using the recognition method above, we developed a keyboard typing application named *AR-keyboard*. A virtual keyboard is overlaid on captured images, and the user's hand image is also overlaid on the keyboard. By doing so, it is possible for users to perform key typing with the sense that there were a real keyboard under the user's hand. When a key pressing action is detected, the key on the keyboard at the pressed position is typed. A user can operate AR-keyboard not only in the air but at any space behind the display such as on the desk and on his/her knees.

Figure 5.31 shows the screenshot of AR-keyboard and sequential images of key typing. Multiple fingers are used for typing and multi-finger typing is realized. Also, high responsivity is realized by performing fast image processing with a high frame rate of 120 fps, which enables users to perform comfortable key typing with small delay.

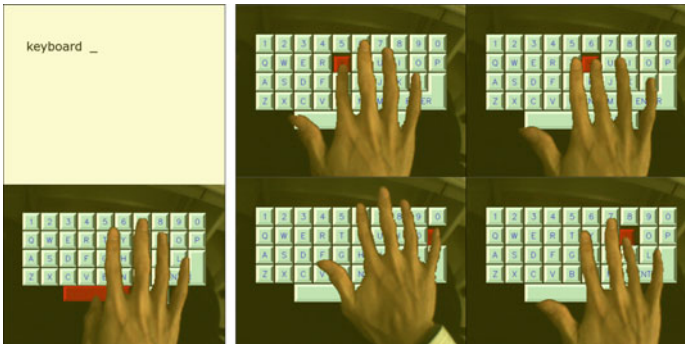

Fig. 5.31 Screenshot of AR-keyboard and multi-finger typing images [28]. Multiple fingers are used for typing

5.7.2 High-Speed Gaze Controller for High-Speed Computer-Human Interaction

5.7.2.1 Introduction

Gesture recognition based on computer vision is one promising approach for Computer-Human Interaction (CHI), since computer vision can recognize the operator's gesture without any contact nor bindings. In particular, if high-speed vision system, that can acquire images typically at 1000 fps, is adopted as the computer vision, very quick and smooth interaction is able to be achieved.

Most computer vision is usually used with a fixed field of view (FOV). Thus the area of interaction is limited within the FOV of the vision system. And there also exists a trade-off between the interaction area and the image resolution of the target, such as a human hand or body. If the interaction area is set to be much larger than the target size, the ratio of the target size to the FOV becomes smaller. Thus the vision system has to recognize the target from the coarser image. This trade-off can be solved by a pan/tilt camera that can control their gaze by two-axis rotational mechanical platform, commonly used for the monitoring and security purposes. However, if the high-speed vision is mounted on the conventional pan/tilt platform, the slow steering speed becomes system bottleneck and limits the advantage of the high-speed vision.

A new type high-speed gaze controller was developed originally to solve this problem. The goal was set to achieve a high-speed pan/tilt camera with the ability to change its gaze direction extremely quickly comparable to the frame rate of the high-speed vision. Conventional mechanical approach is difficult to achieve this goal due to the large inertia of the vision system usually composed of a camera lens and a housing of an imager. Thus, a special optical component that can steer the direction of the vision system was developed, so that the mounted camera was able to change its gaze direction without any physical movement. The developed component was also able to steer the projection direction of a conventional projector. And this function

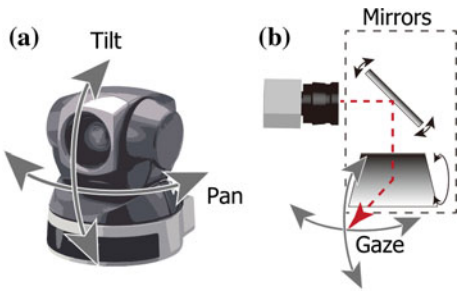

Fig. 5.32 Illustration of a general pan/tilt camera (**a**), and the high-speed gaze controller (**b**)

can realize a new projection mapping method on a dynamic object, including human body and hand.

