[Invited] Image sensing for human-computer interaction

Takashi Komuro[1]
[1] Graduate School of Science and Engineering, Saitama University
255 Shimo-okubo, Sakura-ku, Saitama, 338-8570 Japan
E-mail: komuro@mail.saitama-u.ac.jp

Abstract Cameras play an important role in human-computer interaction (HCI) systems. However, standard cameras are often diverted to such systems and they are not always suitable for HCI. In this presentation, we introduce examples of HCI systems that use special cameras such as high-speed cameras, depth cameras, and multiple cameras.

Keywords: user interfaces, virtual/augmented reality, latency, user-perspective projection

1. Introduction

With the progress of hardware and computer vision technology, HCI systems using cameras are becoming practical. However, standard cameras are often used in such systems, which are not necessarily suitable for HCI. In this presentation, we introduce examples of HCI systems that use special cameras such as high-speed cameras, depth cameras, and multiple cameras to realize higher performance and new features.

2. HCI using high-speed cameras

Reducing latency is important in HCI systems to improve their operability. In particular, HCI systems that use cameras tend to have slow response since it takes long time to perform image capturing, transfer, and processing.

On the other hand, we have developed some HCI systems using high-speed cameras to realize high response. Figure 1 shows a user interface that allows a user to type on a virtual keyboard with his/her multiple fingers in the space behind a mobile device [1]. 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 (augmented reality) technology, which overlays CG on real images.

Fig. 1. AR typing interface

The prototype system consists of a 4.3-inch 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. The system uses optical flow information to recognize typing action. To reduce computational cost, optical flow is calculated only in the regions around fingertips.

VR/AR systems that allow manipulation of virtual objects using users' hand have been developed, but there is a problem of the delay between hand movement and object movement. If the hand is displayed together with virtual objects on the display, delay of up to a certain extent is corrected in the brain. However, if the hand in real space and objects in virtual space are within the same field of view, the delay is perceived visually as positional gap. In such a case, requirement for time delay becomes more severe.

With the background above, we developed a 3D tabletop user interface system that uses a high-speed stereo camera to synchronize real hands and virtual objects, both temporally and spatially, with a high degree of accuracy [2].

Figure 2 shows the developed system. It consists of an upwardly-oriented multi-view autostereoscopic display and a high-speed stereo camera installed above the display. The autostereoscopic display is used for presenting 3D images to users without them having to wear special glasses. To reduce delay, we used cameras having a frame rate of 200 fps (Grasshopper from Point Grey Research Inc.).

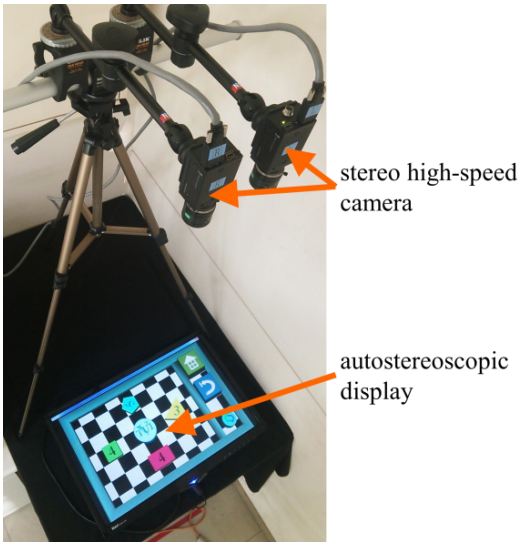

Fig. 2. 3D tabletop user interface

We measured the latency of the system using another high speed camera. While the latency when the camera was operated at 30 fps was 49.0 ms, the latency when the camera was operated at 200 fps decreased to 28.3 ms. We can also see from Fig. 3 that the positional deviation between the finger and the virtual object was reduced.

The results of a user study we conducted showed that there was a significant difference in the time required to complete an object moving operation between the 200 fps and 30 fps cases.

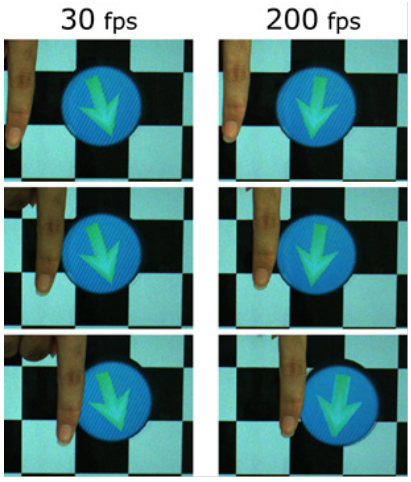

Fig. 3. Positional deviation between finger and virtual object

3. HCI using depth cameras

In AR systems using mobile devices, there is a problem that the images of the scene that are displayed on the screen have different positions and sizes from the real ones. We developed a system that allows users to see the scene behind the device as if the device were transparent [3]. By superimposing a virtual object in the scene and performing collision detection, this system also allows users to interact with virtual objects such as grasping and lifting a virtual object as shown in Fig. 4.

