Journal of the Institute of Industrial Applications Engineers Vol.6, No.2, pp.73–79, (2018.4.25)
DOI: 10.12792/JIIAE.6.73 Online edition: ISSN 2187-8811 Print edition: ISSN 2188-1758

Paper

A Specification Method of Character String Region in Augmented Reality

TAKASHI HIGASA*† Non-member, SHIN-ICHI ITO*† Non-member
MOMOYO ITO*‡ Non-member, MINORU FUKUMI*‡ Member

(Received July 5, 2017, revised December 8, 2017)

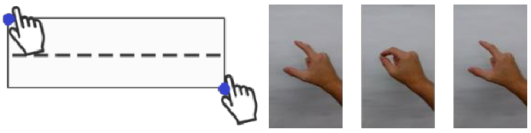
Abstract: This paper proposes a method to enter characters and/or character string in an augmented reality using a gesture motion. The proposed method detects the region of character string using the gesture motion. It consists of five phases; template generation, skin color detection, hand region detection, gesture motion extraction and designation of character string region. The template image consists of two fingers because a gesture is to take hold the tips of the first and second fingers. In the skin color detection, we extract the skin color on the basis of values in saturation by using threshold processing. The hand region is detected by calculating areas and detecting the area with the maximum value as a hand. The gesture motion is extracted using template matching. In order to show the effectiveness of the proposed method, we conduct experiments for character string specification.

Keywords: Augmented Reality, HSV Color System, Gesture Motion, Quadrangle for character string region

1. Introduction

Studies related to virtual reality, augmented reality (AR) and mixed reality have received a lot of attention in recent years. AR is a technology to expand reality environment by adding some information. AR technology is aimed at constructing a new user interface and making it useful for work support and information posting. Furthermore, AR service is expected to be applied to various fields such as medical, education, construction, tourism and entertainment. Although head mount displays (HMDs) for AR lack image quality and clarity compared with display monitors, they can add some information from a distant position to a wall surface, a floor surface and so on in real environment. Then, these interfaces can be utilized as a Graphical User Interface (GUI). These operations are intuitive. In AR contents, manipulation of virtual objects often uses hands. The manipulation performed by human hands is intuitive. OmniTouch is used by attaching a special equipment to the right shoulder [1]. It can be used not only on flat surfaces such as notes and desks, but also on slightly rounded things such as arms and thighs. However, there are problems; training for operation by a special input method, trouble to wear equipment, and securing a place for keyboard projection. On the other hand, the typing system in thin air has been developed to recognize the movements of fingers in the air and enable input operations without special equipment[2]. In this system, it is possible to operate the portable terminal in the air without wearing special equipment and the operation of one finger is supported. However, it is difficult to do intuitive input operation because the investigation is an indirect thing through the cursor. The tapping system using any fingers displays a keyboard on the AR and input characters [3]. However, it is not an intuitive input operation because there is no haptic feedback of the soft keyboard. In this way, users often use the soft keyboard in AR when searching some information that is key words, characters, words, character string, and so on. If characters and/or character string are retrieved by specifying the region of character string using a gesture motion, the issue of the soft keyboard is solved. The final goal of our studies is to create a character string retrieval system. In the proposed system, the region of character string is specified by a gesture motion. It is important to extract a gesture motion to construct the proposed system. There are lots of techniques to extract the gesture motion [4]. In general, the techniques for gesture motion extraction include skin color detection, hand detection, fingertip detection, noise reduction, gesture recognition, fingertip tracking, and so on. The images acquired from a camera is usually an RGB image. It is difficult to detect the gesture motion and track fingertip because the hand region differs depending on individuals and the RGB image is influenced by illumination change. The RGB color system is then converted to another color system and skin color regions are detected. Furthermore, there are lots of systems to detect gesture motions and fingertips. They use feature quantities of scale invariant feature transform (SIFT) and histograms of oriented gradients (HOG). However it is not suitable for gesture interface requiring real time property because feature quantities detection takes a long processing time. On the one hand, there is template matching to detect gesture motions and fingertips. Although it takes a long processing time to detect gesture

† Corresponding Author: {c501737043, s.ito}@tokushima-u.ac.jp
‡ Corresponding Author: {momoito, fukumi}@is.tokushima-u.ac.jp
* Tokushima University, Hakodate College
2-1, Minami-Josanjima, Tokushima 770-8506, Japan



(a) Region designation (b) Geature motion

Figure 1: Gesture motion to specify a region of character string.

