Hand Gesture Recognition Based on Optimal Segmentation in Human-Computer Interaction

Md Abdur Rahim[1], Abu Saleh Musa Miah[2], Abu Sayeed[3], Jungpil Shin[4*]

[1, 3]School of Computer Science and Engineering, University of Aizu, Aizuwakamatsu, Fukushima, Japan
[2]Department of Computer Science and Engineering, Bangladesh Army University of Science and Technology, Bangladesh
[3]Department of Computer Science and Engineering, Rajshahi University of Engineering and Technology, Bangladesh
[*]Corresponding Author: Email: jpshin@u-aizu.ac.jp

2020 3rd IEEE International Conference on Knowledge Innovation and Invention (ICKII) | 978-1-7281-9333-5/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICKII50300.2020.9318870

Abstract

In recent years, hand gesture recognition (HGR) systems have been extensively developed with the technologies of human-computer interaction (HCI), which enables regular interaction with machines. However, the significance of progress in HGR continues to advance, although the problem of hand segmentation and recognition is challenging due to the unfavorable environment, background illumination, hand size, and shape. To overcome this, we propose an optimal segmentation method for identifying hand gestures from input images, improving recognition performance. For segmenting hand gestures, we compared the segmentation methods of YCbCr, SkinMask, and HSV (hue, saturation, and value). The CR component is extracted from YCbCr, then binarization, erosion, and hole filling are performed. Color segmentation is applied to the SkinMask process that detects pixels that match the color of the hand. In the HSV process, threshold masking determines the dominant features. The Softmax classification is used to classify hand gestures where features are extracted through convolutional neural network (CNN). The proposed segmentation methods are applied to a benchmark dataset and the result shows an improvement in recognition accuracy over state-of-the-art systems.

Keywords: human-computer interaction, segmentation, convolutional neural network, hand gesture

Introduction

Hand gesture recognition has an important significance in the efficiency of human-computer interaction technology, which is used as a favorable interface in various adverse situations. It introduces a new dimension to HCI, which can replace touch system functions for a healthier and safer environment for communication with people and devices. Hand gestures are recognized as an important medium in many HCI applications, such as virtual reality, sign language recognition, aerial handwriting, and robotic instruction. However, the problem of identifying and segmenting the hand is challenging because of the complex background, different illumination, size, and shape of the hand. Many scholars explored a dynamic approach that uses gestures to increase the usability of advanced technology. 3D sensors are used to obtain feature data to detect hand gestures [1]. However, it requires a lot of time to analyze the total number of features. Thus, the comparative analysis of hand gestures with the hidden Markov model was proposed to compare with other classified methods [2]. The advantages and disadvantages of various classification technologies are highlighted in Ref. [2]. A system that could be used to identify the word sign was proposed with hybrid segmentation for detecting hands, acquisition of features using CNN feature fusion, and classified by SVM [3]. Reference [4] proposed a system that could detect and recognize hand postures in complex backgrounds. However, the previous methods had low recognition accuracy due to processing speed and computational complexity. Thus, hand gesture recognition based on skeletal data was proposed in [5]. However, there are problems with unwanted tremor, the variation of bone size, self-occlusion skeletal information. Recognizing dynamic hand gestures, the classification based on the hidden conditional neural field (HCNF) was suggested in the human-computer interaction [6]. However, the difficulty of gesture recognition is the changes for shape, direction, lighting condition as it requires real-time performance accuracy and compatibility. Therefore, in this study, we propose an optimal hand segmentation method that recognizes the gestures or instructions among the most demanding environments and applications, which reduce the gap between the human and computer, machine, or robot that can accomplish a specific task.

Proposed Methodology

The overall architecture of the proposed system as mentioned is shown in Fig. 1. The architecture comprises multiple segmentation methods, including CNN and softmax classifier for the purpose of feature extraction and classification. Therefore, we describe the proposed methodology including (A) preprocessing of different segmentation methods, (B) feature extraction, and classification.

A. Preprocessing Method

To process the input images, the Gaussian blurring approach is applied to smooth the input image and reduce noise. A set of segmentation methods is compared with each other for hand gesture detection and recognition.

