FILTERING AND REDUCING DIMENSION IN THE RECOGNITION OF VIETNAMESE STATIC SIGN LANGUAGE

Tran Thi Minh Hanh[*], Pham Xuan Trung[**], Ho Phuoc Tien[***]

The University of Danang, University of Science and Technology

[*]t2mhanh@gmail.com, [**]trung.phamxuan@gmail.com, [***]hptien@yahoo.com

Abstract - Sign language is the primary language for the deaf in communicating with normal people. In this paper, Vietnamese Static Sign Language (VSL) using filtering and dimension - reducing methods combined with Neural network has been proposed. This approach begins with pre-processing steps, feature extraction and classification to recognize and to show the results of the relevant letter. In this paper, the hand's region is segmented from the background using skin color; after that, filters are used to reduce noise/ to smooth noise before extracting the features using the down-sampling method. These features are then used to train and to test with the back-propagation Neural Network. The implementation is performed on the database that was built with some conditions. The combination of the above-mentioned algorithms based on self-built databases results in a relatively high outcome.

Key words - recognition; hand gestures; skin colour; Neural network; PCA; sign language.

1. Introduction

Deaf people account for a rather high proportion. According to statisticst in 2009, we have 360 million deaf people. In Vietnam, the deaf made up one million in 2009. Most of the deaf are also unable to speak; hence it is difficult for them to communicate with other normal persons. Sign language had been developed in the hearing-impaired community for a long time to help them to communication with each other and also with the normal. However only a small proportion of normal people can understand this communication means. Therefore, the development of the automatic sign language translation to natural language is highly expected to improve the communication means among humans.

For many recent years, in the world, there have been many innovative methods to solve this problem. A real-time HMM-based system has been designed for recognizing sentence-level American Sign Language [1], [2]. The subject wears distinctly coloured gloves on both hands, and sits in a chair in front of the camera. Omer Rashid *et al.* [3] proposed to use Support Vector Machine (SVM) combined with two moments based approaches namely Hu-moment along with geometrical description of finger and Zenike moment. Feature extraction is invariant to translation, rotation and scaling with the accuracy rate of 98.5% using Hu-Moment with geometrical features and 96.2% recognition rate using Zernike moment for ASL alphabets and numbers. Chung Huang *et al.* [4] also used SVM for recognizing Taiwanese Sign Language. Ali Karami *et al.* [5] used Wavelet transform and multi-layered perceptron Neural Network for recognizing 32 static alphabets of Persian Sign Language. The colour images are cropped, resized, and converted to gray-scale images instead of hand segmentation. Recognition is performed on bare hands with an accuracy rate of 94.06%. Trong-Nguyen Nguyen *et al.* [6] use PCA and Neural network for recognizing 24 gestures of alphabet (without J and Z) and show an accuracy rate of this combination of 94.3%. Generally, feature extraction is very important in sign language recognition and can determine the performance of the whole system.

In Vietnam, vision-based Vietnamese Sign Language recognition is still a new problem that needs to be solved. As feature extraction plays a determining role in such a recognition system, in this paper, we want to examine and evaluate various methods used for the feature extraction step in the Vietnamese alphabet Sign Language recognition system. The proposed system is represented in Figure 1.

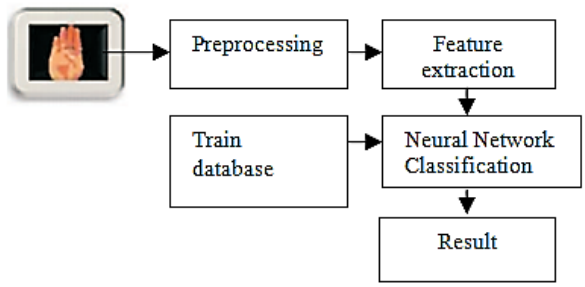

Figure 1. *Proposed system for Vietnamese alphabets Sign Language recognition*

First, a RGB image goes to pre-processing module. The first step in this module is lighting balance then a hand is extracted from the background using skin segmentation in YCbCr space. Binary image, which is the result of skin segmentation, is then de-noised (noise may relate shadow, lighting condition or other objects that have the same color as human skin). Second, the dimension reduction methods are used for feature extraction. Aiming at a real-time recognition system, we focus on methods that have low computational complexity, simple calculation and relative high accuracy.