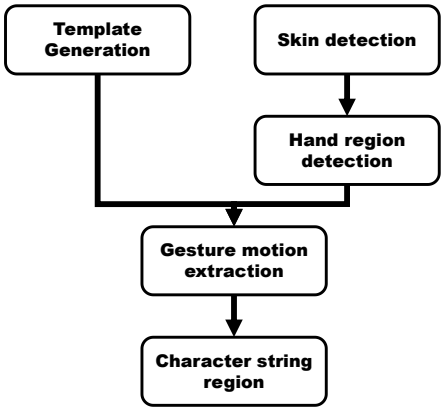


Figure 2: Flowchart of the proposed method.

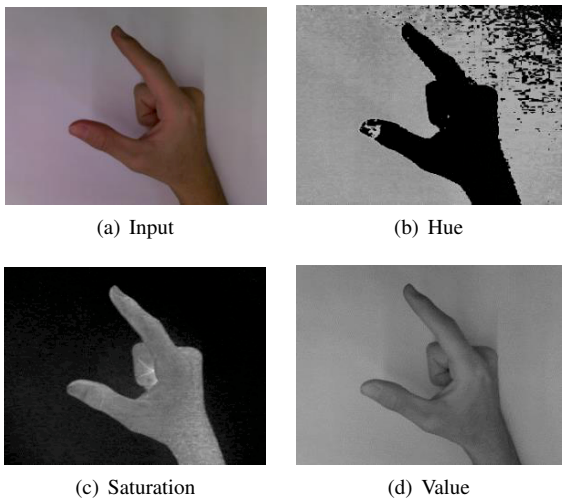
motions and fingertips, their accuracies become high. The cause of a long processing time is due to raster scan. The raster scan is to perform template matching on the entire image. Therefore, narrowing the scope of raster scan keeps real time property. We detect a candidate area of a hand to narrow the scope. The proposed method specifies a character string region using a gesture motion and consists of five phases; template generation, skin color detection, hand region detection, gesture motion extraction and designation of character string region.

In order to show the effectiveness of the proposed method, we conduct experiments for character string specification.

2. Proposed Methods

We propose a method to specify a region of character string by a gesture motion to perform a string search in AR. The proposed method sets up two points by the gesture motion. The gesture motion is to take hold the tips of the first and second fingers as shown in Fig. 1. Then, the region of a quadrangle that the two points are vertex of the quadrangle is specified. The proposed method detects the quadrangle region of character string using a gesture motion as shown in Fig. 1. It consists of five phases; template generation, skin color detection, hand region detection, gesture motion extraction and designation of character string region. The flowchart of the proposed method is shown in Fig. 2.

2.1 Template Generation The proposed method extracts a gesture motion using a template matching method. The template is created by taking a picture. Then, the first and second fingers are taken hold. The background color is white when taking a picture. The size of the template image is 130×60 pixels.



(a) Input (b) Hue

(c) Saturation (d) Value

Figure 3: HSV color system conversion.

2.2 Skin Color Detection The skin color is extracted to recognize hands. The skin color area is extracted using the HSV color system and a threshold processing, because the skin color is mixed color. The proposed method employs the HSV color system to detect the skin color. The HSV color system is composed of three components of hue, saturation and brightness. Fig. 3 shows a sample image of HSV color system conversion. In this paper, the skin color is extracted by using values in saturation. The threshold of the saturation for extracting the skin color is $S_{min} < S < S_{max}$. S_{min} and S_{max} are determined according to development environment.