5.7.2.2 Saccade Mirror and 1 ms Auto Pan/Tilt (APT) Technology

To realize a high-speed gaze controller, it's important to eliminate moving mechanical parts with large inertia. We focused on two-axis rotational mirrors for gaze control as shown in Fig. 5.32b, because:

- The inertia of a mirror can be considerably reduced by adopting small size mirrors.
- Flat mirror is known as one of an ideal optical components because they have no chromatic aberration and can easily achieve high optical performance with precise flat shape.

However, direct coupling of two-axis rotational mirrors and a camera requires a significant area of the mirror and reduces quick response ability.

To solve this problem, a pupil shift lens was inserted between the camera and rotational mirrors. The pupil is a position that limits the incident rays to the camera, which corresponds to the pinhole of the pinhole camera model. The pupil shift lens can shift the pupil to another place in open space in front of the camera lens. Since the ray bundle at pupil has the minimum cross-sectional area, even a small mirror can reflect all the necessary incident ray. For the details, refer the reference [29]. This type new high-speed gaze controller was named "Saccade Mirror" [29].

The saccade mirror can be applied to a noble video shooting system for moving objects by coupled with the high-speed vision system. It can capture the certain moving target as if it stopped and was fixed at the center of the field of view, and this technology was named "1 ms Auto Pan-Tilt (APT)" [30]. Figure 5.33 shows the photograph of a prototype and system connection of the 1 ms APT (a), and the captured image sequence with full high-definition size of a basket ball passed between two persons. The thrown ball is almost always tracked at the center of the FOV. Due to the high-speed response of the whole system, the 1 ms APT system had

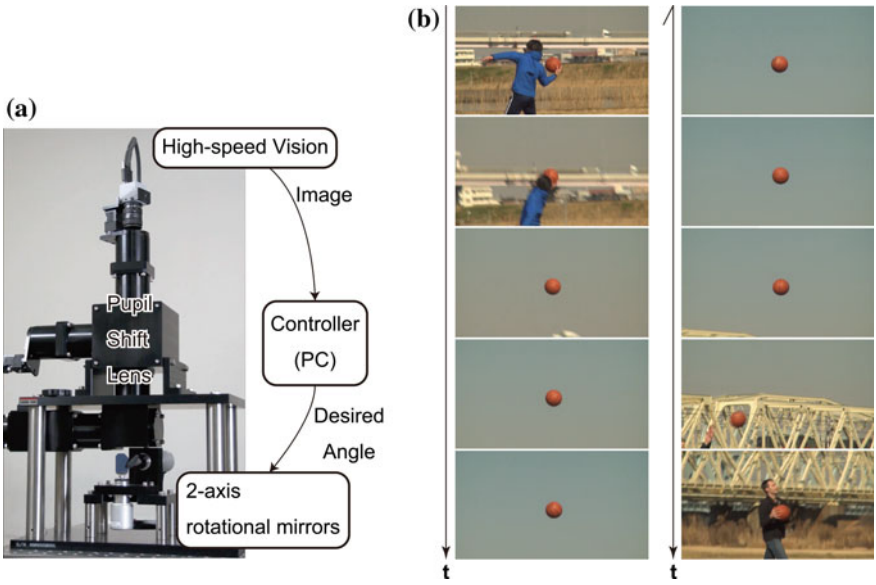

Fig. 5.33 Photograph of the prototype and schematic figure of the system connection (**a**), and captured image sequence while tracking a thrown basket ball (**b**)

also demonstrated the stable tracking of a high-speed table tennis ball in rally, and a quickly rolling yoyo [31, 32].

5.7.2.3 Lumipen: Active Projection Mapping on Dynamic Objects

The saccade mirror was explained as a gaze controller in above. However, if a projector is mounted on the saccade mirror instead of a camera, the saccade mirror can also control the projecting direction. Due to the principle of reversibility of light rays, the structure of the projector is very similar to the structure of the camera. Only the direction of the light ray is inverse. And this is the reason that the saccade mirror can work with projectors.

If the projection direction is always kept at the center of a moving object based on the 1 ms APT technology, the projected image would appear on the object as if it is printed on its surface. Because the conventional projectors have considerable timing delay ($\sim$100 ms) from the input of an image signal and the projection of the image, this delay makes it very difficult for the conventional projectors to fit the projection image position onto the object position precisely, and there exists the position shift between the object and projected image when the target is moving. The saccade mirror has much shorter delay of $\sim$3.5 ms and make it possible to project the given image precisely on the surface of the moving object.