ISBN: 978-1-7281-9333-5

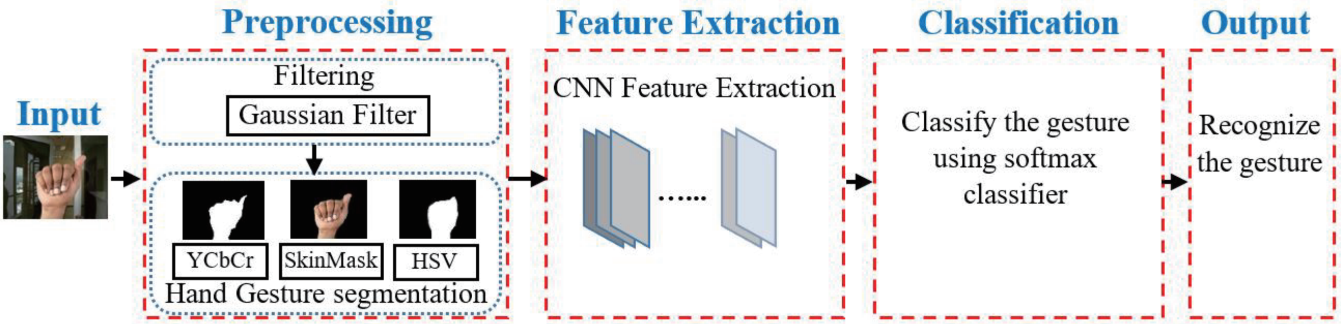

Fig. 1 The architecture of the hand gesture recognition system.

Segmentation of YCbCr Color Space

We performed segmentation in the YCbCr color space. It represents the color value of a luminance element (Y) and two chrominance elements Cb (chrominance-blue) and Cr (chrominance-red). The algorithm calculates the RGB value of each pixel by considering the pixels in the image as one-dimensional arrays. RGB values are then converted to YCbCr color space. However, the grayscale image has two different colors, black and white which are defined by 0 and 255. In this experiment, the threshold value was defined as 128. Moreover, pixel values were defined as 0-127 to 0 and 128-255 to 255. Therefore, we applied erosion to the binary image to remove the pixels in the foreground. The size of the foreground shrinks, and the hole in the region becomes larger. Then, we filled those holes and took the filled image as segmented images. Figure 2 shows the steps of the YCbCr segmentation of an image.

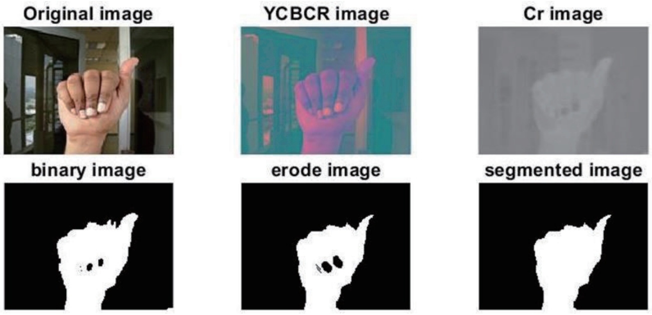

Fig. 2 YCbCr segmentation process of an input image.

Segmentation of SkinMask

To detect a hand, the system detects pixels in a specific frame that matches the color of the hand and performs color segmentation. After capturing gesture images, we accomplished post-processing on images to highlight contours and edges using a binary threshold, blurring, and gray scaling. We performed skin mask mode operation for image segmentation. In skin mask mode, the input images are converted to HSV. However, the range of H, S, V values is measured based on the range of skin color. Then, we applied erosion followed by the dilation. Therefore, Gaussian blur masks everything except the skin-colored things, smoothening the noises. Using this output as a mask on the original input, we applied grayscale on it. This mode is applied under challenging conditions with variable light and multiple subjects. Figure 3 shows the processing image of the skin mask mode.