2.3 Hand Region Detection The proposed method carries out contour detection to recognize hands (shown in Fig. 4). Firstly, the saturation image is converted into a binary image. The object is represented with white and the background is represented with black. Secondly, the binary image is scanned sequentially from the upper left. Then when finding a white pixel, the pixel is taken as the first outline pixel and the starting point to detect outline. Thirdly, the region contiguous to the starting point is searched to find outline pixels counterclockwise. Then, the first object area is set as the first outline pixel. Finally, the closed curve obtained is taken as the contour when it returns to the starting point. The extracted contours are stored in a hierarchical structure (shown in Fig. 5). (a)-(c) are an input image, a resulting example and a hierarchical structure, respectively. First, the outermost contour is extracted from Fig. 5(a). The extracted outermost contours are stored in level 1 of the hierarchical structure. Next, the outline in the outermost contour is extracted. If there is a contour, it is stored in level 2 of the hierarchical structure and this operation repeated. In this paper, we refer to the information of the hierarchy of level 1 because the outermost contour shows a hand. The each contour-enclosed superficial content in the hierarchy of level 1 is calculated. We regard the region of the maximum superficial content as a hand region.

2.4 2.4 Gesture Motion Extraction A gesture motion is detected by using a template matching method.

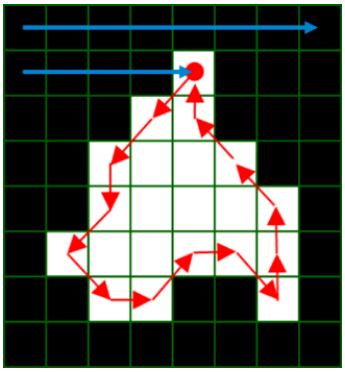
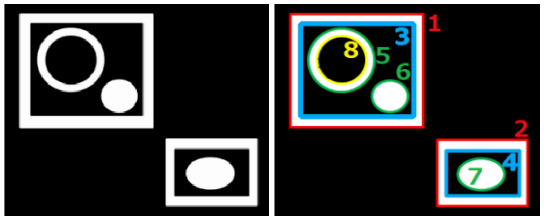
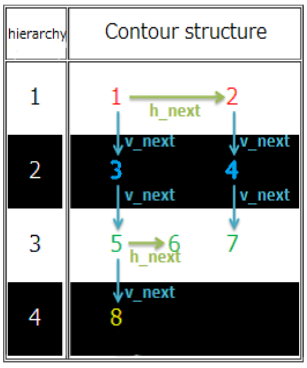


Figure 4: Contour detection. The red circle is a starting point.



(a) Input image (b) Contour detecting result



(c) Hierarchical structure in the contour extraction method

Figure 5: Contour detection algorithm. (a)-(c) are an input image, a resulting example of extracted the contours and ah hierarchical structure in the contour detection method extraction. The red circle is a starting point.

The proposed method has pre-processing for easy template matching. The pre-processing consists of two phases; blurring processing and processing of designating a region to be searched by template matching. The Gaussian filter is used to make a blurry image. Using the Gaussian filter, a filtering value of luminance values is calculated around the target pixel. In other words, we calculate the luminance value of the target pixel using a function of Gaussian distribution. The function of the Gaussian distribution is expressed by equation (1).

$$G(x,y) = \frac{1}{2\pi\sigma^2}\exp\{-\frac{x^2+y^2}{2\sigma^2}\} \quad (1)$$

where x and y are distances from the origin in the horizontal axis, and the vertical axis, respectively. σ is the standard deviation of the Gaussian distribution.

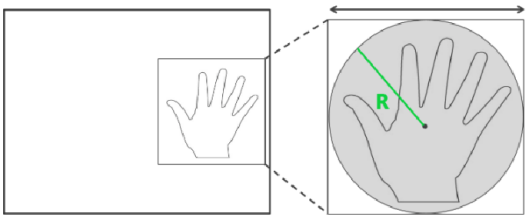


Figure 6: Hand region detection

The proposed method narrows a searching region for template matching. Then, the centroid of the detected hand region is calculated. Moreover, the radius R from the centroid is calculated (shown in Fig. 6.) as follows;

$$R = \alpha\sqrt{\frac{S}{\pi}} \quad (2)$$

where S and α are a hand region area and a bias to create a searching region, respectively. The proposed method makes the quadrangle surrounding the circle with radius R and it is set as a search region. (shown in Fig. 6).

