Vision-based 3D Input Interface Technologies

Takashi Komuro

Graduate School of Information Science and Technology, The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan

ABSTRACT

Vision-based user interfaces are technologies where human motion is detected by a camera or cameras and is used as input. In this paper we describe advantages and disadvantages, technological issues, and related research of vision-based user interfaces as well as examples of 3D input interface systems that we developed.

1. INTRODUCTION

User interfaces (UIs) of current computers are mostly with a keyboard and a mouse. They are suitable for fast operations which principally include text input, but they require some training to use properly and also require a place to put the devices. Therefore, touch panel UIs, which are simpler and easier to use, are prevalent in mobile and public information devices.

However, touch panels generally have low position accuracy, which makes precise operations difficult. In addition, the finger position is detected only after the screen is touched and it is difficult to predict incorrect operation in advance. Consequently, fast operation was impossible in text input and menu selection, which was the bottleneck that prevents information devices from becoming more sophisticated.

Vision-based UIs using a camera or cameras have a potential both to solve these problems and to realize simpler and easier-to-use interfaces than touch panel UIs. In this paper we describe advantages and disadvantages, technological issues, and related research of vision-based user interfaces as well as examples of 3D input interface systems that we developed.

2. VISION-BASED UIS

User interfaces where human motion is detected by a camera or cameras and is used as input are called vision-based UIs. Vision-based UIs have the following advantages: 1) unrestrained and noncontact, 2) separation of display space and operation space, 3) high spatial resolution, 4) capable of obtaining 3D information. On the other hand, they have some disadvantages such that there is no tactile feedback and that users often get tired during long time use. In addition, there still remain some technological issues on temporal resolution, response time, and stability against background and disturbance.

Typical vision-based UIs are those where a user operates a device with his/her gestures. Microsoft's Project Natal is a gesture control system which obtains human motion with a 3D camera and uses it for game control. Gesture controlled TVs, which Hitachi and Toshiba both developed, enable remote control of home appliances by human gestures. These systems mainly use the advantage of vision-based UIs that they are unrestrained and noncontact, but they do not remove the bottleneck of information devices.

Canesta Projection Keyboard [1] is a system which projects a keyboard image on a flat surface was developed to assist typing input for mobile devices. It uses another advantage of vision-based UIs that display space and operation space are separated and is a pioneer technology that removes the bottleneck of information devices. However, it requires a place to fix the device and a flat surface to project a keyboard on, which restricts the environment of usage.

We therefore focus on other advantages of vision-based UIs that they have high spatial resolution and that are capable of obtaining 3D information, and aim to remove the bottleneck that conventional information devices have by actively taking the advantages. We show development examples of such 3D input interface systems.

3. ZOOMING TOUCH PANEL

Information devices with a touch panel are often seen in our life such as ATMs in banks, ticket vending machines at stations, and ordering system in restaurants, but touch panels are difficult to operate precisely as well as it is unsanitary. We developed a touch panel with a new interface [2], which measures three dimensional position of the finger near the panel using two cameras, displays a cursor on the panel, and zooms the screen according to the finger depth. Figure 1 shows the appearance of the system.

We implemented a demonstration program of ticket vending machines at stations which enables a user to select the destination directly on the map. When a user moves his/her finger close to the panel, the screen is zoomed in around the finger position, which makes it easy to select items. In this way, this system enables more precise operation than existing touch panels and can increase amount of information that can be displayed.

Fig. 1 Zooming Touch Panel

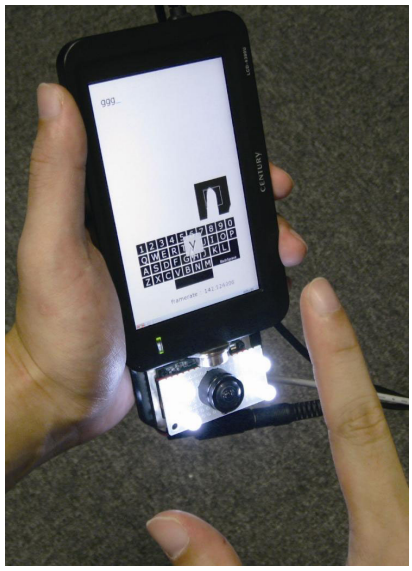
4. 3D INPUT INTERFACE FOR MOBILE DEVICES

Recently mobile devices have become small and are difficult to have input interface that has wide operation area on their surface. For example, conventional interface such as a keypad or a touch panel on a cell phone has limited operation area, which causes low operability. We therefore propose a vision-based 3D input interface for mobile devices where users can operate the device by the movement of a fingertip in the air.

