A Hand Tracking for a Human Computer Interaction System by the Modified Block Matching Algorithm

Jin Ok Kim

School of Information and Communication Engineering, Sungkyunkwan University,
300, Chunchun-dong, Jangan-gu, Suwon, Kyunggi-do, 440-746, KOREA
jinny@ece.skku.ac.kr

Abstract. A GUI (graphical user interface) has been a dominant platform for HCI (human computer interaction). A GUI-based interaction has made computers simpler and easier to use. The GUI-based interaction, however, does not easily support the range of interaction necessary to meet users' needs that are natural, intuitive, and adaptive. In this paper, the modified BMA (block matching algorithm) is proposed to track a hand in a sequence of an image and to recognize it in each video frame in order to replace a mouse with a pointing device for a virtual reality. The HCI system with 30 frames per second is realized in this paper. The modified BMA is proposed to estimate a position of the hand and segmentation with an orientation of motion and a color distribution of the hand region for real-time processing. The experimental result shows that the modified BMA with the YCbCr (luminance Y, component blue, component red) color coordinate guarantees the real-time processing and the recognition rate. The YCbCr color coordinate requires less bits to be coded than the RGB (red, green, blue) color coordinate with omitting the luminance of each pixel color and has less sensitivity to surrounding circumstances. The hand tracking by the modified BMA can be applied to a virtual reality or a game or an HCI system for the disable.

1 Introduction

Hand gesture recognitions from video images have been studies with a lot of interests as a means of providing simple and intuitive human-computer interfaces [1] [2]. Recently as the new wave of portable computing, the era of pen computer without keyboard and the core technology of pen computing consist of hand-written characters and gesture recognitions [3] [4] [5]. Most of previous gesture-based applications, however, need a special hardware such as a mouse, a stylus pen, or a data glove in order to input a gesture. Although current gesture-input devices are well improved, it is more or less unnatural to make gestures by using special hardware, especially for a novice who is afraid to use a computer device. In this paper a method is proposed to recognize gestures by extracting hand regions and to control a window manager from these recognition results. Processing color-image with a skin-color model and its labeling are used to extract a hand

image. Its trajectory is obtained by using the modified BMA (block matching algorithm) from each frame of a sequence of an input image [6] [7] [8]. When this trajectory is analyzed to recognize a gesture according to the predefined window-management rules, the gesture-recognition technique is used to control the display by the window manager [9] [10] [11].

2 Hand Segmentation

The hand-region segmentation using color information consists of constructing a model-database and extracting a process of hand region segmentation. Since gray coordinate includes only the normal information of each pixel, the hand region included in some kind of textural image can be hardly extracted. Thus, in this paper, the hand region is extracted by the color coordinate that uses the YCbCr (luminance Y, component blue, component red) color coordinate that can separate luminance and chrominance. The YCbCr color coordinate requires less bits to be coded than the RGB (red, green, blue) color coordinate with omitting the luminance of each pixel color and is less sensitive to the circumstance.

2.1 Hand Region Segmentation

The hand-region segmentation algorithm involves the use of color information in a fast low-level region segmentation process [12] [13] [14]. The aim is to classify the pixels of an input image into skin color and non-skin color [15] [16] [17] [18]. A skin-color reference map is used in the YCbCr color space.

The skin-color region can be identified by a presence of a certain set of chrominance (i.e., Cb and Cr) values narrowly and consistently distributed in the YCbCr color space. $\text{Range}\{\text{Cb}\}$ and $\text{Range}\{\text{Cr}\}$ are denoted as the respective ranges of Cb and Cr values that correspond to skin color that subsequently defines the skin-color reference map.

The different skin color perceived from the video image cannot be differentiated from the chrominance information of the image region. Such a map derived from Cb and Cr values remains effective regardless of a skin-color variation. The difference of skin color of all races is mainly due to the darkness or fairness of the skin. The feature is characterized by the difference in the brightness of the color that is governed by Y but not Cr and Cb.