The template matching is used to detect the gesture motion. In this paper, Normalized Cross-Correlation (NCC) is used for calculation of similarity. The function of the NCC is expressed by equation (3).

$$R_{NCC} = \frac{\sum_{j=0}^{N-1}\sum_{i=0}^{M-1} I(i,j)T(i.j)}{\sqrt{\sum_{j=0}^{N-1}\sum_{i=0}^{M-1} I(i,j)^2 \times \sum_{j=0}^{N-1}\sum_{i=0}^{M-1} T(i,j)^2}} \quad (3)$$

where $T(i,j)$ and $I(i,j)$ are the luminance values of a template image and a frame image, respectively. M and N are the width and the height of the frame image, respectively. A gesture motion is extracted when the similarity is larger than a certain value.

2.5 2.5 Designation of character string region

The start and end points to specify a character string region are determined using a gesture motion. The algorithm for determining the start and the end points is shown in Fig. 7. We visualize the start and the end points determined in the user interface. The start and the end points are displayed in the red and the green circles, respectively (Fig. 8). The character string region is the quadrangle surrounded by the start and the end points (shown in Fig. 8). In Fig. 8, the white quadrangle shows the specified region.

3. Experiments of Threshold Decision for Template Matching

The threshold for template matching is decided through experiments because the threshold of the template matching is significantly related to gesture motion extraction. We conducted experiments to investigate the threshold on R_{NCC}. At first, the threshold was changed from 0.91 to 1.00 at 0.01 intervals to detect a candidate threshold. Then, the threshold was changed with a focus on the candidate threshold at 0.001 intervals to decide the threshold more accurately.

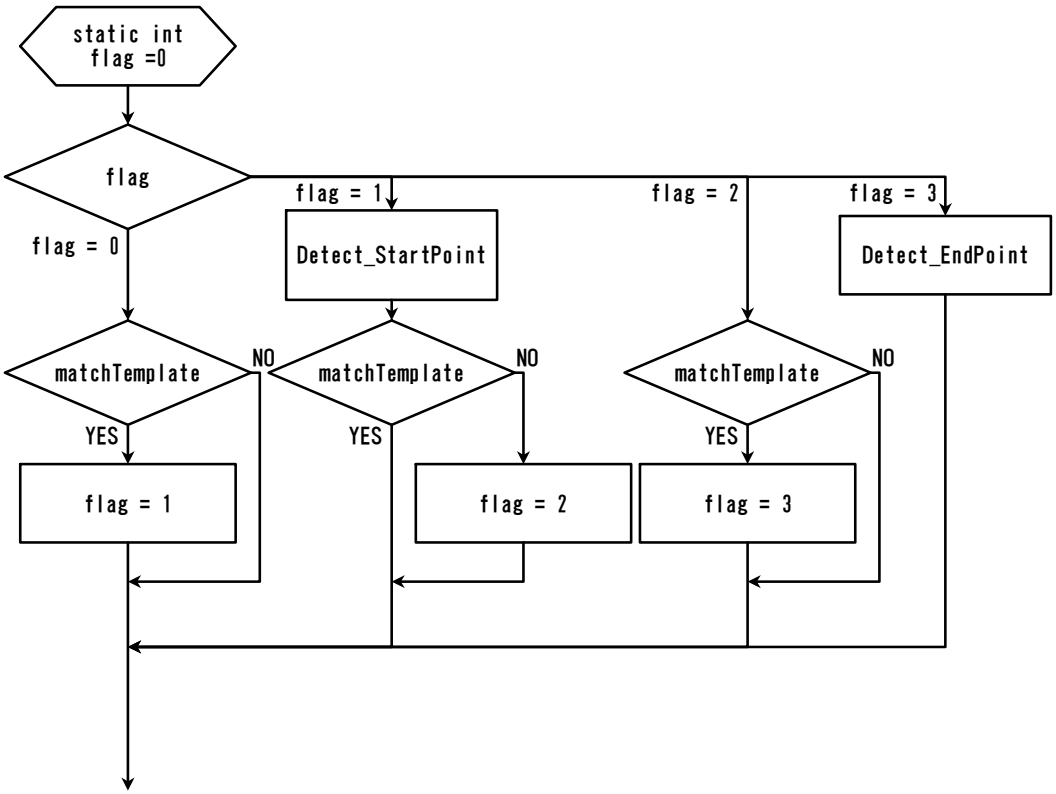


Figure 7: Procedure of determination of start and end points.

