Fast Tracking of Hands and Fingertips in Infrared Images for Augmented Desk Interface

Yoichi Sato
Institute of Industrial Science
University of Tokyo
7-22-1 Roppongi, Minato-ku
Tokyo 106-8558, Japan
ysato@cvl.iis.u-tokyo.ac.jp

Yoshinori Kobayashi Hideki Koike
Graduate School of Information Systems
University of Electro-Communications
1-5-1 Chofugaoka, Chofu
Tokyo 182-8585, Japan
yosinori, koike@vogue.is.uec.ac.jp

Abstract

In this paper, we introduce a fast and robust method for tracking positions of the centers and the fingertips of both right and left hands. Our method makes use of infrared camera images for reliable detection of user's hands, and uses template matching strategy for finding fingertips. This method is an essential part of our augmented desk interface in which a user can, with natural hand gestures, simultaneously manipulate both physical objects and electronically projected objects on a desk, e.g., a textbook and related WWW pages. Previous tracking methods which are typically based on color segmentation or background subtraction simply do not perform well in this type of application because an observed color of human skin and image backgrounds may change significantly due to projection of various objects onto a desk. In contrast, our proposed method was shown to be effective even in such a challenging situation through demonstration in our augmented desk interface. This paper describes the details of our tracking method as well as typical applications in our augmented desk interface.

1. Introduction

One of the important challenges in Computer Human Interactions is to develop more natural and more intuitive interfaces. Graphical user interface (GUI), which is a current standard interface on personal computers (PCs), is well-matured, and it provides an efficient interface for a user to use various applications on a computer. However, many users find that the capability of GUI is rather limited when they try to do some tasks by using both physical documents on a desk and computer applications. This limitation comes from the lack of seamless integration between two different types of interface. One is interface for using physical objects such as books on a desk. The other is GUI for using computer programs. As a result, users have to keep switching their focus of attention between physical objects on a desk and GUI on a computer monitor.

One of the earliest attempts to provide seamless integration between those two types of interface, i.e., interface for using physical objects and GUI for using computer programs, was reported in Wellner's DigitalDesk [13]. In this work, the use of a desk equipped with a CCD camera and a video projector was introduced.

Inspired by this DigitalDesk, we have proposed an augmented desk interface in our previous work[4] in order to allow a user to perform various tasks by manipulating both physical objects and electronically displayed objects simultaneously on a desk. In basic demonstrations, our augmented interface system was shown to be able to provide intuitive interface for using physical objects and computer programs simultaneously. Unfortunately, however, applications of the proposed system were limited to rather simple ones mainly due to a limited capability of monitoring user's movements, e.g., hand gestures, in a non-trivial environment in real-time. Therefore, a user was allowed to use only a limited range of hand gestures on an uncluttered desk.

In this paper, we introduce a new method for tracking a user's hands and fingertips reliably at video-frame rate. Our method makes use of infrared camera images for reliable detection of user's hands, and uses template matching strategy for finding fingertips accurately. The use of an infrared camera is especially advantageous for our augmented desk interface system where observed color of human skin and image background change significantly due to projection onto a desk. In contrast, previous methods for finding hand and fingertips are typically based on color segmentation or background subtraction, and therefore those methods would have difficulties in such challenging situations.