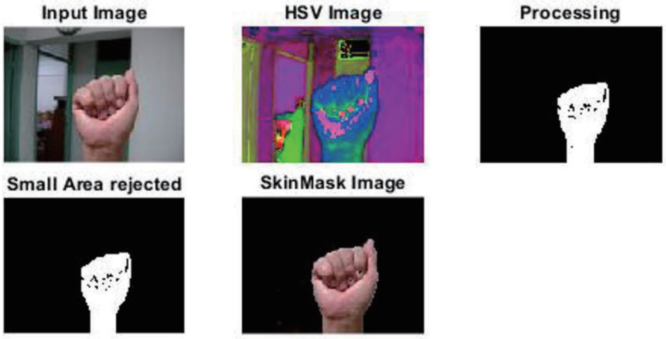

Fig. 3 SkinMask segmentation process of an input image

Segmentation of Hue, Saturation, and Value (HSV) Color Space

We analyzed the spatial properties of HSV colors by emphasizing the conceptual concept of color combination, the values of significance, and intensity of an image pixel. Hue (H), saturation (S), and values (V) in HSV determine the color, depth or purity of the hand, and the light of the image. Hue defines an angle with a continuous representation of color over a range of red axes, and a radial distance measured for saturation. When the saturation is low, the color intensity can be estimated by the gray values while for high saturation, the color intensity can be approximated by its hue. We set threshold masking to determine the dominant features. To define high and low threshold masks when hue, saturation, and values pixels are set to 1, and the remaining are 0. Depending on the training dataset, threshold values are fixed based on various experiments. The low and high saturation limits are $S_{low} = 0.01$ and $S_{high} = 1$, respectively, while the low and high values are $V_{low}=0$ and $V_{high} = 1$. The lower and higher components of hue are $h_{low} = 0$ and $h_{high} = 1$, respectively in case of no threshold. If Eq. (1) is satisfied, the color of the hand is identified.

$$P_{gesture}(l,k) = \begin{cases} 1, if\ a < H(l,k) < b, c < S(l,k) < d, e < V(l,k) < f \\ 0, \quad otherwise \end{cases} \qquad (1)$$

where H (l, k), S (l, k), and V (l, k) are hue, saturation, and value element, and the h_{low}, h_{high}, s_{low}, s_{high}, v_{low}, and v_{high} are expressed as a, b, c, d, e, f, respectively. Figure 4 shows an example of HSV color analysis of a hand gesture. The segmented image is provided to determine the features.

ISBN: 978-1-7281-9333-5

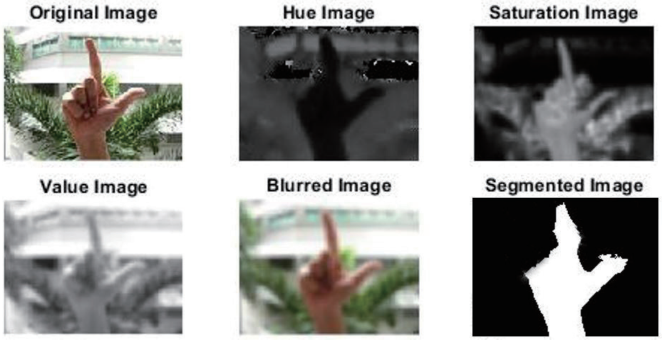

Fig. 4 HSV segmentation analysis.

B. Feature Extraction and Classification

A convolutional neural network (CNN) is a deep learning method that takes an input image, determines the importance of different aspects of the image and is able to distinguish it from each other [7]. It has a sequence of layers, and each level of the convolution transforms one volume of activation into another form with a different function. Figure 5 illustrates the structure of the CNN feature extraction and classification process. To extract features, we used the processed images as input. The level of the convolution was used to extract the high-level properties and move them to a specific length of kernel step. Furthermore, we used a pooling layer like Max Pooling to find the dominant features for effective training of the model. It reduces the spatial size of the convoluted feature and takes measures the maximum value covered by the kernel. The level of convolution is used to learn the non-linear combinations in a fully connected layer. Then, the input image is converted using Flatten function into a feature vector and applied in the training for iterations. However, the proposed architecture differentiates between expected and low-level features in images and classify them using the softmax classifier.

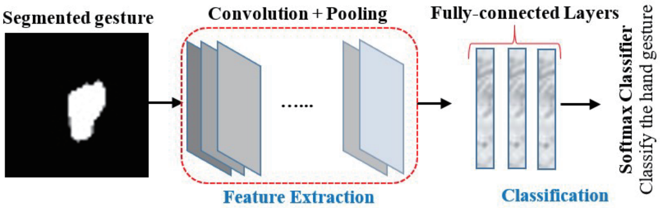

Fig. 5 Structure of the feature extraction and classification.