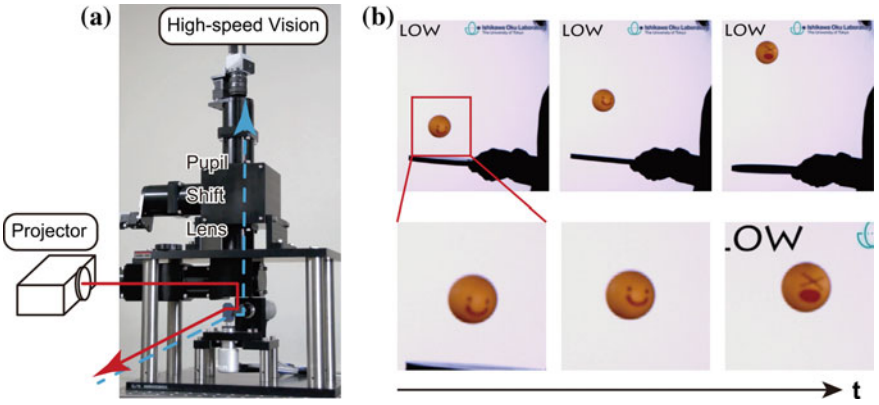

Fig. 5.34 Schematic figure of the system connection (**a**), and the projected image sequence on a lifting table tennis ball (**b**). A facial expression image was stably superimposed on the surface of quickly moving ball

This active projection technology was named "Lumipen" [33, 34]. Figure 5.34a shows a configuration of the Lumipen system. The projector is mounted on the saccade mirror sharing the optical axis with the high-speed vision so that the both instruments share the field of view/projection. Stable tracking control enables the projector always project an image on the surface of the target object. Figure 5.34b shows the projected image sequence on the lifting table tennis ball. The facial expression was always projected on the bouncing ball.

5.7.3 Integrated Systems

5.7.3.1 Concept and Related Systems

We have designed various types of the information space integrating high-speed sensing technology, high-speed display technology and human model.

As one of the promising conceptual architecture, "Invoked Computing" is proposed [35]. Direct interaction with everyday objects augmented with artificial affordances is clearly a very efficient approach leveraging natural human interaction capabilities. Hence the idea of conceiving ubiquitous computing as an invisible world can be "condensed" on real objects. Ubiquitous computing field actually is described as an "enchanted village" in which people discover hidden affordances in everyday objects. In "Invoked Computing", we explore the reverse scenario: a ubiquitous intelligence capable of discovering affordances suggested or represented symbolically by human beings (as actions and scenarios involving objects and drawings). We propose the following example: taking a banana and bringing it closer to the ear. The

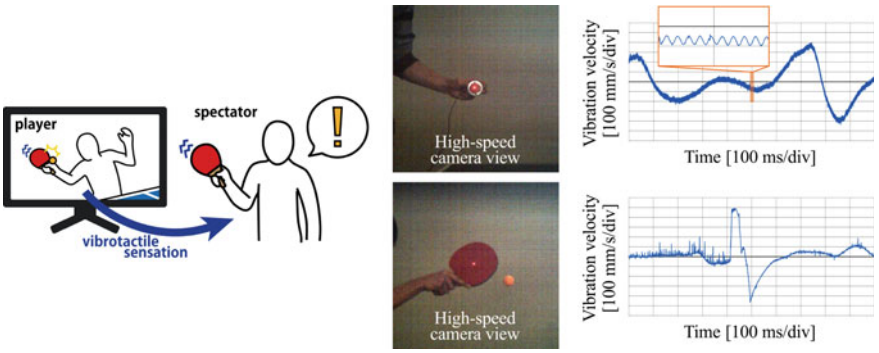

Fig. 5.35 VibroTracker: a vibrotactile sensor tracking objects [36]

gesture is clear enough: directional microphones and parametric speakers hidden in the room would make the banana function as a real handset on the spot.