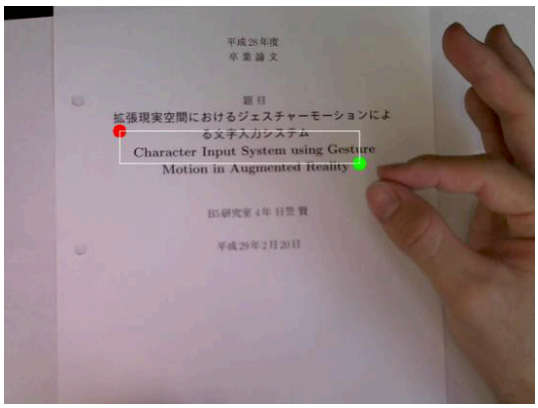


Figure 8: Determination of a start and an end points. The red and the green circles are the start and the end points, respectively. The white quadrangle shows the specified region.

3.1 Experiment a Environment The subject sat on a chair in our laboratory. The number of experiment times is 10. The size of words of a journal title was 300×50 pixels. S_{min} and S_{max} for skin color detection were 200 and 255, respectively. α for hand region detection was 1.6. σ for the Gaussian filter was 11. The number of success times of gesture motion extraction was calculated to detect the candidate threshold and decide the threshold.

3.2 Experimental results and discussions Table 1 shows the results for the candidate threshold detection and threshold decision. In table 1(a), "$\times$" represents failure. The gesture motions could not be extracted when R_{NCC} were 0.91 to 0.97. The cause of the failure is that gesture motion was always extracted even if the subject was not doing the gesture motion. These results suggest that the threshold of R_{NCC} was small. Then, the numbers of success times were 10, 8 and 0 when the threshold of R_NCC were 0.98, 0.99 and 1.00, respectively. The numbers of failure times were 2 and 10 when the threshold of R_{NCC} were 0.99 and 1.00. At the time of this failure, the gesture motion was not detected. The thresholds were regarded as severe when the threshold of R_{NCC} were 0.99 and 1.00 respectively. We regarded 0.98 as the candidate threshold of R_{NCC} because the highest number of success times was given at 0.98. Then, table 1(b) shows the results of the threshold decision by changing with a focus on the candidate threshold at 0.0001 intervals. The number of success times was 10 when R_NCC are 0.981 to 0.984. The number of success times was more than 10 times when R_NCC were 0.985 to 0.989. These results suggest that R_{NCC} can be set to an adequate threshold value. The reason is that one gesture motion was extracted as two gesture motion. These result suggest that the threshold was severe set. Since the threshold values 0.984 and 0.985 were borderline, it was considered unsuitable. Therefore, this paper regarded the value 0.983 as an appropriate threshold.

4. Experiments of Character String Specification

In order to show the effectiveness of the proposed method, we conducted experiments to specify the character string.

Table 1: Experimental results for the candidate threshold detection and the threshold decision.

(a)

threshold	0.91	0.92	0.93	0.94	0.95	0.96	0.97	0.98	0.99	1.00
times	×	×	×	×	×	×	×	10	8	0

(b)

threshold	0.981	0.982	0.983	0.984	0.985	0.986	0.987	0.988	0.989
times	10	10	10	10	12	12	14	12	11

4.1 Experiment Environment The subjects were 4 healthy people (average age : 22 years old). The subject sat on a chair and conducted a task for experiments in our laboratory. The task was to select the region of a printed journal title by a gesture motion. The number of experiments per him/her was five. The number of iterations per experiment was free. The subjects repeatedly performed to select the region until a successful conclusion was obtained. The size of words of the journal title was 300×50 pixels. The line spacing was about 15 pixels. S_min andS_max for skin color detection were 200 and 255, respectively. α for hand region detection was 1.6. σ for the Gaussian filter was 11. Moreover, the threshold of template matching (R_{NCC}) was 0.983.