2. Proposed method

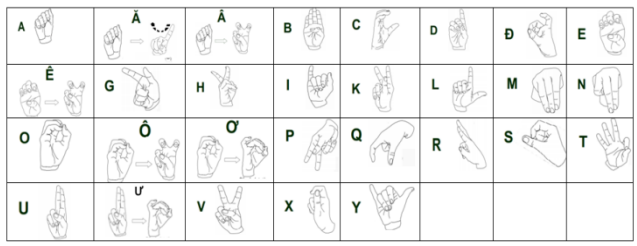

Figure 2. *Vietnamese alphabet Sign Language [7]*

As the present paper considers Vietnamese sign language recognition, the Vietnamese alphabet sign language is shown in the Figure 2.

We see that the special Vietnamese characters (ă, â, ê,

ô, ơ, and ư) are presented by two consecutive gestures. Therefore, Vietnamese characters can be recognized by combining two recognition results of two successive signs.

The following will present the main steps of our proposed method.

2.1. Preprocessing

The RGB image that is captured from webcams or cameras contains not only the hand region but also the background. Therefore, pre-processing is an important module that segments the hand region, removes noise and decreases the impact of illumination on recognition accuracy. In this module, many steps are carried out:

Step 1: Lighting balancing;

Step 2: Hand skin segmentation;

Step 3: Noise removing;

Step 4: Edge smoothing.

2.1.1. Lighting balancing

In this paper, the authors used skin color feature in order to detect the hand region from the background, therefore thresholds for Cb and Cr have to be fixed. Lighting is the main factor that affects these two above channels when converting from the RGB space to the YCbCr space. The effect of this conversion directly impacts the hand segmentation result. Therefore, lighting balancing is the first step that needs to be considered. In this paper, lighting balancing was carried out using Gray World and Modify Gray World methods [8], [9].

2.1.2. Hand segmentation

Segmenting hand from the background was carried out using skin color. The aim of this step is to create the binary image in which the hand region is represented by pixels with intensity of 1 and background's pixels with intensity of 0. First, a RGB image was converted into the YCbCr space as in [10]:

$$Y = 0.299R + 0.587G + 0.114B \quad (1)$$

$$Cb = -0.1687R - 0.3313G + 0.5B + 128 \quad (2)$$

$$Cr = 0.5R - 0.4187G - 0.0813B + 128 \quad (3)$$

Figure 3. *Hand segmentation using YCbCr color space. (a) Original image – (b) YCbCr image – (c) image after segmentation*

The thresholds for hand segmentation were empirically chosen in equation 4 and 5. Similar thresholds in [11], [12] are adjusted to be suitable for our database.

$$132 \le Cr \le 157 \quad (4)$$

$$106 \le Cb \le 128 \quad (5)$$

This resulting binary image was then used to remove noise in the next step.

2.1.3. Noise moving

Outside hand noise

Hand segmentation using skin color also causes noises, which are outside and inside the hand palm. We use a threshold, represented by N that is the number of pixels in a region to determine whether the region is noise.

From our experimental results, outside noise cancelling was implemented by determining the regions in which the number of pixels with intensity equal to 1 is smaller than N. Regions that have less than N pixels (intensity of 1) are misunderstood with hand region because their colors are similar to skin color. After such determining, pixel values of1 of these regions are converted to 0 in order to remove noise.

In this paper, N is equal to 500.

Inside hand palm noise

The main causes that create noise inside the hand palm are shadow, light obscurity. In order to eliminate this noise, firstly, all pixel values of binary image are converted from 0 to 1 and vice versa. Using a threshold of 300 pixels with an intensity value of 1, the noise inside the hand palm is determined and removed by converting this region pixel value to 0. Finally, converting all pixel values was carried out again to recreate the binary image after removing the noise.

Figure 4. *Noise removing for binary image. Left: before noise removing, right: after noise removing*

2.1.4. Edge smoothing

Edge is one of the main elements that feature hand image information. Therefore, smoothing edge without losing information was also considered. Mathematical morphology can be used to smooth the contour, break narrow isthmuses, and eliminate thin protrusions and small holes.