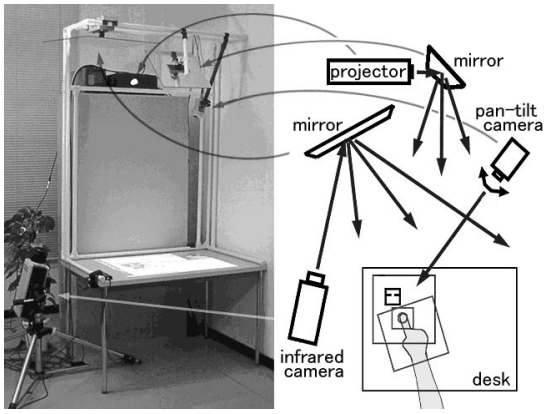

Figure 1. Overview of our augmented desk interface system

This paper is organized as follows. In Section 2, we describe the previously proposed methods for tracking human hands and fingertips; we also include the limitations of these methods. In Section 3, we explain our method for fast and robust tracking of a user's hand location and fingertips by using an infrared camera. In Section 4, we show examples of tracking results by our proposed method. In Section 5, we describe our augmented desk interface system which is based on our hand and fingertip tracking method. Finally, in Section 6, we present our conclusions and discussion for future research directions.

2. Related works

In this section, we give a brief overview of the previously proposed methods for tracking a human hand and its fingertips, and examine the limitations of these methods.

The use of glove-based devices for measuring the location and shape of a user's hand has been widely studied in the past, especially in the field of virtual reality. Angles of finger joints are measured by some sort of sensor, typically mechanical or optical. The position of a hand is determined by an additional sensor. One of the most widely known examples of such devices is DataGlove by VPL Research [14] which uses optical fiber technology for flexion detection and a magnetic sensor for hand position tracking. A good survey of glove-based devices can be found in [12].

In general, glove-based devices can measure hand postures and locations with high accuracy and high speed. However, the use of glove-based devices is not suitable for some types of applications such as human computer interfaces because those devices may limit a user's motion due to the physical connection to their controllers.

For this reason, a number of methods based on computer vision techniques have been studied by other researchers in the past. One approach is to use some kind of markers attached to a user's hand or fingertips, so that those points can be easily found. For instance, color markers attached to the fingertips are used in the method reported in [1] to identify locations of fingertips in input images. Maggioni [5] presented the use of a specially marked glove for hand tracking. The glove has two slightly off-centered, differently colored circular regions. By identifying those two circles with a single camera, the system can estimate hand position and orientation.

In another approach, image regions corresponding to human skin are extracted typically either by color segmentation or by background image subtraction. The main challenge of this approach is how to identify the image regions in input images. Since the color of human skin is not completely uniform and changes from person to person, the methods based on color segmentation often produce unreliable segmentation of human skin regions. To avoid this problem, some methods requires a user to wear a glove of a uniform color. On the other hand, the methods based on background image subtraction have difficulties when applied to images with a complex background.

After the images regions are identified in input images, the regions are analyzed to estimate the location and orientation of a hand, or to estimate locations of fingertips. For instance, in the method by Maggioni etal.[6], a shape of the contour of an extracted hand region is used for determining locations of fingertips. Segen and Kummar [11] introduced a method which fits a line segment to a hand region contour to locate the side of an extended finger.

All of the methods based on extraction of hand regions face a common difficulty when they are applied in the situation assumed in our application, i.e., augmented desk interface. Since our augmented desk interface system can project various objects such as texts or figures with different colors onto a user's hand on the desk, hand regions cannot be identified by color segmentation or background segmentation.

Another approach used in hand gesture analysis is to use a three dimensional model of a human hand. In this approach, in order to determine the posture of the hand model, the model is matched to a user's hand images which have been obtained by using one or more cameras. The method proposed by Rehg and Kanade[9] is one example based on this approach. Unlike other methods which do not use a three dimensional hand model, the method proposed by Rehg and Kanade can estimate three dimensional posture of a user's hand.

However, this approach faces several difficulties such as self-occlusion of a hand or high computational cost for es-

timation of hand posture. Due to the high degrees of freedom of the hand model, it is very difficult to estimate the hand configuration from a two dimensional image even if images are obtained from multiple viewpoints.

In addition to the methods mentioned in this section, a large number of methods were proposed in the past. A good survey of hand tracking methods as well as algorithms for hand gesture analysis can be found in [3] and [8].