Also, in order to enhance the experience in the information space, we newly have developed "VibroTracker" by integrating the "Saccade Mirror" and vibration meter [36]. For example, it is exciting merely to watch sports events, but simulating the haptic sensations experienced by a player would make spectating even more enjoyable. This is not peculiar to sports events. In addition to video and audio, the ability to relive the sensations experienced by others would also offer great entertainment value at temporal and spatial distances. The existing systems have some problems in measuring vibrations. A contact-type vibrometer deforms the original vibrations and is a burden to wear or carry. Even with a non-contact sensor like a microphone, it is difficult to measure slight vibrations of a fast-moving target against the surrounding noise. Our VibroTracker system solved these issues by using a laser Doppler vibrometer (LDV) and a high-speed optical gaze controller (Saccade Mirror), enabling users to relive the vibrotactile sensations experienced by others. Figure 5.35 shows the concept and the results of the developed system.

Moreover, we have developed a new type of wearable user interface involving high-speed vision technologies. The current trend towards smaller and smaller mobile devices may cause considerable difficulties in using them. Based on this background, we propose an interface called "Anywhere Surface Touch", which allows any flat or curved surface in a real environment to be used as an input area [37]. Figure 5.36 shows the developed system. The interface uses only a single small camera and a contact microphone to recognize several kinds of interaction between the fingers of the user and the surface. The system recognizes which fingers are interacting and in which direction the fingers are moving. Additionally, the fusion of vision and sound allows the system to distinguish the contact conditions between the fingers and the surface. Evaluation experiments showed that users became accustomed to our system quickly, soon being able to perform input operations on various surfaces.

Fig. 5.36 Anywhere surface touch: utilizing any surface as an input area with a wearable device [37]

We also have developed two more systems which integrates the subsystems described in the previous sections.

5.7.3.2 Visual and Tactile Cues for High-Speed Interaction

This is a new system which integrates two subsystems. Figure 5.37 shows the developed system. One subsystem can extract a 3-dimensional position of a moving object every 2ms, and project pictures (e.g. a screen of a video game or a computer, a movie etc.) on the moving object at the same time, using two "1ms Auto Pan-Tilt" systems described in Sect. 5.7.2 which can track an object in 3-dimensional space without delay by high-speed vision and two rotational mirrors. Another subsystem can display tactile sensation on an object depending on its 3-dimensional position, especially a particular position on a palm of a hand, using "Airborne Ultrasound Tactile Display (AUTD)" described in Sect. 5.4. This time, we realized a demonstration that papers around us and our hands are transformed into a screen of a computer or a smartphone, and we can feel even tactile sensation. In a sense, this system can be regarded as a moving object version and a tactile sensation version of projection mapping technology.

This system recognizes our hands and objects existing in the environment at a high-speed beyond human's ability of recognition by high-speed image processing technology, and it is possible to use the system to display and input information without uncomfortable feelings such as a delay. Thus, this system shows that we can use an object as a tool for human interfaces even if the object is moving. While we aim to embed intelligent function into objects such as conventional computers and