4.2 Evaluation method This paper employed operation accuracy for specifying the region to evaluate the proposed method. The evaluation methods made a comparison between the results of regions specified using mouse operation and the proposed method. In operation evaluation, we regarded it as success when the coordinate differences between the mouse operation and the gesture operation on both x-coordinate and y-coordinate became 35 pixels or less.

$$\begin{aligned} |M_{Sx} - P_{Sx}| &\leq 35 \\ |M_{Sy} - P_{Sy}| &\leq 35 \\ |M_{Ex} - P_{Ex}| &\leq 35 \\ |M_{Ey} - P_{Ey}| &\leq 35 \end{aligned} \tag{4}$$

where M and P show the mouse operation and the proposed method, respectively. S, E, x and y are the start position, the end position, x-coordinate and y-coordinate, respectively.

4.3 Experimental Results Figure 9 shows examples of results of the mouse operation and the gesture operation for specifying the region of the journal title. The red and the green circles and the white quadrangle are the same as those in Fig. 8. The yellow quadrangles are the results obtained by the mouse operation. Table 2 shows experimental results of the operation evaluation. Table 2(a)-(d) show the results of the subjects A, B, C and D, respectively. "S or F" represents success or failure. The number of the success times for specifying the region was 17 (subjects A, B and D : 4 times, subject C : 5 times). The number of the failed times was three (subjects A, B and D : one time). In the subject B, the coordinate differences of the end point became smaller by repeated experiments. In the subject D, the coordinate differences ($M_{Ex} - P_{Ex}$, $M_{Ey} - P_{Ey}$) was small compared to those of other subjects.

Table 2: Results of the operation precision evaluation

(a) Subject A

S or F	$M_{Sx} - P_{Sx}$	$M_{Sy} - P_{Sy}$	$M_{Ex} - P_{Ex}$	$M_{Ey} - P_{Ey}$
S	-18	-7	-18	-35
F				
S	-29	-29	-3	-8
S	-4	-15	-5	-10
S	-1	-23	-7	-12

(b) Subject B

S or F	$M_{Sx} - P_{Sx}$	$M_{Sy} - P_{Sy}$	$M_{Ex} - P_{Ex}$	$M_{Ey} - P_{Ey}$
F				
S	-17	-13	-32	-31
S	-10	-19	-19	-30
S	-21	-20	-21	-10
S	-14	-33	-10	-2

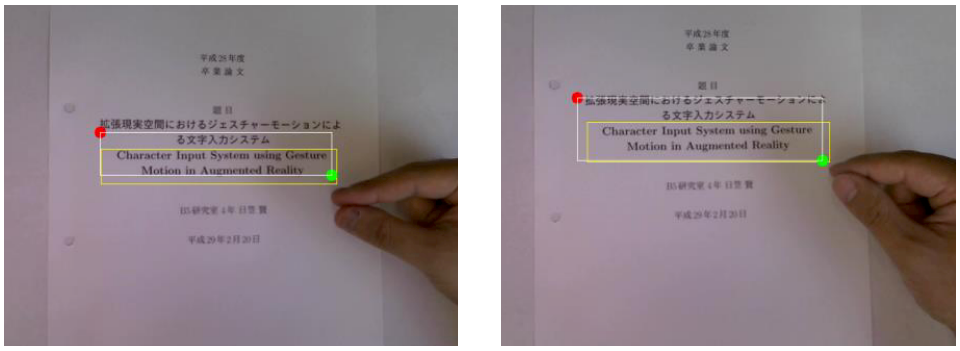
(c) Subject C

S or F	$M_{Sx} - P_{Sx}$	$M_{Sy} - P_{Sy}$	$M_{Ex} - P_{Ex}$	$M_{Ey} - P_{Ey}$
S	-9	-15	-7	-27
S	3	-12	-24	-19
S	3	-9	-22	-10
S	-9	-19	-23	-7
S	-12	-3	-33	12