3. Real-time tracking of fingertips in IR images

Unfortunately, none of the previously proposed methods for hand tracking provides the capability necessary for our augmented desk interface system. To realize a natural and intuitive interface to manipulate both physical objects and electrically projected objects on a desk, the system needs to be able to track a user's hand and fingertip locations in complex environments in real-time without relying on markers or marked gloves attached to the user's hand.

In this work, we propose a new method for tracking a user's palm center and fingertips by using an infrared camera.[1] The use of an infrared camera is especially advantageous for our augmented desk interface system where observed color of human skin and image background change significantly due to projection onto a desk. Unlike regular CCD cameras which detect light in visible wavelengths, an infrared camera can detect light emitted from a surface with a certain range of temperature. Thus, by setting the temperature range to approximate human body temperature, image regions corresponding to human skin appear particularly bright in input images from the infrared camera.

3.1 Extraction of left and right arms

To extract right or left arms, an infrared camera is installed with a surface mirror so that user's hands on a desk can be observed by the camera as shown in Fig.1.

The video output from the infrared camera is digitized as a gray-scale image with 256×220 pixel resolution by a frame grabber on a PC. Because the infrared camera is adjusted to measure a range of temperature that approximates human body temperature, e.g,. typically between 30^o and 34^o, values of image pixels corresponding to human skin are higher than other image pixels. Therefore, image regions which correspond to human skin can be easily identified by binarization of the input image with a threshold value. In our experiments, we found that a fixed threshold value for image binarization works well for finding human skin regions regardless of room temperatures. Fig.2(a) and (b) show one example of an input image from the infrared camera, and a region of human skin extracted by binarization of the input image, respectively.

If other objects on a desk happen to have temperatures similar to that of human skin, e.g., a warm cup or a note PC, image regions corresponding to those objects in addition to human skin are found by image binarization. To remove those regions other than human skin, we first remove small regions, and then select the two regions with the largest size. If only one large region is found, we consider that only one arm is observed on the desk.

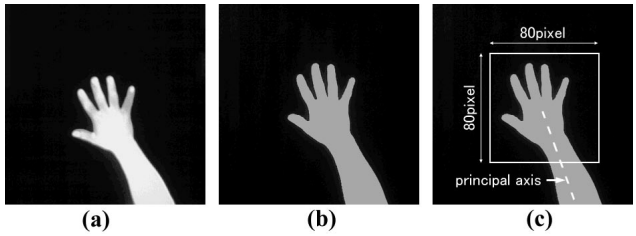

Figure 2. Extraction of hand region

3.2 Finding fingertips

Once regions of user's arms are found in an input image, fingertips are searched for within those regions. Compared with the extraction of users' arms, this search process is more computationally expensive. Therefore, a search window is defined in our method, and fingertips are searched for only within the window instead of over entire regions of users' arms.

A search window is determined based on the orientation of each arm which is given as the principal axis of inertia of the extracted arm region. The orientation of the principal axis can be computed from the image moments up to the second order as described in [2]. Then, a search window of a fixed size, e.g., 80×80 pixels in our current implementation, is set so that it includes a hand part of the arm region based on the orientation of the arm. (Fig.2(c)) We found that a fixed size for the search window works reliably because the distance from the infrared camera to a user's hand on a desk remains relatively constant.

[1] The use of an infrared camera was examined in another study for human body posture measurement [7]. In that case, human body postures were determined by extracting human body regions in infrared images, and then by analyzing contour of the extracted regions. While our method is used for accurate estimation of distinct feature points such as palm centers and fingertips, their method was designed to estimate a rough posture of a human body, e.g., the orientation of a body and the direction of two arms.

Once a search window is determined for each hand region, fingertips are searched for within that window. The overall shape of a human finger can be approximated by a cylinder with a hemispherical cap. Thus, the projected shape of a finger in an input image appears to be a rectangle with a semi-circle at its tip.