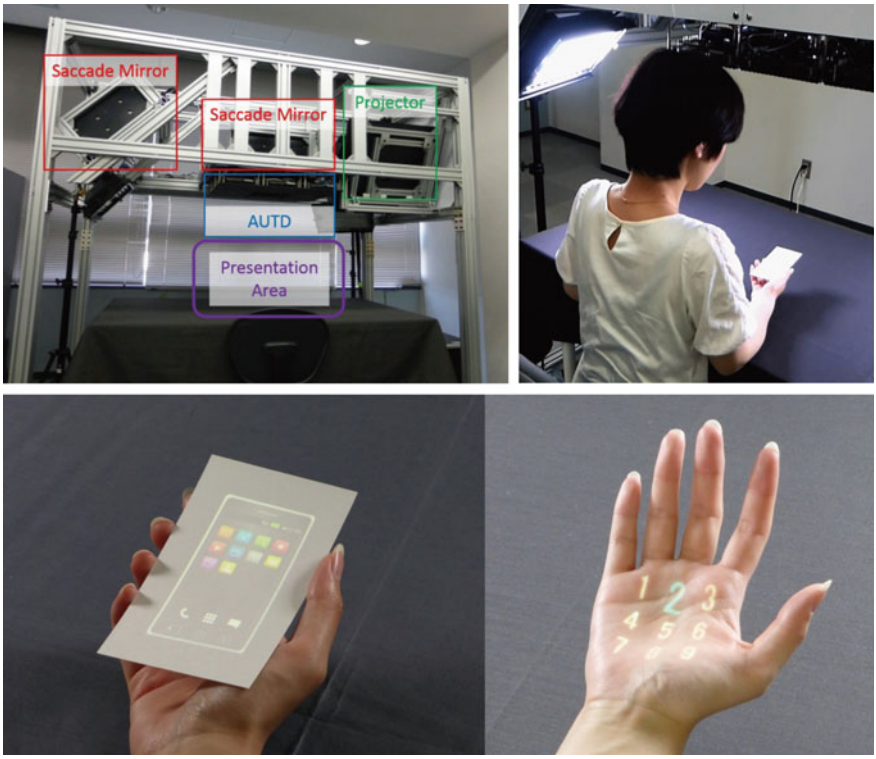

Fig. 5.37 Visual and tactile cues for high-speed interaction

smartphones, this system embeds information into existing environments and objects; it points the way toward future dramatic changes in our information environment.

5.7.3.3 AIRR Tablet: Floating Display with High-Speed Gesture UI

This system integrates high-speed and high-brightness floating display and high-speed 3D gesture recognition. For the aerial image, we use a display technology called AIRR technology described in Sect. 5.5.3 and the 3D high-speed hand tracking and gesture recognition [38] made it possible to manipulate the aerial image in high speed.

We believe that this will be the next generation of user-friendly 3D display technology. Previous methods to generate aerial image are based on lenses and mirror arrays. By using AIRR, much wider viewing angle is achieved. Furthermore, by using our newly developed LED display, bright image can be formed even under strong room lighting. In addition, the image can be viewed by multiple persons simultaneously (Fig. 5.38).

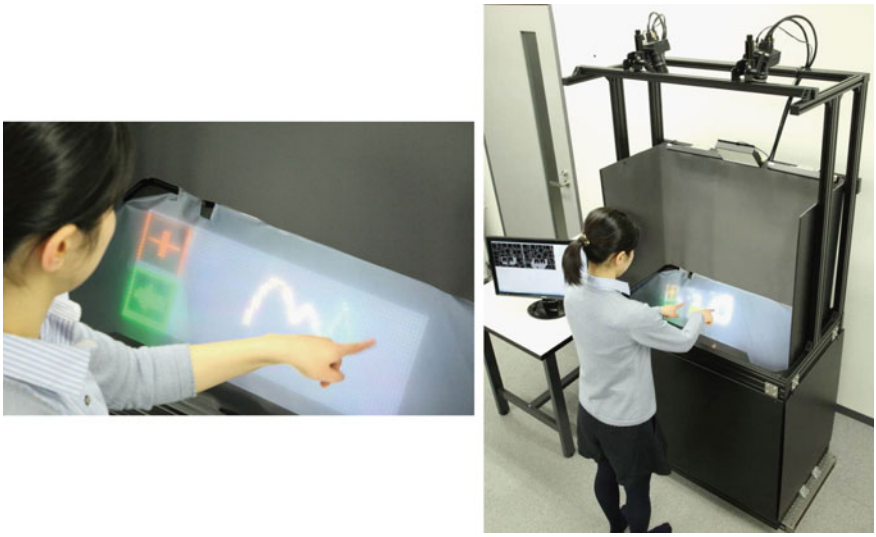

Fig. 5.38 AIRR Tablet: floating display with high-speed gesture UI

The high-speed 3D gesture recognition utilize super high-speed stereo cameras, which makes it possible to recognize gesture and track 3D position (500 fps) with extremely small latency. Not only the user can expand and rotate the floating screen, even if we perform extremely fast action such as punching it can still be detected. It can be said that high speed operation on floating image will become the next generation of information environment.

The system we integrated is called "AIRR Tablet" which recognizes hands or any other objects in high speed beyond human perception. We achieve input and output without any delay, and we show that we can turn the empty space into a large tablet. Unlike conventional computers and smartphones, we can perform operations without any physical collision. It enables high-speed 3D input and output.