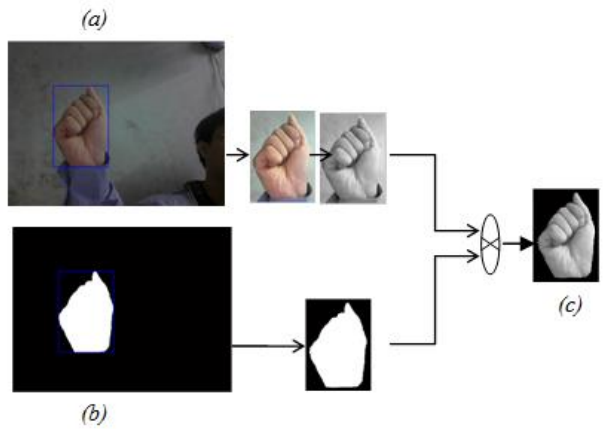

Figure 5*. Pre-processing result: (a) original image, its hand region RGB image and gray-scale image; (b) segmented with noise removed image (binary image) and hand region crop image, (c) gray-scale hand image*

After carrying out these steps, according to the location of the hand region in a binary image (Fig. 5b), the original image and the binary image are cropped to the hand region. The RGB hand image is converted to gray-scale hand image. This image is multiplied with the binary hand image

in order to create the final hand region image (Fig. 5c).In order to tackle with different resolutions of the hand region in each original image, in the final step, the final hand region image is resized to 112×92 pixels.This image was used to extract the feature in the next step.

2.2. Feature extraction

In this section, we consider some filtering methods to improve the quality of hand images and to facilitate the recognition step. Mean and Median filters can reduce noise in the gray image of the hand portion extracted from the background. A Mean filter (sized 3×3) is efficient for noise smoothing. Median filter is also considered to avoid blurring while, at the same time, carrying out noise removing. This is a nonlinear process especially useful for reducing impulse, salt-and-pepper noise. Such a filtered image may then be used for further steps of feature extraction (for example, down-sampling or PCA, see more at the end of this subsection).

On the other hand, if edge-based features of hand images are preferred, edge detection filters can be applied to gray hand images. These filters can be Gradient (derivation approximation), Sobel, Laplacian filters, and many others. For the sake of low computational complexity, Gradient and Sobel filters will be considered in our experiment.

Dimension reduction methods are also used to combine with filtering methods for feature extraction. PCA [6] is one of the most popular methods used to reduce the large dimensionality of the data space to the smaller intrinsic dimensionality of feature space. In the present paper, PCA applied to gray hand images is used as a base-line method for our comparison. We can also combine PCA with the above edge-based detection filters such as Gradient or Sobel filters. In this case, PCA is applied to the output of these filters.

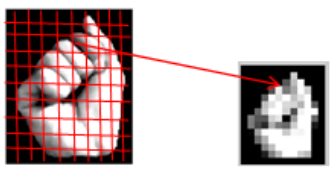

***Figure 6**. Down-sampling result with L = 8 (a) hand segmented image (112×92 pixels) and (b) image after down-sampling with size of 14×12 pixels. One pixel in this image corresponds to a square (red square) in the image in Figure 6a.*

Different from PCA, which creates Eigen space from all train database, in this paper a simple down-sampling method is proposed for feature extraction. A hand image is divided into blocks of L×L pixels, where L is the down-sampling factor. The mean value of each block is calculated and then becomes the value of the corresponding pixel in the down-sampled image (Fig. 6). It is worth noting that each pixel in the down-sampled image corresponds to an L×L block in the input hand image.

Similarly to the above PCA-based methods, we can also combine down-sampling with Gradient or Sobel filter.

These different feature extraction methods will be tested in the experiment.

2.3. Neural network classifier

To complete the sign language recognition method, a neural network classifier is presented in the following.

In this work a three layer feed-forward neural network is used for classification. The input layer consists of features vectors extracted by the dimension reducing algorithm. The hidden layer consists of neurons with weights, the summation function and the transfer function. The latter is a log sigmoid function [13] and is non-linear with the output value between 0 and 1.

In this paper, we select the number of neurons in a hidden layer that is same size as the input layer. The number of neurons in the output layer is the number of signs needed to identify in database. The value of any output node of the positive "1" will correspond to one sign in the database, while the other output nodes will have the value "0". The corresponding sign is the output result of the input image. We have 23 alphabets and 3 accent marks so 26 neurons are chosen at the output layer.

The supervised learning and back propagation algorithm are used for training neural network [13] and the gradient descent method for updating weights. The optimization was based on the mean square error (MSE) with regularization. MSE, gradient error and epochs are the criteria to stop the training. For training phase, the back-propagation learning technique was used. The weights and biases of the NN are updated and adjusted during the training of all patterns.

The parameters are set as follows: learning rate lr = 0.01, value error mse = 1e-10, minimal gradient error = 1e-10, epochs = 1000. The weights are randomly initialized.

