Fast Finger Tracking System for In-air Typing Interface

Kazuhiro Terajima
The University of Tokyo
7-3-1 Hongo, Bunkyo-ku,
Tokyo, 113-8656 Japan
Kazuhiro_Terajima@ipc.i.u-tokyo.ac.jp

Takashi Komuro
The University of Tokyo
7-3-1 Hongo, Bunkyo-ku,
Tokyo, 113-8656 Japan
Takashi_Komuro@ipc.i.u-tokyo.ac.jp

Masatoshi Ishikawa
The University of Tokyo
7-3-1 Hongo, Bunkyo-ku,
Tokyo, 113-8656 Japan
Masatoshi_Ishikawa@ipc.i.u-tokyo.ac.jp

CHI 2009, April 4 – 9, 2009, Boston, MA, USA
ACM 978-1-60558-247-4/09/04.

Abstract
We developed a system which performs 3D motion tracking of human's hand and fingers from images of a single high-frame-rate camera and that recognizes his/her typing motion in the air. Our template-matching-based method using hand textures reduces background effect and enables markerless tracking. In addition, use of a high-frame-rate camera enables recognition of rapid typing motion which is difficult to track using standard cameras. In order to realize real-time recognition, we developed hardware which parallelizes and accelerates image processing. As a result, we achieved real-time recognition of typing motion with the throughput of 138 fps (frames per second) and the latency of 29 ms.

Keywords
Vision-based UI, portable device, embedded computer vision.

ACM Classification Keywords
H5.2. Information interfaces and presentation (e.g., HCI): User Interfaces - Input devices and strategies.

Introduction
In recent years, portable information devices have become small and it is becoming difficult to have

enough space for character input and pointing operations on the device. Thus users are required to make delicate input operations on limited surfaces such as small keyboards and touch panels. To overcome this problem, an interface where 3D motion of a hand and fingers are input to the device may be effective.

While there has been much research on such interface systems, most of the systems require an operator to wear some physical device[1-4], or require flat space for operation[5]. On the other hand, some systems use a camera and realize non-contact hand gesture recognition. 3D shape reconstruction using multiple cameras[6] could obtain detailed hand shape, but it is ill-suited to portable devices in terms of size. Some hand gesture recognition system using a single camera use hand pose database[7,8], which requires large memory and it is difficult to implement in a portable device. 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.

In this research, we aim to realize recognition of hand and finger poses from the images of typing motion in the air which are captured by a single camera. The system works even in the presence of cluttered background without using markers. This will lead to a novel input interface illustrated in Fig. 1, which allows unconstrained, non-contact and spacious typing operations. In addition, use of high-frame-rate camera enables recognition of rapid typing motion which is difficult to track using standard cameras.

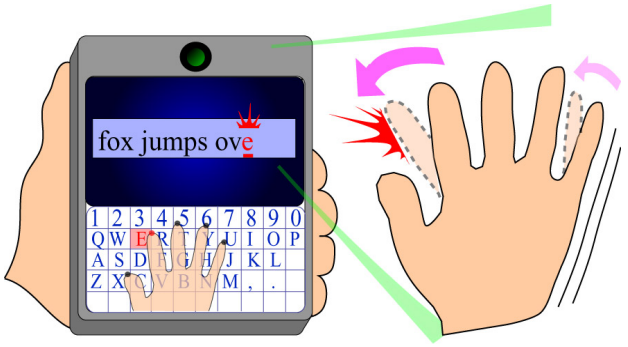

figure 1. In-air typing interface.

Algorithm for hand and finger tracking

For tracking of hand and finger motion, our algorithm assumes that the palm and each finger are flat, and estimates relative motion between hand/fingers and the camera. Our algorithm introduces a structural model of a hand and refines the search range by applying pose recognition of the palm and fingers in a stepwise fashion, which enables stable tracking.

Initialization

The template images for tracking are generated at initialization stage. In this study, the palm and finger areas are selected by manually pointing the feature points of the hand with the mouse. This process might be fully automated by a more sophisticated algorithm.

Preprocessing

Considering the use of the handheld device in our living environments, it is highly likely that illumination condition is not always constant. 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 (the pixel aspect ratio is the same as the template image). 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.