References

1. I. Ishii, T. Tatebe, Q. Gu, Y. Moriue, T. Takaki, K. Tajima, 2000 fps real-time vision system with high-frame-rate video recording. *Proceedings IEEE International Conference on Robotics and Automation*, pp. 1536–1541 (2010)
2. Y. Liu, H. Gao, Q. Gu, T. Aoyama, T. Takaki, I. Ishii, High-frame-rate structured light 3-D vision for fast moving objects. J. Robot. Mechat. **26**(3), 311–320 (2014)
3. S. Inokuchi, K. Sato, F. Matsuda, Range imaging system for 3-D object recognition. *Proceedings Conference on Pattern Recognition*, pp. 806–808 (1984)
4. J. Chen, T. Yamamoto, T. Aoyama, T. Takaki, I. Ishii, Simultaneous projection mapping using high-frame-rate depth vision. *Proceedings of IEEE International Conference on Robotics and Automation*, pp. 4506–4511 (2014)

5. M. Shimojo, T. Araki, S.Teshigawara, A. Ming, M. Ishikawa, A net-structure tactile sensor covering free-form surface and ensuring high-speed response. *Proceedings of IEEE International Conference on Intelligent Robots and Systems*, pp. 670–675 (2007)
6. H. Arita, Y. Suzuki, H. Ogawa, K. Tobita, M. Shimojo, Hemispherical net-structure proximity sensor detecting azimuth and elevation for guide dog robot. *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 653–658 (2013)
7. K. Hasegawa, H. Shinoda, Aerial display of vibrotactile sensation with high spatial-temporal resolution using Large-Aperture airborne ultrasound phased array. *Proceedings of IEEE World Haptics Conference 2013*, pp. 31–36 (Daejeon, Korea, 2013)
8. T. Iwamoto, M. Tatezono, H. Shinoda, Non-contact method for producing tactile sensation using airborne ultrasounders. *Haptics: Perception, Devices and Scenarios: 6th International Conference, Eurohaptics 2008 Proceedings* (*Lecture Notes in Computer Science*), pp. 504–513 (2008)
9. T. Hoshi, T. Masafumi, T. Iwamoto, H. Shinoda, Noncontact tactile display based on radiation pressure of airborne ultrasound. IEEE Trans. Haptics **3**(3), 155–165 (2010)
10. J. Awatani, Studies on acoustic radiation pressure. I (General considerations). J. Acoust. Soc. Am. **27**, 278–281 (1955)
11. P.J.J. Lamoreet, H. Muijser, C.J. Keemink, Envelope detection of amplitude-modulated high-frequency sinusoidal signals by skin mechanoreceptors. J. Acoust. Soc. Am. **79**, 1082–1085 (1986)
12. H. Yamamoto, M. Tsutsumi, K. Matsushita, R. Yamamoto, K. Kajimoto, S. Suyama, Development of high-frame-rate LED panel and its applications for stereoscopic 3D display. Proc. SPIE **7956**, 79560R (2011)
13. S. Farhan, S. Suyama, H. Yamamoto, Hand-waving decodable display by use of a high frame rate LED panel. *Proceedings of IDW '11*, vol. 3, pp. 1983–1986 (2011)
14. H. Yamamoto, K. Sato, S. Farhan, S. Suyama, Hand-waving steganography by use of a high-frame-rate LED Panel. *SID 2014 DIGEST*, pp. 915–917 (2014)
15. K. Sato, A. Tsuji, S. Suyama, H. Yamamoto, LED module integrated with microcontroller, sensors, and wireless communication. *Proceedings of the International Display Workshops***20**, 1504–1507 (2013)
16. H. Yamamoto, S. Suyama, Aerial 3D LED display by use of retroreflective sheeting. Proc. SPIE **8648**, 86480Q (2013)
17. H. Yamamoto, Y. Tomiyama, S. Suyama, Floating aerial LED signage based on aerial imaging by retro-reflection (AIRR). Optics Express **22**(22), 26919–26924 (2014)
18. C.B. Burckhardt, R.J. Collier, E.T. Doherty, Formation and inversion of pseudoscopid images. Appl. Opt. **7**, 627–631 (1968)
19. T. Tokimoto, K. Sato, S. Suyama, H. Yamamoto, High-frame-rate LED display with pulse-width modulation by use of nonlinear clock. *Proceedings of 2013 IEEE 2nd Global Conference on Consumer Electronics*, pp. 83–84 (2013)
20. S. Kitazawa, T. Kohno, T. Uka, Effects of delayed visual information on the rate and amount of prism adaptation in the human. J. Neurosci. **15**, 7644–7652 (1995)
21. H. Tanaka, K. Homma, H. Imamizu, Physical delay but not subjective delay determines learning rate in prism adaptation. Exp. Brain Res. **208**, 257–268 (2011)
22. T. Honda, M. Hirashima, D. Nozaki, Adaptation to visual feedback delay influences visuomotor learning. PLoS ONE **7**, e37900 (2012)
23. T. Ishikawa, Y. Sakaguchi, Both movement-end and task-end are critical for error feedback in visuomotor adaptation: a behavioral experiment. PLoS ONE **8**, e55801 (2014)
24. S. Cheadle, A. Parton, H. Muller, M. Usher, Subliminal gamma flicker draws attention even in the absence of transition-flash cues. J. Neurophys. **105**, 827–833 (2011)
25. Y. Nakajima, Y. Sakaguchi. Abrupt transition between an above-CFF flicker and a stationary stimulus induces twinkle perception: evidence for high-speed visual mechanism for detecting luminance change. J. Vis. **13**, 311 (2013)
26. M. Sinico, G. Parovel, C. Casco, S. Anstis, Perceived shrinkage of motion paths. J. Exp. Psychol. Hum. Percept Perform. **35**, 948–957 (2009)