3. Experiments and results

3.1. Database

The database includes gestures of 23 alphabets and 3 accent marks were built in order to apply for recognizing Vietnamese alphabet Sign Language. For each alphabet sign language, 100 images are captured with 4 lighting conditions and pose angles by using Sony Vaio SVT14113cxs and DELL INSPIRON webcams with distance varying from 50cm to 80 cm. Therefore, the total number of images in the database for 26 hand signs (23 alphabets and 3 accent marks) is 2600 images.

3.2. Training and testing

The database is divided into 2 parts: 1300 images are used for the training phase, the 1300 other images are used for testing phase. All the experiments are conducted with this division.

While assessing the performance of different feature extraction methods, the experiment aims at:

- Comparing the efficiency of down-sampling methods with PCA. In this part, many different factor (L = 4, 6, 8, 10, 12, 14, 16) are used for down-sampling methods. The dimension of the new image is reduced from 10304 (112×92 pixels) to 644 (28×23 pixels), 304 (19×16

pixels), 168 (14×12 pixels), 120 (12×10 pixels), 80 (10×8 pixels), 56 (8×7 pixels), 42 (7×6 pixels), respectively. Therefore, the number of Eigen vectors in PCA method is equal to dimensions after using down-sampling methods with different factor L.

- Comparing the efficiency of reducing dimension on a gray-scale image with employing this method on an edge-based image. Gradient and Sobel filters are used to create edge-based images. That leads to different scenarios in Table 1 such as Gradient + down-sampling, Sobel+down-sampling, Gradient + PCA, Sobel + PCA.

3.3. Recognition results

Table 1. *Average accuracy rate of recognition*

down-sampling factor L	4	6	8	10	12	14	16
dimension	644	304	168	120	80	56	42
down-sampling	89.4	**89.9**	**89.7**	**90.0**	**90.2**	**90.1**	**91.1**
Gradient+ down-sampling	**83.4**	**85.8**	**88.6**	**88.1**	**90.2**	**87.5**	**86.1**
Sobel + down-sampling	**83.4**	**87.1**	**88.0**	**86.6**	**89.2**	**86.8**	**84.0**
PCA	83.3	85.4	84.3	86.5	88.2	86.7	88.1
Gradient + PCA	**77.1**	**77.8**	**77.0**	**76.7**	**76.8**	**77.1**	**75.9**
Sobel +PCA	78.4	78.6	79.3	77.9	76.8	78.2	77.0

Table 1 shows the average accuracy rate of Vietnamese Sign Language recognition when changing the dimension of segmented hand image and filtering methods. The results show that down-sampling method (with or without filtering methods) gives relative higher accuracy in comparison with PCA-based methods. Moreover, the accuracy rate is stable when decreasing the dimension. The reason is that the down-sampling method with factors L represents global information of an image and also eliminates unnecessary information that related to details or noise. This method helps to hold the shape of the hand (global information). Being different from face or other patterns, which require complex textures, hand shape is very important in sign language recognition. Gesture's form is more important than the details information of the gesture. The increase of L holds global information, hence the results are definitely better. However, the accuracy decreases when L increases to more than 16 because of losing much information in the image. Yet it is very interesting to note that down-sampling has very low computational complexity.

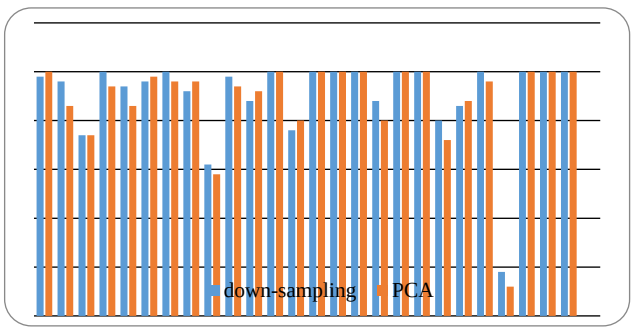

Figure 7. *Recognition rate with 23 alphabets and 3 accent marks*

This result also shows that implementing the down-sampling on gradient-based (edge-based) image does not improve the recognition accuracy. This result is the same with PCA methods. While edge-based images can efficiently represent the contours or boundaries and, hence, to some extent the form/shape of a hand image without storing much of its data, using edge information for recognition is not as good as simply using the raw images. Of course, it would be interesting to confirm this conclusion by testing other edge detection filters and, particularly, some more sophisticated edge detection methods.