With the skin-color reference map, the color segmentation can be taken by (1). The segmentation requires only the chrominance component of the input image.

$$\text{Out} = \begin{cases} 1, & \text{if } [\text{Cr}(x, y) \in \text{Range}\{\text{Cr}\}] \cup [\text{Cb}(x, y) \in \text{Range}\{\text{Cb}\}] \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

The output pixel is classified as a skin color and set to one if both the Cb and Cr values at that point fall inside their respective ranges. Otherwise, the pixel is classified as a non-skin color and set to zero.

A median filter is used to reduce some blurs of edges, replacing the current point in the image with the median of the brightness in its neighborhood. Since the median of the brightness in the neighborhood is not affected by individual noises of spikes, the median filter can eliminate an impulse noise quite well. And furthermore, it can be applied iteratively since median filtering does not blur edges much.

2.2 Edge Extraction

The convolution masks in Figure 1 are used to detect edges. It is very easy to compute as it uses only a 3×3 neighborhood of the current pixel.

-1	0	0
0	0	0
0	0	1

0	0	1
0	0	0
-1	0	0

Fig. 1. Convolution Masks for Edge Detection

Since it has a property of a correlation mask, the magnitude of the edge is computed by (2).

$$\text{Threshold} = |P(i+1, j+1) - P(i-1, j-1)| + |P(i-1, j+1) - P(i+1, j-1)| \quad (2)$$

The primary advantage of the convolution masks is fast computational time despite of its high sensitivity.

3 Motion Estimation

3.1 The Block Matching Algorithm

The BMA is a popular correlation-based approach to motion estimation and motion tracking. Direction and trajectory of hand gestures are derived upon a distribution and a direction of the extracted motion vectors. Hence, the distribution of the motion vectors is statistically processed and applied to a hand gesture recognition.

An image is partitioned into a set of nonoverlapped and equally spaced and rectangular blocks, assuming that the translation motion within each block is uniform. Although this simple model considers translation motion only, other types of motion, such as rotation and zooming of large objects, may be closely

approximated by the piecewise translation of these small blocks provided that these blocks are small enough.

Motion vectors for these blocks are estimated by finding their best matched counterparts in the previous frame. Since the motion of each block is described by only one motion vector, the side information on motion vectors decreases and the rectangular shape information is known to both of the encoder and the decoder. The block size needs to be chosen properly. In general, the smaller is the block size, the more accurate is the approximation. It is apparent, however, that the smaller block size leads to more motion vectors. As a compromise, a size of 16×16 is considered to be a good choice. This has been specified in international video coding standards such as H.261, H.263 and MPEG-1 and MPEG-2.

3.2 Matching Criteria

The matching of the blocks can be quantified according to various criteria including the maximum cross-correlation, the minimum MSE (mean square error), the minimum MAD (mean absolute difference), and the maximum MPC (matching per count).

In the minimum MSE criterion, the MSE is defined by (3).

$$\text{MSE}(d_1, d_2) = \frac{1}{N_1 N_2} \sum_{(n_1, n_2) \in \beta} [F_k(n_1, n_2) - F_{k+1}(n_1 + d_1, n_2 + d_2)]^2 \quad (3)$$

where β denotes an $N_1 \times N_2$ block for a set of candidate motion vectors (d_1, d_2) . The estimate of the motion vectors is taken to be the value of (d_1, d_2) which minimizes the MSE

$$[d_1, d_2]^T = \arg \min \text{MSE}(d_1, d_2) \quad (4)$$

Minimizing the MSE criterion can be viewed as the optical flow constraint on all pixels of the block. In fact, expressing the displaced frame difference,

$$df \ d_k, d_{k+1}(n_1, n_2) = [F_k(n_1, n_2) - F_{k+1}(n_1 + d_1, n_2 + d_2)] \quad (5)$$

The minimum MSE criterion, however, is not commonly used in hardware implementations because it is difficult to realize the square operation in hardware. Instead, the minimum MAD criterion defined by (6) is widely used in practice.

$$\text{MAD}(d_1, d_2) = \frac{1}{N_1 N_2} \sum_{(n_1, n_2) \in \beta} |F_k(n_1, n_2) - F_{k+1}(n_1 + d_1, n_2 + d_2)| \quad (6)$$