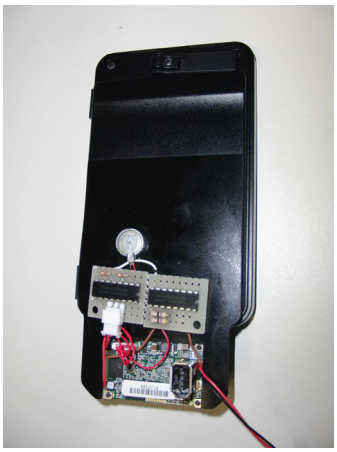
In order to realize such an interface system, there are some technological issues such as stable recognition of the position of a fast-moving fingertip and accurate detection of keystroke actions using a single camera. Our system stably estimates 3D positions of the fingertip by using a binarized image of the fingertip as a template image and by tracking the fingertip in input images about translation, rotation and scale. Since the scale of the fingertip is inversely proportional to the distance between finger and camera, the 3D position of the fingertip can be estimated.

In mobile devices the distance between the camera and the fingertip is close and the fingertip moves fast in the images. We therefore use a high-frame-rate camera to realize stable tracking. Small keystroke gestures in the air are detected by applying a bandpass filter to the scale change of the fingertip images.

To study the effectiveness of the proposed interface, we built a prototype system [3][4]. The appearance of the system is shown in Fig. 2. The system consists of a compact IEEE1394 high-frame-rate camera Firefly MV (Point Grey Research Inc.) with a lens having a focal length of 1.9 mm, four white LEDs, a PC and a small USB display. The frame rate of the camera is 144 fps with an image size of 752 x 180 pixels. Using the wide-angle lens, the angle of view of 90 degrees is acquired. Since the images obtained through a wide-angle lens is distorted, we applied distortion correction to the obtained images.



(a)



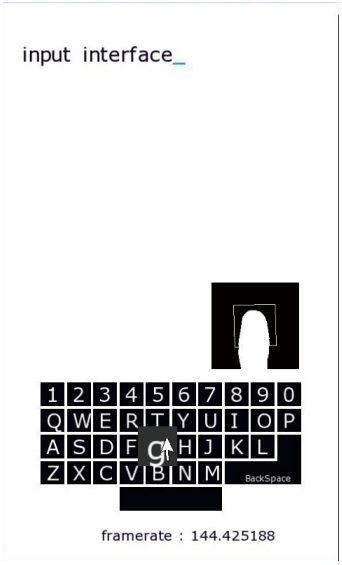
(b)

Fig. 2 3D Input Interface for Mobile Devices
(a) front (b) back

To realize input interface in which we can type letters fast, tactile feedback is important. With tactile feedback, users can quickly recognize that the input action of the users is detected. From this respect, we attached a vibration motor on the back side of the display which vibrates for a short time when a keystroke action is detected.

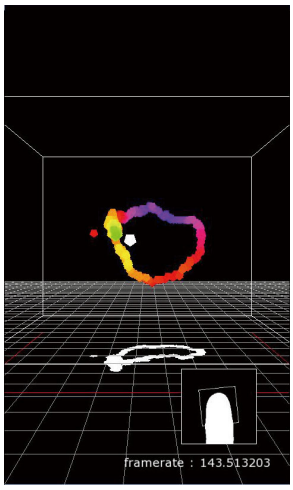
We implemented several applications using the system shown in Fig. 3. In-air Typing Keyboard is an application where users can type letters in the air. The pointer of software keyboard moves according to the position of the fingertip. If the pointer is located on the target key and a keystroke action is detected, the target character is input. In Zooming Picture Viewer, users can zoom and scroll the picture on the display with the 3D position of the fingertip. The user can grab the picture by a keystroke action, and then the picture zooms and

scrolls on the screen according to the 3D position of the fingertip. When a keystroke action is carried out again, the picture is released and fixed on the screen. In 3D Painter, users can draw lines in 3D space with fingertip using 3D position directly. The drawn figure can be displayed from different views using the finger movement.



(a)

(b)



(c)

Fig. 3 Application Examples:
(a) In-air Typing Keyboard, (b) Zooming Picture Viewer, (c) 3D Painter

As seen above, by obtaining 3D position with a single camera, the proposed interface solves the problem of small operation area in mobile devices as well as realizes new mobile applications which we have never seen before.

5. FAST HAND TRACKING FOR VISION-BASED UIS

Vision-based hand gesture recognition often uses a silhouette image of hand, which wouldn't work well in the presence of cluttered background. Even if skin color extraction is used, it would fail to distinguish hand region from face region.

We therefore propose a template-matching-based hand tracking method using hand/finger textures [5], which reduces background effect. Our algorithm assumes that the palm and each finger are flat, and estimates relative motion between hand/fingers and the camera. To achieve stable tracking, we introduce a structural model of a hand. The pose of the palm is represented by 3 translational and 3 rotational parameters with its origin at wrist. The pose of each finger is represented by 2 angle parameters.