When considering the accuracy rate of all 26 signs, we obtained the result in figure 7. Two methods (down-sampling, PCA) are chosen to show their recognition rate. In this figure, **?** notation represents the mark that combines with O, U to create Ơ, Ư and ~ notation represents the mark that combines with A to create Ă.

Some alphabets are recognized with perfectly high results (100% for A, T, M, O, P, Q, S, T, X and 3 marks). However, there are still some alphabets that give low results, especially with Y. In fact, the images used to test the Y sign language are different from the images used to train this sign: the angle's difference is about 45^0. This result shows that the condition for capturing the image much affects the recognition accuracy. If a better training image database is carefully selected, i.e. including many cases of illumination and posing an angle condition, the recognition rate might be improved.

4. Conclusion

In this paper, we evaluate the influence of filtering and dimension reduction for Vietnamese Sign Language recognition while implementing a complete model for this purpose. The experimental results show that the performance of the proposed method concerning down-sampling is relatively high (91%). Such a feature extraction method with low computational complexity might be effective for real-time recognition applications.

This paper considered the recognition accuracy of 23 Vietnamese alphabets Sign Language and 3 accent marks which is the foundation for developing recognition methods for 29 Vietnamese alphabets Sign Language. Further work will be focused on improving the algorithm for recognizing Vietnamese Sign Language for short sentences.

Acknowledgement

This study has been supported by the 3DCS Teaching Research Team at DUT. We kindly thanks for the contributions from Vinh Q. Nguyen, Thanh C. Nguyen and Trung M. Luong.

REFERENCE

[1] T. Starner and A. Pentland, *"Real-time American Sign Language recog-nition from video using hidden Markov models*", MIT Media Lab., MIT, Cambridge, MA, Tech. Rep. TR-375, 1995.

[2] J. Weaver, T. Starner, and A. Pentland, "Real-time American Sign Lan-guage recognition using desk and wearable computer based video", *IEEETrans. Pattern Anal. Mach. Intell*., vol. 33, no. 12, pp. 1371–1378, Dec. 1998

[3] Omer Rashid, Ayoub Al-Hamadi, Bernd Michaelis, "Utilizing Invariant Descriptors for Finger Spelling American Sign Language Using SVM" 6th International Symposium, *ISVC 2010, Las Vegas, NV, USA, November 29-December 1, 2010. Proceedings*, Part I pp 253-263.

[4] Chung-Lin Huang, Bo-Lin Tsai, "A Vision-Based Taiwanese Sign Language Recognition", *Pattern Recognition (ICPR), 2010 20th International Conference*, pp 3683 – 3686, Aug. 2010.

[5] Ali Karami, BahmanZanj, AzadehKianiSarkaleh, "Persian sign language (PSL) recognition using wavelet transform and neural networks", *Expert Systems with Applications***,** Volume 38, Issue 3, Pages 2661–2667, March 2011.

[6] Trong-Nguyen Nguyen, Huu-Hung Huynh, Jean Meunier, " Static Hand Gesture recognition using Pricipal Component Analysis combined with Artificial Neural Network", *Jounal of Automation and Control Engineering*, Vol. 3, No. 1, pp. 40-45, February, 2014

[7] Center for Research and Education of the Deaf and Hard of Hearing (CED), http://trungtamkhiemthinh.org/

[8] Phil Chen, Dr. Christos Grecos, *"A Fast Skin Region Detector"*, Department of EEE, Loughborough University.

[9] Alvarez, Sergio, David F. Llorca, Gerard Lacey, and Stefan Ameling. "Spatial Hand Segmentation Using Skin Colour and Background Subtraction", *Trinity College Dublin's Computer Science Technical Report, Dublin, November, 2010.*

[10] Manel Ben Abdallah, AmeniSessi, mohamadKallel, M.S.Bouhlei M, "Different Techniques of Hand Segmentation in the Real Time*", International Journal of Computer Applications & Information Technology*, Vol. II, Issue I, January 2013.

[11] Oleg Rumyantsev, Matt Merati, Vasant Ramachandran, *"Hand Sign Recognition through Palm Gesture and Movement"*, Stanford University.

[12] AvinashBabu.D, Dipak Kumar Ghosh, Samit Ari, *"Color Hand Gesture Segmentation for Images with Complex Background"*, Department of Electronics and Communication Engineering National Institute of Technology Rourkela –Rourkela - India.

[13] Tom M.Mitchell, "Machine Learning'', McGraw-Hill science/ Engineering /Math , 1997.

(The Board of Editors received the paper on 07/07/2014, its review was completed on 04/09/2014)