Experimental Results

A. Dataset Description

In this work, we considered the benchmark dataset [4]. The dataset contained ten hand postures of different hand shapes and sizes on different backgrounds. Forty subjects (aged from 22 to 56 years) participated in performing hand postures and 5 images were collected from each subject. A total of 2000 hand postures (40 people in 10 classes with 5 images per class) were used for gesture segmentation and recognition. The hand postures images were 160 x 120 pixels in size with complex backgrounds. Figure 6 shows the sample images from the NUS-II dataset.

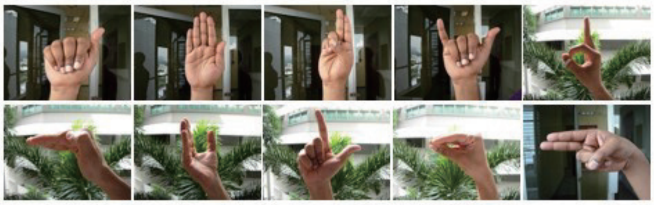

Fig. 6 Representative images from NUS-II (subset A) [4], showing gesture classes a to j.

B. Hand Gesture Segmentation and Recognition

We segmented the hand from dataset images through the proposed multiple segmentation methods. The segmented images were then used as input to the CNN architecture. YCbCr, SkinMask, and HSV segmented images are shown in Fig. 7. The proposed CNN architecture extracted features and classified hand gestures using the softmax classifier. We separated the data into two random groups such as training and testing datasets. The 80% of dataset images were used as training and the rest were testing. In this experiment, 400 images were used to evaluate the classification performance. We tested the accuracy of recognition by both unsegmented and segmented hand posture images. Table 1 represents the comparison of the recognition accuracy of unsegmented hand gestures and segmented hand gestures. Figure 8 shows the recognition accuracy based on the different segmentation approaches.

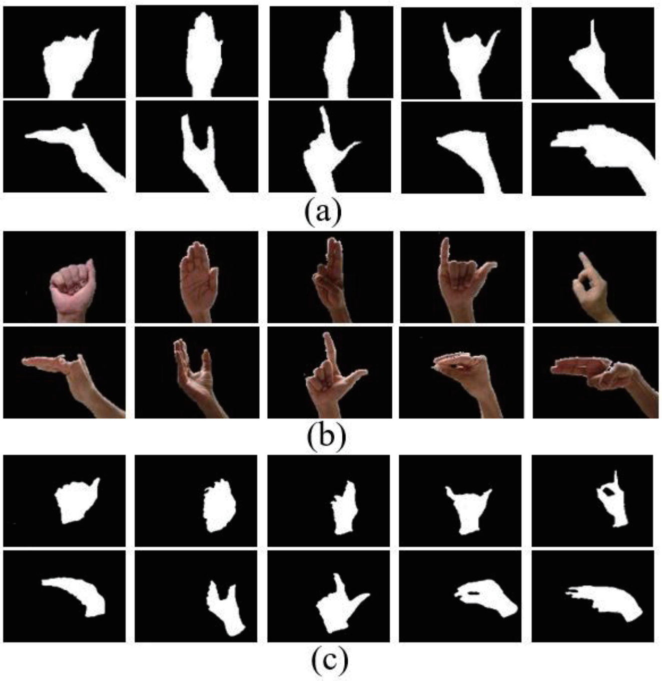

Fig. 7 Example of segmented images from datasets. (a) YCbCr segmentation, (b) SkinMask segmentation, and (c) HSV segmentation.

According to the experimental results, the HSV segmentation method achieved better results than the other two methods owing to the use of different thresholds for different regions of the same image. Therefore, the gestures of the hand were well-segmented. The system achieved the highest recognition accuracy using the HSV segmentation method in “b”. The average accuracy of a different method of segmented gestures was 95.27%, 96.33%, and 97.43%, respectively.

However, the recognition accuracy was increased compared to unsegmented gestures. Table 2 represents the comparison of recognition accuracy.