The template images of a hand and each finger are captured at initialization stage, which cancels individual difference. The input images are preprocessed so as to reduce the effect of illumination variation. First of all, the hand region, which is predicted from the pose of hand/fingers in the previous frame, is cropped and scaled from an image to fit into 4096 pixels. Then, the image is denoised with a smoothing filter, and is converted to a grayscale image with enhanced texture by calculating the gradient in each of the RGB components. Finally the brightness is normalized with histogram equalization. A high-speed camera is used to narrow down the search range. We also refine the search range by applying pose recognition of the palm and fingers in a stepwise fashion.

We developed a hardware system with FPGA coprocessor. The coprocessor, which is implemented in a Xilinx FPGA Virtex4 XC4VLX25, can execute template matching with perspective transformation 260,000 times per second. The coprocessor consists of PEs (Processing Elements), control circuits, and interface circuits. The PEs perform main computation and each PE has an arithmetic pipeline and local memory. The pipeline has 12 stages and processes perspective transformation and matching operation by hard-wired logic. The template image and the current input image are stored in the local memory. The computation is parallelized using a 6-way SIMD organization. We used a commercially-available USB camera Artray ARTCAM-036MI (up to 200 fps @ 320 x 240 pixels) with a wide-angle lens (f=4.2 mm).

Using the developed system, we conducted a demonstration experiment of real-time recognition. Figure 4 shows the recognition result. The system achieved the average frame rate of 138 fps and the latency of 29 ms. The number of candidates for template matching is 200 for a palm and 200 for each finger (1200 in total). The experimental results showed that our template-matching-based method was robust against a cluttered background.

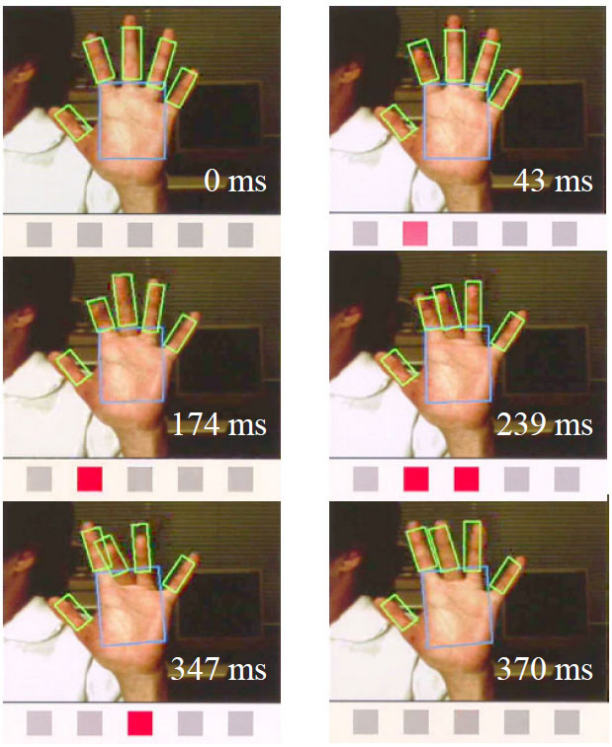


Fig. 4 Hand Tracking Result

5. CONCLUSION

As 3D input interfaces that solve the bottleneck of information devices by taking the advantages of vision-based UIs, we showed development examples of the touch panel interface which zooms the screen according to the depth position of a fingertip and the mobile input interface which uses fingertip movement in the air detected by a single camera as input.

Among the technological issues of vision-based UIs, temporal resolution and response time can be improved using a high-frame-rate camera. Meanwhile, a new region extraction algorithm is required to get more stability against background and disturbance.

Another advantage of vision-based UIs is that a large amount of information can be obtained. The future challenge is to achieve higher functionality and more conformable operability by complete pose estimation of the hand and fingers.

REFERENCES

[1] H. Roeber, J. Bacus, and C. Tomasi., “Typing in thin air the canesta projection keyboard - a new method of interaction with electronic devices,” CHI’03 extended abstracts, pp. 712-713 (2003)

[2] N. Fukuoka, T. Komuro, M. Ishikawa, “Zooming Touch Panel: Improving the Functionality of a Touch Panel using Small Cameras,” Proc. IPSJ Interaction 2007, pp.33-34 (2007) (in Japanese)

[3] Y. Hirobe, T.Niikura, Y. Watanabe, T. Komuro, M. Ishikawa, “Vision-based Input Interface for Mobile Devices with High-speed Fingertip Tracking,” Adj. Proc. ACM UIST 2009, pp.7-8 (2009)

[4] T. Niikura, Y. Hirobe, A. Cassinelli, Y. Watanabe, T. Komuro, M. Ishikawa, “In-air Typing Interface for Mobile Devices with Vibration Feedback,” Proc. ACM SIGGRAPH 2010 (2010)

[5] K. Terajima, T. Komuro, M. Ishikawa, “Fast Finger Tracking System for In-air Typing Interface,” CHI’09 extended abstracts, pp.3739-3744 (2009)