Based on this observation, fingertips are searched for by template matching with a circular template as shown in Fig.3 (a). In our proposed method, normalized correlation with a template of a fixed-size circle is used for the template matching. Ideally, the size of the template should differ for different fingers and different users. In our experiments, however, we found that the fixed size of template works reliably for various users. For instance, a square of 15×15 pixels with a circle whose radius is 7 pixels is used as a template for normalized correlation in our current implementation.

While a semi-circle is a reasonably good approximation of the projected shape of a fingertip, we have to consider false detection from the template matching. For this reason, we first find a sufficiently large number of candidates. In our current implementation of the system, 20 candidates with the highest matching scores are selected inside each search window. The number of initially selected candidates has to be sufficiently large to include all true fingertips.

After the fingertip candidates are selected, false candidates are removed by means of two types of false detection. One is multiple matching around the true location of a fingertip. This type of false detection is removed by suppressing neighbor candidates around a candidate with the highest matching score.

The other type of false detection is a matching that occurs in the middle of fingers such as the one illustrated in Fig.3 (b). This type of falsely detected candidates is removed by examining surrounding pixels around the center of a matched template. If multiple pixels in a diagonal direction are inside the hand region, then it is considered not to exist at a fingertip, and therefore the candidate is discarded.

By removing these two types of false matchings, we can successfully find correct fingertips as shown in Fig.3 (c).

3.3 Finding centers of palms

The center of a user's palm needs to be determined to enable recognition of various types of hand gestures. For example, the location of the center is necessary to estimate how extended each finger is, and therefore it is essential for recognizing basic gestures such as click and drag.

In our proposed method, the center of a user's hand is given as the point whose distance to the closest region boundary is the maximum. In this way, the center of the hand becomes insensitive to various changes such as opening and closing of the hand. Such a location for the hand's center is computed by morphological erosion operation of an extracted hand region. First, a rough shape of the user's palm is obtained by cutting out the hand region at the estimated wrist as shown in Fig.4 (a). The location of the wrist is assumed to be at the pre-determined distance, e.g., 60 pixels in our case, from the top of the search window and perpendicular to the principal direction of the hand region.

Then, a morphological erosion operator is applied to the obtained shape of the user's palm until the area of the region becomes small enough. As a result, a small region at the center of the palm is obtained. Finally, the center of the hand region is given as the center of mass of the resulting region as shown in Fig.4 (c).

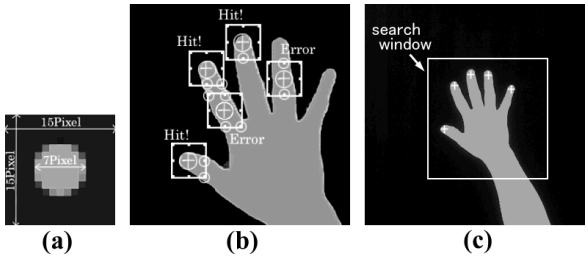

Figure 3. Template matching for fingertips

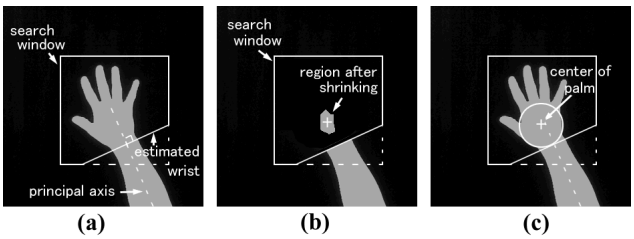

Figure 4. Center of a user's palm

4. Performance

We have tested our proposed method by using the system shown in Fig.1. An infrared camera (NIKON Thermal Vision LAIRD-3A) is installed with a surface mirror, so that a user's hand on a desk can be observed. Input images from the infrared camera are processed as described in this paper in Section3 on a personal computer (Hardware: PentiumIII 450MHz, OS: Linux kernel 2.0.36) with a general-purpose image processing board (HITACHI IP-5010).