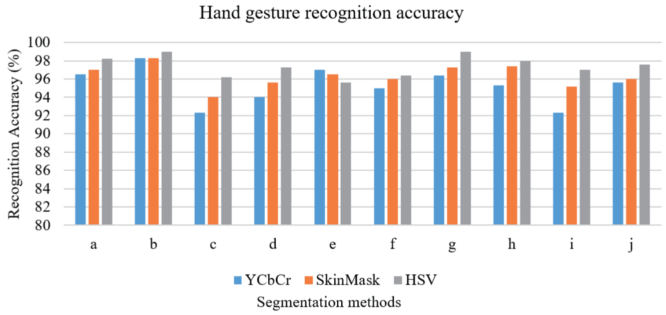

Fig. 8 Accuracy of recognizing hand gestures using different segmentation methods.

TABLE I
ACCURACY OF UNSEGMENTED AND SEGMENTED HAND GESTURES

Gestures	Recognition accuracy (%)			
	Unsegmented	Segmentation method		
		YCbCr	SkinMask	HSV
a	90.7	96.5	97	98.2
b	96	98.3	98.3	99
c	85.7	92.3	94	96.2
d	89.7	94	95.6	97.3
e	93.2	97	96.5	95.6
f	91.2	95	96	96.4
g	86.5	96.4	97.3	99
h	90	95.3	97.4	98
I	83.5	92.3	95.2	97
J	93	95.6	96	97.6
Avg.	89.95	95.27	96.33	97.43

TABLE II
COMPARISON OF RECOGNITION ACCURACY WITH PREVIOUS SYSTEMS

Reference	Method	Reported accuracy (%)
Ref. [4]	Bayesian model	94.36
Ref. [5]	Skeleton Distance measurement	95.91
Ref. [8]	CNN	90
Proposed	Segmentation and CNN	97.43

Conclusion

This paper highlights the significant contribution of hand gesture recognition based on the proposed segmentation methods. We compared several methods for segmentation and classification using the benchmark image dataset. To recognize the hand gestures, the preprocessing of input images was implemented using various segmentation methods such as YCbCr, SkinMask, HSV. The preprocessed images were then provided for the extraction of features. We proposed an efficient feature extraction strategy, CNN that combines hand motion and cues for gesture recognition. The gestures are categorized using softmax classification. The experimental results achieved the accuracy of the average recognition of 95.27%, 96.33%, 97.43% using YCbCr, SkinMask, HSV proposed segmentation methods. Moreover, we compared the recognition of segmented and unsegmented gestures. The results showed that the HSV segmentation method was better to recognize the gesture because it distinguished colors from intensity as robustness to the lighting invariant.

References

[1] J. Jia, et al., *Real-time hand gestures system based on leap motion*, Concurrency and Computation: Practice and Experience, vol. 31, no. 10, pp. e4898, May 2019.

[2] K.M. Sagayam and D.J. Hemanth, *Hand posture and gesture recognition methods for virtual reality applications: a survey*, Virtual Reality, vol. 21, no. 2, pp. 91-107, Jun 2017.

[3] M.A. Rahim, et al., *Non-Touch Sign Word Recognition Based on Dynamic Hand Gesture Using Hybrid Segmentation and CNN Feature Fusion*, Applied Sciences, vol. 9, no. 18, pp. 3790, Sep 2019.

[4] P.K. Pisharady, et al., *Attention based detection and recognition of hand postures against complex backgrounds*, International Journal of Computer Vision, vol. 101, no. 3, pp. 403-419, Feb 2013.

[5] M.A. Rahim, et al., *Human-Machine Interaction based on Hand Gesture Recognition using Skeleton Information of Kinect Sensor*, In Proceedings of the 3rd International Conference on Applications in Information Technology, ACM, pp. 75-79, Nov 2018.

[6] W. Lu, et al., *Dynamic hand gesture recognition with leap motion controller*, IEEE Signal Processing Letters, vol. 23, no. 9, pp. 1188-1192, Jul 2016.

[7] S.F. Chevtchenko, et al., *A convolutional neural net-work with feature fusion for real-time hand posture recognition*, Applied Soft Computing, vol. 73, pp. 748-766, Dec 2018.

[8] S. Ahlawat, et al., *Hand Gesture Recognition Using Convolutional Neural Network*, In International Conference on Innovative Computing and Communications, Springer, pp. 179-186, 2019.