Pose Estimation

To improve stability of tracking and to reduce processing, our algorithm searches pose parameters in two stages. At first stage, the pose of the palm $\mathbf{X}_{\text{hand}}$ with its origin at wrist is estimated. Then, the bending and spreading angles of each finger $\mathbf{X}_{\text{finger}}$ with their origin at the base of finger are estimated. Each pose parameter contains the following components and the total degrees of freedom are 16 (six for the palm and two each for the five fingers).

$$\mathbf{X}_{\text{hand}} = \begin{bmatrix} \varphi_{\text{hand}} & \theta_{\text{hand}} & \psi_{\text{hand}} & x_{\text{hand}} & y_{\text{hand}} & z_{\text{hand}} \end{bmatrix}$$

$$\mathbf{X}_{\text{finger}} = \begin{bmatrix} \varphi_{finger} & \psi_{finger} \end{bmatrix}$$

Then, pose candidates are prepared. Using the pose in the previous frame $\mathbf{X}(t-1)$ and its first derivative $\dot{\mathbf{X}}(t-1)$, n candidates $\hat{\mathbf{X}}_i(t)\,(i=1,\ldots,n)$ are generated by the following equation, where $\mathbf{m}_i$ is a Gaussian noise with $\mathbf{0}$ mean.

$$\hat{\mathbf{X}}_i(t) = \mathbf{X}(t-1) + \dot{\mathbf{X}}(t-1)\Delta t + \mathbf{m}_i$$

A template image $T(x,y)$ is perspectively transformed by the parameters of each candidate using the following equations, where f is the focal length of the lens, $\mathbf{R}$ and $\mathbf{t}$ are rotation matrix and translation vector formed from the pose parameters, $\mathbf{n}$ is a normal vector to the plane, and d is a distance between the origin and the plane.

$$s\begin{pmatrix} x' \\ y' \\ 1 \end{pmatrix} = \begin{pmatrix} f & 0 & 0 \\ 0 & f & 0 \\ 0 & 0 & 1 \end{pmatrix} \mathbf{H} \begin{pmatrix} x \\ y \\ 1 \end{pmatrix}$$

$$\mathbf{H} = \mathbf{R} + \frac{\mathbf{t} \cdot \mathbf{n}^{\mathsf{T}}}{d}$$

For each candidate, SAD (Sum of Absolute Difference) between the transformed template and input image is calculated. The candidate whose SAD is minimum is selected as the current pose parameter $\mathbf{X}(t)$.

Experiment

To evaluate the stability of the algorithm described above, we conducted experiments applying the algorithm on the pre-recorded images captured by a high-frame-rate camera. 320 x 240 pixel, 120 fps, 6 second video images are used as input. The number of candidates for template matching is 200 for a palm and 200 for each finger (1200 in total). Figure 2 and 3 shows the results of hand tracking. The hand is stably tracked even in the presence of background and under varying illumination.

Next, we conducted a recognition experiment of typing gesture with a finger bending inside as shown in Fig. 4 repeated (a) 1.2 times per second and (b) 5 times per second. The images are captured at 120 fps, 60 fps, and 30 fps. The red lines in Fig. 5 indicate raw data of estimated finger angles.The periodic pulses are seen in the graph, which reflects typing gesture of the finger. Some pulses are missing in the graph of (a)-30 fps and (b)-60 fps and no pulses are seen in the graph of (b)-30 fps. These results show that a frame rate of at least 60 fps is required to recognize slow typing gesture

repeated 1.2 times per second and 120 fps is required to recognize rapid typing gesture repeated 5 times per second.

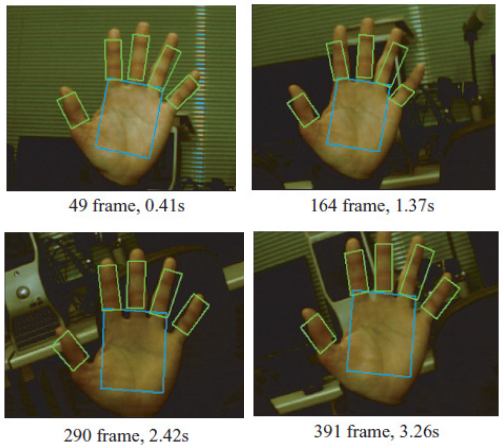

figure 2. Tracking result in cluttered background.