(d) Subject D

S or F	$M_{Sx} - P_{Sx}$	$M_{Sy} - P_{Sy}$	$M_{Ex} - P_{Ex}$	$M_{Ey} - P_{Ey}$
S	1	1	-12	-1
S	-3	2	-8	0
S	-2	-2	-15	-5
S	5	2	3	8
F				

4.4 Discussions We confirmed that the number of the success times was 17. In the proposed method, the skin color was detected by a threshold processing on the saturation, hand region was detected using contour detection algorithm and the gesture motion was extracted by template matching. These results suggest that the accuracy of each approach is high, and the parameters for each approach are set appropriately. Furthermore, It is observed that the devi-



(a) Subject A (b) Subject B

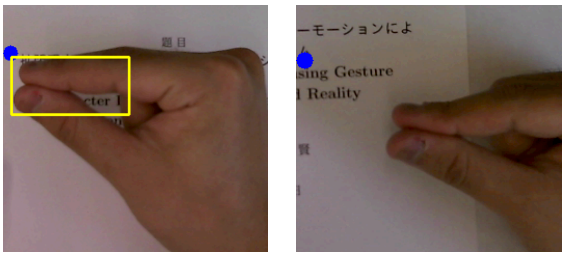
(c) Subject C (d) Subject D

Figure 9: Result of the operation for specifying the region of the journal title. The Red and the green circles and the white quadrangles are the same as those in Fig. 8. The Yellow quadrangle are the results by the mouse operation.

ation of the designated points is negative for the points by the mouse operation. These results suggest that the designated points are on the upper left from the mouse operation point. This is due to the fact that the specified point was at the upper left corner of the template image. We confirmed that the number of the failed times are three. These results suggest that the hand region was not successfully detected. The hand region is judged to be the maximum area of the skin color region. When there is a hand region at the edge of the screen, the region is small and it is not judged well as a hand area. Furthermore, it was not easy to do the gesture motion, and therefore the start and end points were not stable. Fig. 10 shows an undetection example of the subject A. The cause of this was because a part of the hand region became dark. In the results of the subject B, the coordinate differences became smaller by repeated experiments. These results suggest that the subject B adjusted gradually the gesture motion. Fig. 11 shows an undetection example of the subject B. It is considered this result was caused by the absence of the finger area. In the results of the subject D, the coordinate differences were small because the subject D corrected positioning of the start and the end points in a special way. That special way was to continue to close the two fingers and move the designated point.

5. Conclusion

In this paper, we proposed a method to specify a region of character string by a gesture motion. The proposed method consists of five phases; template generation, skin color detection, hand region detection, gesture motion extraction and designation of character string region. In the template generation, we created an image when two fingers touch each other as a template image. The skin color was then detected by a threshold processing on the saturation in the HSV color system. Next, the hand region was detected using a contour detection algorithm. After hand region detection, the gesture motion was extracted by using template matching. In order to show the effectiveness the proposed method, we conducted experiments for character



(a) Start point (b) End point

Figure 10: Subject A failing result. (a) was a range frame to perform template matching when the subject A specified the start point. The yellow square indicates the positions matched with the template. (b) was a range frame to perform template matching when the subject A tried to designate an end point.

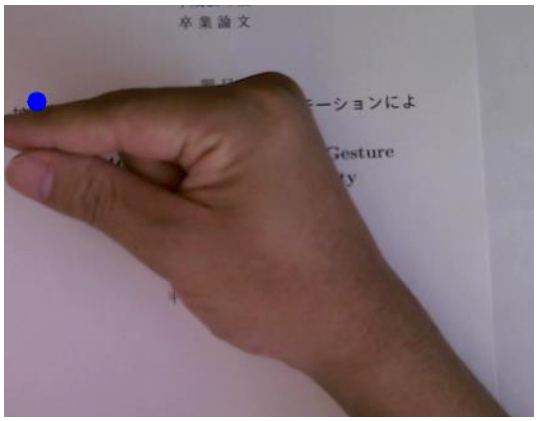


Figure 11: Range frame to perform template matching when the subject B failed

string specification. In the experimental results, we could obtain good results (the number of the success times was 17). However, we had three failed times. It is thought that these results are due to the fact where the region of fingers went out of the screen because the character of the journal title was the edge of the screen. Furthermore, it was not easy to do the gesture motion, and therefore the start and the end points were not stable. The future work will improve the way of doing a gesture motion because it was not easy to do the gesture motion in this paper.