Several examples of tracking results are shown in Fig.5. These results show that our proposed method successfully finds centers of palms and locations of fingertips. Centers of palms are found reliably regardless of opening of a hand. Also, fingertips are found successfully by our method even when fingers are not fully extended. This is a case where the previously proposed methods based on shape of contour of hand regions often have difficulties.

While we have not yet done any careful optimization of the codes, the current implementation of our system is running almost in real-time, e.g., approximately 25-30 frames per second for one hand. The system suuccesfully finds fingertips and palm centers for both left and right hands. However, in this case, processing speed becomes somewhat lower due to the doubled area for searching for fingertips. If two hands are tracked, the system runs at approximately 15 frames per second.

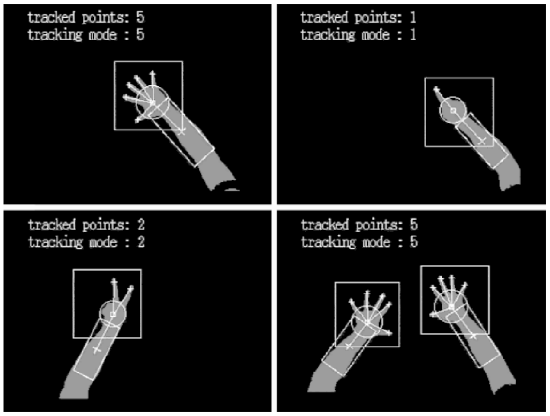

Figure 5. several examples of tracking results

5. Augmented desk interface system

The proposed method for tracking hands and fingertips in infrared images was successfully used for our augmented desk interface system. As shown in the system overview in Fig.1, the system is equipped with an LCD projector, an infrared camera, and a pan-tilt camera. The LCD projector is used for projecting various kinds of digital information such as computer graphics objects, text, or a WWW browser on a desk.

For alignment between an image projected onto a desk by the LCD projector and an input image from the infrared camera, we determine a projective transformation between those two images through initial calibration of the system. The use of projective transformation is enough for calibration of our system since imaging/projection targets can be approximated as to be planer due to the nature of our application. In addition, a similar calibration is also carried out for the pan-tilt camera so that the camera can be controlled to look toward a desired position on the desk.

The pan-tilt camera is controlled to follow a user's fingertip whenever the user points at a particular location on the desk with one finger. This is necessary to obtain enough image resolution to recognize real objects near a user's pointing finger. Currently-available video cameras simply do not provide enough image resolution when the entire table is observed. In our current implementation of the interface system, a two-dimensional matrix code [10] is used for identifying objects on the desk. (See Fig.7 for example). More sophisticated computer vision methods would be necessary for recognizing real objects without any markers.

Using our augmented desk interface system, we have tested various applications in which a user manipulates both physical objects and electronically projected objects on the desk. Fig.6 shows one example of such applications. In this example, a user can manipulate a projected object on a desk using both left and right hands. By bending his/her forefinger at an end of the object, a user can grab the object's end. Then, a user can translate, rotate, and stretch the object by two-handed direct manipulation.

Fig.7 shows another application example of our augmented desk interface system. In this application, a user can browse WWW pages associated with physical documents on a desk by simply pointing to those documents with his/her forefinger. With the pan-tilt camera, which follows a user's forefinger on the desk, a small two-dimensional matrix code attached to a physical document is recognized. Once a physical document is found on the desk, associated WWW pages are projected directly next to the document.

6. Conclusions

In this paper, we have proposed a fast and reliable method for tracking a user's palm centers and fingertips for both left and right hands. Our method makes use of infrared camera images and template matching by normalized correlation which is performed efficiently with a general-purpose image processing hardware. In particular, our method is effective for applications in our augmented desk interface system where observed color of human skin and image backgrounds continuously change due to projections by an LCD projector. While previous methods based on color segmentation or background image subtraction would have difficulties for tracking hands or finertips, our proposed method was demonstrated to perform very reliably even in this situation.