3.3 The Modified BMA (Block Matching Algorithm)

Since the BMA assumes that each pixel's movement is same in block and movement is not found over the limits of a search window size, the assumption is a cause of the blocking effect and the measurement error of movement. Thus in order to compensate the blocking effect and the prediction error of motion vector, the size of the search window is set in more appropriate size. Since 100 random variables were made in the search window and a center point of correlation window is located in each point, we can first find a point whose MAD has a minimum value. And the search window is made in 16×16 size by using above method, finally, 50 random variables are fixed in most of proper motion vectors. Many problems, such as motion vector locations, which are caused by an initial value in three steps or 2-D algorithm can be solved by this algorithm and can be found in type of the proper distribution.

Because the BMA is acquired under assumption that all pixel's movement is same in a block, it is difficult to find the movement of specific part. But an indicating point and a click point we need are tips of finger. The difference between the tip of finger and the back of hand is classified to find tips of finger. When any specific point is indicated, movement of tips is relatively bigger than movement of back. After the size of motion vectors is first represented in histogram, the proper value is fixed in the upper distribution and threshold. Finally, if those values are calculated, motion vector of tip is acquired. It is recognized that just a movement of tip is a point and the open hand is a click. Recognition method is represented in a histogram of the motion vector movement.

4 Experiment

4.1 Image Grabber

An image grabber is made by BT878 of Rockwell. The BT878 video and audio capture chip has a multi-function PCI (peripheral component interconnect) device. The video function features DMA (direct memory access), PCI bus master for analog NTSC/PAL/SECAM composite, S-video and digital CCIR656 video capture [10]. The audio function features a completely independent DMA/PCI bus master for FM radio or TV sound capture. The main features of the BT878 are NTSC/PAL/SECAM video decoding, multiple YCbCr and RGB pixel formats supported on the output, VBI (vertical blanking interval).

The BT878 integrates an NTSC/PAL/SECAM composite and S-video decoder scaler, DMA controller and PCI Bus master on a single device. The BT878 can place the video data directly into a host memory for video capture applications and into a target video display frame buffer for video overlay applications. The BT878 contains a pixel data FIFO to decouple the high speed PCI bus from the continuous video data stream. The video data input may be scaled, color translated and burst-transferred to a target location on a field basis. This allows for simultaneous preview of one field and capture of the other field. Alternatively, the BT878 is able to capture both fields simultaneously or preview both fields

Table 1. Comparison of Computing Time (ms) and Searching Pixels

	Full Search	Three Steps	The modified BMA
Time	129	100	45
Pixels	2304	270	150

simultaneously. The fields may be interlaced into memory or sent to separate field buffers.

4.2 Hand Region Segmentation

The YCbCr color coordinate is an international standard for video signals and also BT878 chip generates color format of YCbCr (4:2:2). So processing time is reduced since there is no need to computation that converts RGB to YCbCr.

The values of Cb and Cr on the histogram are segmented for the whole image of the hand region. Here are values of Cb and Cr from the experiment to print out total 30 prints because of 10 prints for each 3 persons.

$$88 \leq C_b \leq 120 \quad \text{and} \quad 130 \leq C_r \leq 160 \quad (7)$$

In case of the BMA segmentation only with values of Cb and Cr, it could find wrong motion vectors due to lots of noises. The modified BMA is used to get motion vectors. The input image size is 320×240 pixels and the initial block size is 16×16 pixels. The search window size is 48×48 pixels. At the initial search window, 100 seeds were created and a 16×16 search window was created to be a center of the window. The final motion vector could be found with additional 50 seeds.

Table 1 shows that the modified BMA has excellent properties in computing time and searching pixels. So this period is enough to realize real-time working for the finger motion mouse on this paper. Figure 2 shows a hand region and its histogram of YCbCr and Figure 3 shows a hand region and its edge.

4.3 Gesture Recognition

Without real-time and on-line processing for the HCI system, working at finger gesture is completely vain. Therefore, any application is worthless if it takes longer computation time than existing mouse does regardless its algorithm performance.

A system that executes 30 frames was built for a reference. The result of motion vector is applied to an HCI system. The experimental result shows that hand region segmentation and motion vector acquisition with the YCbCr color coordinate and the modified BMA can be realized for the real-time and on-line processing.