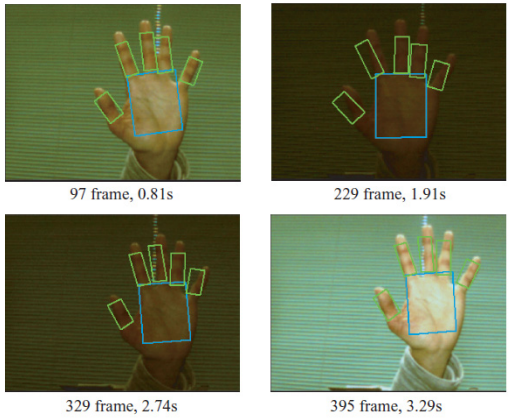

figure 3. Tracking result under varying illumination.

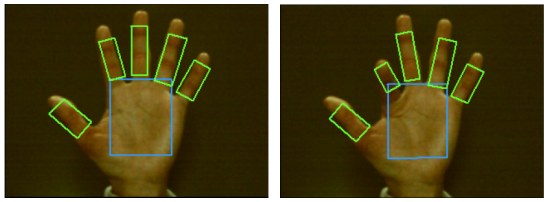

figure 4. Typing Gesture.

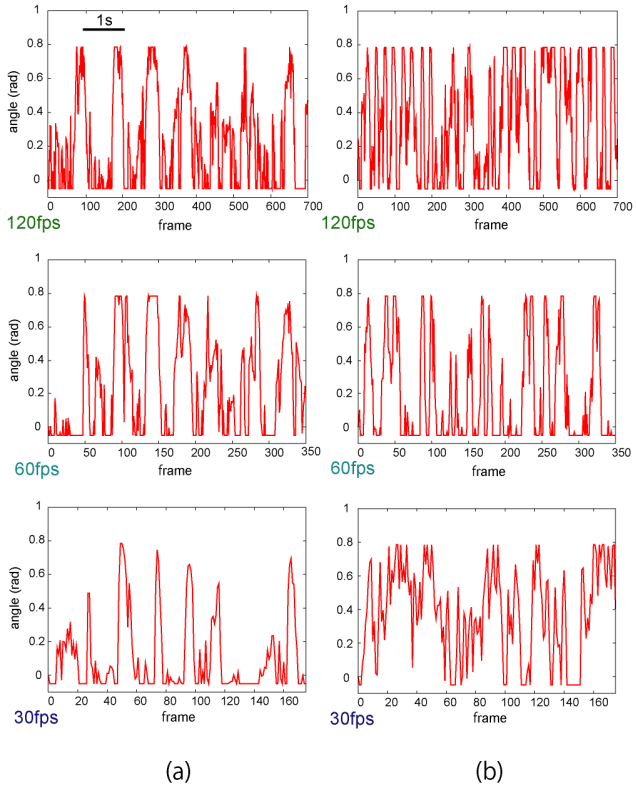

figure 5. Recognition result of typing gesture.

Hardware acceleration

The processing time of the algorithm described above in a PC (Intel Core2Duo E6750 2.66GHz, 2GB memory, Windows XP) is about 0.4 s per frame. To accelerate this, we developed a hardware system with FPGA coprocessor.

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.

In this prototype, we used a standard PC (Intel Core2Duo E6750 2.66GHz, 2GB memory, Windows XP) as a main processor, and Artray ARTCAM-036MI as a high-frame-rate-camera (up to 200 fps @ 320 x 240 pixels, USB 2.0 interface) with a wide-angle lens (f=4.2 mm).

The coprocessor is implemented in a Xilinx FPGA Virtex4 XC4VLX25. The coprocessor works at 180 MHz, and can execute template matching with perspective transformation 260,000 times per second.

Since most of computation is performed in the coprocessor, the system can be implemented in an embedded hardware with lower-performance processor.

Real-time recognition

Using the developed system, we conducted a demonstration experiment of real-time recognition.

Figure 6 shows the appearance of the system and Fig. 7 shows the recognition result. The system achieved the average frame rate of 138 fps and the latency of 29 ms (=4 frames). The processing mainly consists of image preprocessing on the PC, data transfer between the PC and the FPGA, and template matching on the FPGA, and they are properly pipelined.

This result shows that the use of high-frame-rate camera and hardware acceleration enabled real-time recognition of typing gesture in the air.