References

[1] C. Harrison, H. Benko, A. D. Wilson, "Omnitouch: Wearable multitouch interaction everywhere", JProc. the 24th annual ACM symposium on User interface software and technology, ACM, pp.441-450, 2011.

[2] 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 on Human factors in computing systems, ACM, pp.712-713, 2003.

[3] A. Sugiura, M. Toyoura, M. Xiaoyang and S. Nisiguchi, "Kakutyou Genzitu Kan no tame no click inter-face (Intuitive click interface for augmented reality)", Transactions of the Institute of Electronics, Information and Communication Engineers, D 97.9, pp.1426-1436, 2014.

[4] C. Manresa, *et al*, "Hand tracking and gesture recognition for human-conputer interaction", ELCVIA Electronic Letters on Computer Vision and Image Analysis, pp.96-104, 2005.

[5] S. Lenman, L. Bretzner, B. Thuresson, "Using marking menus to develop command sets for computer vision ased hand gesture interfaces, NordiCHI ' 02", Proceedings of the second Nordic conference on Human-computer interaction, pp.239-242, 2002.

[6] C. Manresa, J. Varona, R. Mas and F. J. Perales, "Hand Tracking and Gesture Recognition for Human-Computer Interaction", Electronic Letters on Computer Vision and Image Analysis 5(3), pp.96-104, 2005.

[7] V. I. Pavlovic, R. Sharma and T. S. Huang, "Visual interpretation of hand gesture for human-computer interaction: a review", IEEE Pattern Analysis and Machine Intelligence 19(7), pp.677-695, 1997.

Takashi Higasa (Non-member) received the B. E. degrees from the Tokushima University in 2017. He has studied Information Science and Intelligent System of master ' s program at Tokushima University. He received the best student paper award from the ICISIP in 2017. His research interests include image processing and augmented reality.

Shin-ichi Ito (Non-member) has received the B. E. and M. E. degrees from the Tokushima University in 2002 and 2004, respectively, and the D. E. degree from Tokyo University of Agriculture and Technology in 2007. He has worked at Japan Gain the Summit Co., Ltd. and Tokyo University of Agriculture and Technology as a System Engineer and an Specially Appointed Assistant Professor, in 2004 and 2007, respectively. Since 2009, he has been an Assistant Professor at Tokushima University. His current research interests are EEG analysis, bio- signal processing and information visualization. He is a member of IEEE, IEICE, JSMBE and IEEJ.

Momoyo Ito (Non-member) received B.E., M.E., and Ph.D. degrees in computer science from Akita University, Japan in 2005, 2007, and 2010, respectively. She was an Assistant Professor from 2010 to 2016 at Tokushima University. Since 2016, she has been an Associate Professor at Tokushima University. Her current research interests are human behavior analysis and intelligent transportation systems for active safety. She is a member of IEEE, JSAE, IEICE, IPSJ, and JSME.

Minoru Fukumi (Non-member) received the B.E. and M.E. degrees from the Tokushima University, in 1984 and 1987, respectively, and the doctor degree from Kyoto University in 1996. Since 1987, he has been with the Department of Information Science and Intelligent Systems, Tokushima University. In 2005, he became a Professor in the same department. He received the best paper awards from the SICE in 1995 and Research Institute of Signal Processing in 2011 in Japan, and best paper awards from some international conferences. His research interests include neural networks, evolutionary algorithms, image processing and human sensing. He is a member of the IEEE, IEEJ, RISP, JSAI and IEICE.