(a) Hand Region

(b) Cb of Hand Region

(c) Cr of Hand Region

Fig. 2. Hand Region and its Histogram of YCbCr

It is made to be recognized as a pointing device when the thumb was unfold and moving, as clicking when all fingers were unfold. To avoid working to a slight movement, it was set up not to work for the motion vector less than 20. The HCI system recognizes the magnitude and direction difference of the motion vector between pointing and clicking.

(a) Original Image $F(t-1)$

(b) Original Image $F(t)$

(c) Edge of $F(t-1)$

(d) Edge of $F(t)$

Fig. 3. Hand Region and Its Edge

Figure 4 shows how the HCI system with the modified BMA works. When we point and move, the monitor shows its trace. When we click, it executes a click operation. While a new click clears all previous traces and goes back to initial display, the pointer is located on the center of the display.

5 Conclusion

A GUI has been a dominant platform for HCI. A GUI-based interaction has made computers simpler and easier to use. The GUI-based interaction, however, does not easily support the range of interaction necessary to meet users' needs that are natural, intuitive, and adaptive. In this paper, the modified BMA is proposed to track a hand in a sequence of an image and to recognize it in each video frame in order to replace a mouse with a pointing device for a virtual reality.

Fig. 4. HCI System

The HCI system with 30 frames per second is realized in this paper. The most important criteria for realizing the HCI system are the correct motion vector and its real-time processing. The modified BMA is proposed to estimate a position of the hand and segmentation with a orientation of motion and a color distribution of the hand region for real-time processing. The experimental result shows that the modified BMA with the YCbCr color coordinate guarantees the real-time processing and the recognition rate. The YCbCr color coordinate requires less bits to be coded than the RGB color coordinate with omitting the luminance of each pixel color and has less sensitivity to surrounding circumstances.

The hand tracking by the modified BMA can be applied to a virtual reality or a game or an HCI system for the disable. Our future work will focus on VHDL of the HCI in order to reduce some noises and to get the better response.

References

1. Mitchell, J.L., Pennebaker, W.B., Fogg, C.E., LeGall, D.J.: MPEG Video Compression Standard. Champion and Hall (1996)
2. Haskell, B.G., Puri, A., Netravali, A.N.: Digital Video: An Introduction to MPEG-2. Champion and Hall (1997)

3. Shi, Y.Q., Sun, H.: Image and Video Compression for Multimedia Engineering. CRC Press (2000)
4. Russ, J.C.: The Image Processing Handbook. 3 edn. CRC Press (1999)
5. Bae, C.S., Chun, B.T., Min, B.W.: Window manager control by extraction hand region and analysing its trajectory. *Journal of KISS* **24** (1997)
6. Ahmad, T., Taylor, C.J., Lantitis, A., Cootes, T.F.: Tracking and recognition hand gestures using statistical shape model. *Image and Vision Computation* **15** (1997)
7. Crane, R.: A Simplified Approach to Image Processing. Prentice Hall (1997)
8. Parker, J.R.: Algorithm for Image Processing and Computer Vision. Wiley Computer Publishing (1996)
9. Tekalp, A.M.: Digital Video Processing. Prentice Hall (1995)
10. Bovik, A.: Handbook of Image and Video Processing. Academic Press (2000)
11. Pratt, W.K.: Digital Image Processing. 3 edn. John Wiley (2001)
12. Sonka, M., Hlavac, V., Boyle, R.: Image Processing, Analysis, and Machine Vision. PWS Publishing (1999)
13. Gonzalez, R.C., Woods, R.E.: Digital Image Processing. 2 edn. Addison-Wesley (2002)
14. Jain, A.K.: Fundamentals of Digital Image Processing. Prentice Hall (1988)
15. Forsyth, D.A., Ponce, J.: Computer Vision: A Modern Approach. Prentice Hall (2002)
16. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification. 2 edn. Wiley-Interscience (2000)
17. Abe, S.: Pattern Classification. Springer Verlag (2001)
18. Javidi, B.: Image Recognition and Classification. Marcel Dekker (2002)