Fig. 4. See-through mobile AR system

Figure 5 shows the configuration and principle of the system. A camera for face tracking is attached to the front of a mobile display, and a depth camera for capturing images of a hand and the background is attached to the back of the display. The 3D scene obtained by the depth camera is displayed after changing the center of projection to the user's viewpoint position.

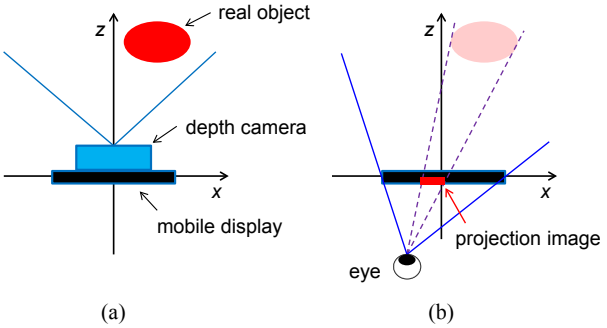

Fig. 5. Projection according to the user's viewpoint

Based on the see-through mobile AR system, we developed a system that allows the user to manipulate a virtual object using the user's own hand and that provides the user with perception of materials [4]. This system supports users to obtain the material appearance of products in on-line shopping and enhances users' willingness to purchase the products. Figure 6 shows a user moving a virtual object by his hand using the system. The user was able to finely control the virtual object and to observe subtle change of the gloss and burnish of the object.

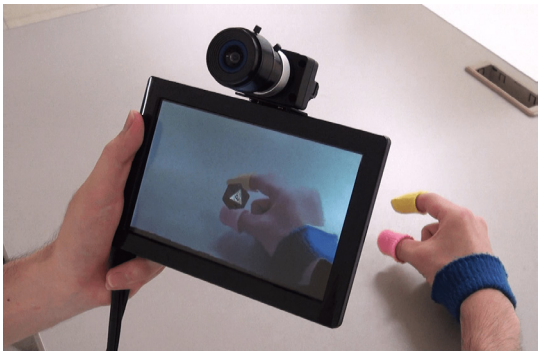

Fig. 6. Mobile AR system for providing material perception

In order to reproduce the material appearance of real objects, we developed a system that measures the shape and reflectance of real objects [5]. The system captures depth and color images of a rotating object that is placed on a turntable using an RGB-D camera. The shape of the object is reconstructed by integrating the depth images of the object captured from different viewpoints. The reflectance of the object is obtained by estimating the parameters of a reflectance model from the reconstructed shape and color images.

Figure 7 shows the measurement setup. We connected Microsoft Kinect to a PC and used it for capturing the images of a target object. We used a small LED light as a single light source and fixed the light to the camera.

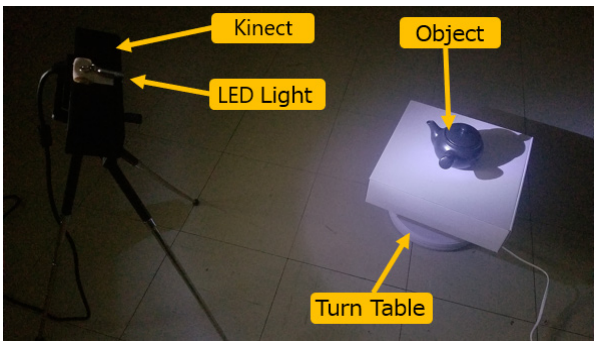

Fig. 7. Measurement setup

The system calculates the correspondence between each vertex of the reconstructed 3D model and color pixels in all the captured frames. By reprojecting the vertex to all the camera viewpoints, it is determined which pixel in each color image the vertex corresponds to. By picking the corresponding pixel values, intensity change at the vertex over all frames is obtained. The reflectance of the target object is estimated from the reconstructed shape, acquired change intensity, and the light source information. We used the Blinn-Phong model as the reflectance model. To obtain the parameters of this model, the least squares method is applied to the reflectance model and observation data.

Figure 8 shows an example of presenting the material appearance of real objects. The user is manipulating the virtual object whose model was generated from the shape and reflectance of a real object, using his own hand.

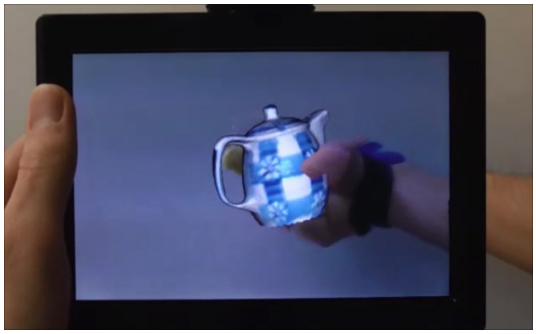

Fig. 8. A user is manipulating an object whose model is generated from a real object

4. HCI using multiple cameras

By using a depth camera, it is possible to reproduce images that are seen from different viewpoints. However, it is difficult to reproduce the scene that is unseen from the camera due to occlusion, or to reproduce specular reflection. By using multiple cameras and capturing images from different positions, more accurate free-viewpoint images can be generated.