27. Y. Nakajima, Y. Sakaguchi, Perceptual shrinkage of motion path observed in one-way high-speed motion. *Proceedings 24th Annual Conference JNNS*, pp. 88–89 (2014)
28. M. Higuchi, T. Komuro, Multi-finger AR typing interface for mobile devices using high-speed hand motion recognition. *Extended Abstracts on ACM SIGCHI Conferene on Human Factors in Computing Systems* (*CHI 2015*), pp. 1235–1240 (2015)
29. K. Okumura, H. Oku, M. Ishikawa, High-speed gaze controller for millisecond-order Pan/tilt Camera. Proc. IEEE ICRA **2011**, 6186–6191 (2011)
30. K. Okumura, K. Yokoyama, H. Oku, M. Ishikawa, 1 ms auto Pan-Tilt–video shooting technology for objects in motion based on Saccade Mirror with background subtraction. Adv. Robot. **29**, 457–468 (2015)
31. YouTube, https://youtu.be/9Q_lcFZOgVo
32. YouTube, https://youtu.be/Of2suN6ijao
33. K. Okumura, H. Oku, M. Ishikawa, Acitve projection AR using high-speed optical axis control and appearance estimation algorithm. *Proceedings of IEEE ICME* (2013). doi:10.1109/ICME.2013.6607637
34. T. Sueishi, H. Oku, M. Ishikawa, Robust high-speed tracking against illumination changes for dynamic projection mapping. *Proceedings of IEEE VR2015*, pp. 97–104 (2015)
35. A. Zerroug, A. Cassinelli, M. Ishikawa, Invoked computing: spatial audio and video AR invoked through miming. *Proceedings of Virtual Reality International Conference* (2011)
36. L. Miyashita, Y. Zou, M. Ishikawa, VibroTracker: a vibrotactile sensor tracking objects. *SIGGRAPH 2013*, Emerging Technologies (2013)
37. T. Niikura, Y. Watanabe, M. Ishikawa, Anywhere surface touch: utilizing any surface as an input area. *The 5th Augmented Human International Conference* (2014)
38. M.S. Alvissalim, M. Yasui, C. Watanabe, M. Ishikawa, Immersive virtual 3D environment based on 499 fps hand gesture interface. *International Conference on Advanced Computer Science and Information Systems*, pp. 198–203 (2014)