Currently, we are extending our method so that not only all fingertips can be found, but also so that those

fingertips can be distinguished from one another, e.g. the fingertip of an index finger and that of a middle finger. Also, based on our tracking method, we are endeavoring to enhance our augmented desk interface system with more sophisticated gesture recongnition capabilities.

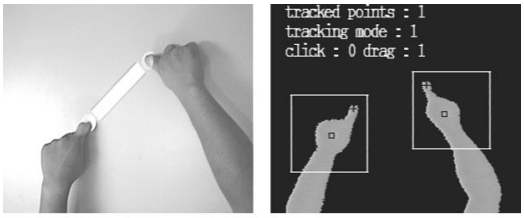

Figure 6. Direct manipulation of a CG object with two hands

Figure 7. Web browsing in our augmented desk interface system

References

[1] R. Cipolla, Y. Okamoto, and Y. Kuno, "Robust structure from motion using motion parallax," *Proc. 1999 IEEE International Conference on Computer Vision*, pp. 374-382, 1993.

[2] W. T. Freeman, D. B. Anderson, P. A. Beardsley, C. N. Dodge, M. Roth, C. D. Weissman, and W. S. Yerazunis, "Computer vision for interactive computer graphics," *IEEE Computer Graphics and Applications*, vol. 18, no. 3, pp. 42-53, May-June 1998.

[3] T. S. Huang and V. I. Palvovic, "Hand gesture modeling, analysis, and synthesis," *Proc. of 1995 IEEE International Workshop on Automatic Face and Gesture Recognition*, pp. 73-79, September 1995.

[4] M. Kobayashi and H. Koike, "Enhanced Desk, integrating paper documents and digital documents," *Proc. 1998 Asia Pacific Computer Human Interaction*, pp. 167-174, 1998.

[5] C. Maggioni, "A novel gestural input device for virtual reality," *Proc. 1993 IEEE Annual Virtual Reality International Symposium*, pp. 118-124, 1993.

[6] C. Maggioni and B. Kammerer, "GestureComputer - history, design and applications," *Computer Vision for Human-Machine Interaction (R. Cipolla and A. Pentland, eds.)*, pp. 23-51, Cambridge University Press, 1998.

[7] J. Ohya, "Virtual kabuki theater: towards the realization of human metamorphosis systems," *Proc. 5th IEEE International Workshop on Robot and Human Communication*, pp. 416-421, November 1996.

[8] V. I. Pavlovic, R. Sharma, and T. S. Huang, "Visual interpretation of hand gestures for human-computer interaction: a review," *IEEE Trans. PAMI*, Vol. 19, No. 7, pp. 677-695, July 1997.

[9] J. Rehg and T. Kanade. "Digiteyes: Vision-based hand tracking for human-computer interaction," *Proc. Workshop on Motion of Non-Rigid and Articulated Objects*, pp. 16-22, Austin, Texas, November 1994.

[10] J. Rekimoto, "Matrix: a realtime object identification and registration method for augmented reality," *Proc. 1998 Asia Pacific Computer Human Interaction (APCHI'98)*, July 1998.

[11] J. Segen and S. Kumar, "Shadow gestures: 3D hand pose estimation using a single camera," *Proc. 1999 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 479-485, June 1999.

[12] D. J. Sturman and D. Zeltzer, "A survey of glove-based input," *IEEE Computer Graphics and Applications*, Vol. 14, pp. 30-39, January 1994.

[13] P. Wellner, "Interacting with paper on the DIGITAL DESK," *Communications of The ACM*, Vol. 36, No. 7, pp. 87-96, July 1993.

[14] T. G. Zimmermann, J. Lanier, C. Blanchard, S. Bryson, and Y. Harvill, "A hand gesture interface device," *Proc. ACM Conf. Human Factors in Computing Systems and Graphics Interface*, pp. 189-192, 1987.