Figure 9 shows a video conference system that generates images of each person from an appropriate viewpoint position and to present the images to the user in order to provide correct gaze directions [6]. In this system, a Kinect sensor is placed on the top of a screen, and three cameras that are placed behind the screen capture images through small holes drilled on the screen.

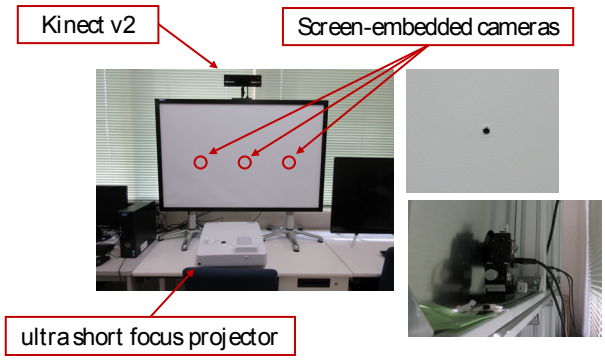

Fig. 9. Video conference system using screen-embedded cameras

The Kinect sensor is used to estimate users' viewpoint positions and viewing directions. Using this information, the system finds the correspondence of who is look at each person. Then, using the depth information from Kinect and images from the screen-embedded cameras, the system generates and presents images of persons that are rendered from appropriate positions. To generate images as natural as possible, the system switches the camera to capture the texture image that is mapped to a person's surface according to the viewpoint position.

A demonstration of the system in a two-to-two video conference is shown in Figure 10. Both the users sitting on the left side (A) and the right side (B) look at the same remote person displayed on the left side of the screen (target). Since the target looks at B, the target is rendered from the viewpoint of B.

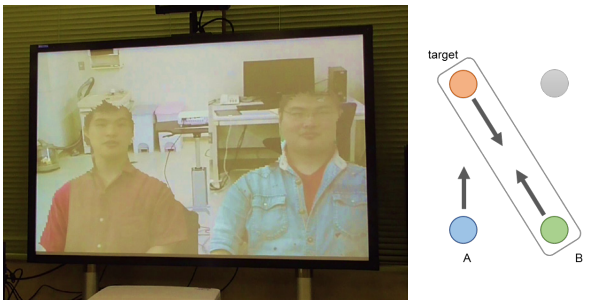

Fig. 10. Demonstration of a two-to-two video conference

As a result, the image of the target that was suitable for B was presented.

5. Conclusion

We introduced examples of HCI systems that use special cameras such as high-speed cameras, depth cameras, and multiple cameras. Requirements for cameras in HCI, such as low latency, depth sensing capability, wide field of view and high-resolution, are different from those for photography. The camera technology for photography is already being saturated, but new image sensing technology for HCI still has room for improvement. We are expecting new HCI systems that are realized by using such new image sensing technology.

References

[1] M. Higuchi, T. Komuro: "Multi-finger AR Typing Interface for Mobile Devices Using High-Speed Hand Motion Recognition", Ext. Abst. ACM SIGCHI Conference on Human Factors in Computing Systems, pp. 1235-1240 (2015)

[2] T. Kusano, T. Komuro: "3D Tabletop User Interface with High Synchronization Accuracy using a High-speed Stereo Camera", Proc. the 2015 ACM International Conference on Interactive Tabletops and Surfaces, pp. 39-42 (2015)

[3] Y. Unuma, T. Komuro: "Natural 3D Interaction using a See-through Mobile AR System", Proc. The 14th IEEE International Symposium on Mixed and Augmented Reality, pp. 84-87 (2015)

[4] R. Nomura, Y. Unuma, T. Komuro, S. Yamamoto, N. Tsumura: "Mobile Augmented Reality for Providing Perception of Materials", Proc. 16th International Conference on Mobile and Ubiquitous Multimedia, pp. 501-506 (2017)

[5] S. Tsunezaki, R. Nomura, T. Komuro, S. Yamamoto, N. Tsumura: "Reproducing Material Appearance of Real Objects using Mobile Augmented Reality", Adj. Proc. 2018 IEEE International Symposium on Mixed and Augmented Reality, pp. 196-197 (2018)

[6] K. Kobayashi, T. Komuro, B. Zhang, K. Kagawa, S. Kawahito: "A Gaze-preserving Group Video Conference System using Screen-embedded Cameras", Proc. the 23rd ACM Symposium on Virtual Reality Software and Technology, Article No. 82 (2017)