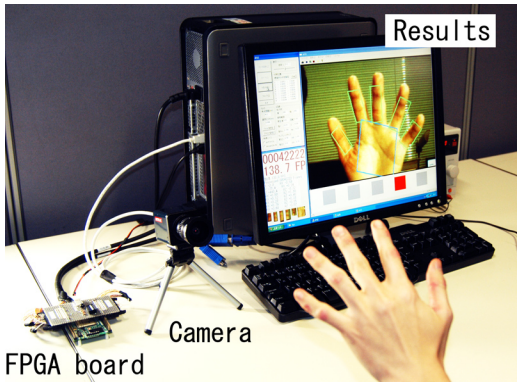

figure 6. In-air typing recognition system.

Conclusion

In this study, we proposed an algorithm that tracks a human hand and recognizes typing motion in the air. The experimental results showed that our template-matching-based method was robust against a cluttered background and varying illumination, and that a high frame rate over 100 fps is required to recognize rapid typing gesture. Using a commercially-available USB high-speed camera and an FPGA board, a prototype

system was constructed and the system achieved real-time recognition of typing motion with the throughput of 138 fps and the latency of 29 ms. Future work includes quantitative evaluation of the algorithm such as accuracy and recognition rate, analysis of required typing speed and accuracy for the proposed interface, and usability evaluation of this typing approach with an actual working system.

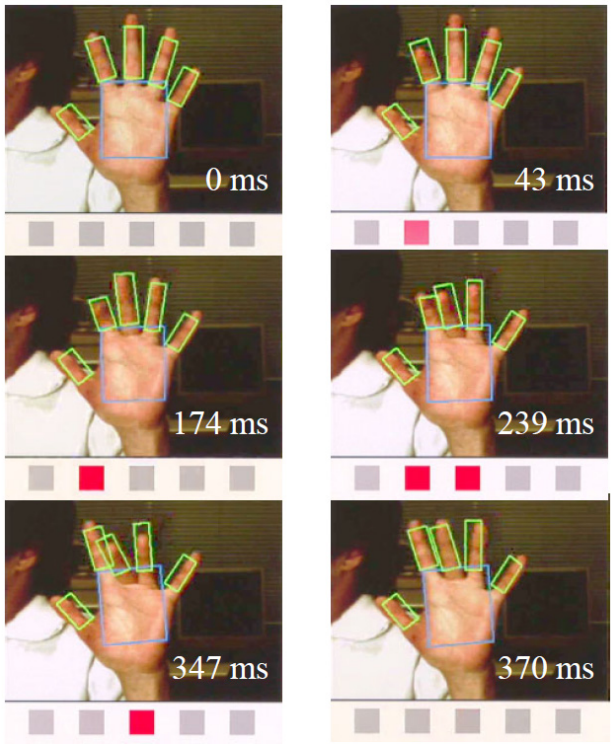

figure 7. Real-time recognition of in-air typing.

References

[1] Kölsch, M. and Turk, M. Keyboards without Keyboards: A Survey of Virtual Keyboards, Proc. Workshop/Symposium on Sensing and Input for Media-centric Systems (2002).

[2] Ahmad, F. and Musilek, P. UbiHand: a wearable input device for 3D interaction, ACM SIGGRAPH Research posters (2006), 159.

[3] Kim, S. and Kim, G. Using keyboards with head mounted displays, Proc. ACM SIGGRAPH international conference on Virtual Reality continuum and its applications in industry (2004), 336-343.

[4] Imura, M., Fujimoto, M., Yasumuro, Y., Manabe Y., and Chihara, K. AirGrabber: virtual keyboard using miniature infrared camera and tilt sensor, Proc. ACM international conference on Augmented tele-existence (2005), 277-277.

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

[6] Ueda, E., Matsumoto, Y., Imai, M. and Ogasawara, T. Hand Pose Estimation for Vision-based Human Interface. IEEE Transactions on Industrial Electronics 50, 4 (2003), 676-684.

[7] Stenger, B., Thayanathan, A., Torr, P. H. S. and Cipolla, R. Model Based Hand Tracking Using a Hierarchical Bayesian Filter, IEEE Transactions on Pattern Analysis and Machine Intelligence 28, 9 (2006), 1372-1384.

[8] Hoshino, H. and Tanimoto, T. Realtime estimation of human hand posture for robot hand control, Proc. IEEE International Symposium of Computational Intelligence in Robotics and Automation (2005), 1-6.