Short Survey on Static Hand Gesture Recognition

Huu-Hung Huynh
University of Science and Technology
The University of Danang, Vietnam

Duc-Hoang Vo
University of Science and Technology
The University of Danang, Vietnam

***Abstract*—This paper presents a survey of methods which have been recently proposed for recognizing static hand gestures. These approaches are first summarized and then are assessed based on a common dataset. Because mentioned methods employ different types of input, the survey focuses on stages of feature extraction and classification. Other former steps, such as pre-processing and hand segmentation, are slightly modified. In experiments, this work does not only consider the recognition accuracy but also suggests suitable scenarios for each method according to its advantages and limitations.**

Keywords—Hand gesture; rank-order correlation matrix, Gabor filter; block; centroid distance; Fourier transform

I. INTRODUCTION

Hand gesture recognition is one of the important problems in vision-related fields such as human-machine interaction, communication, and robotic. There are two gesture types including static and dynamic ones. The application of each gesture type depends on the system objective and gesture definition. Static hand gestures are usually identified based on the hand appearance, e.g. contour and shape, while gestures of the other type are mostly recognized according to the change of hand pose and motion trajectory. In this paper, we focus on some recent methods that recognize static hand gestures. The survey considers four approaches proposed in [1]–[4]. These methods are selected because they used different inputs, described gestures by various features, and employed typical classifiers. Each of such approaches can thus be extended to be appropriate with a wide variation of practical gesture collections.

The work [1] aimed to identify static gestures representing the Vietnamese alphabet, in which a character may be represented by either a single gesture or a combination of two hands. The researchers decomposed this task into a problem of static gesture recognition and a combination based on the alphabetic rules. In this survey, we consider the former task. The input image of the method [1] is a depth map captured by a Kinect 1 at 30 fps and resolution of 640×480. In [2], the authors introduced a system identifying Arabic alphabet and number sign language that helps the communication between hard-of-hearing people. A color camera was employed for data acquisition. The gestures are distinguished according to the hand silhouette and hidden Markov model (HMM) technique. The study [3] also captured input images via a RGB color camera. Differently from [2], the feature extraction was performed in frequency domain, and the stage of classification employed Bayesian techniques. The last considered work [4] combined both spatial and frequency domains for describing the hand characteristics before identifying each input gesture by the simple nearest neighbor method. Details of these methods are presented in the next section.

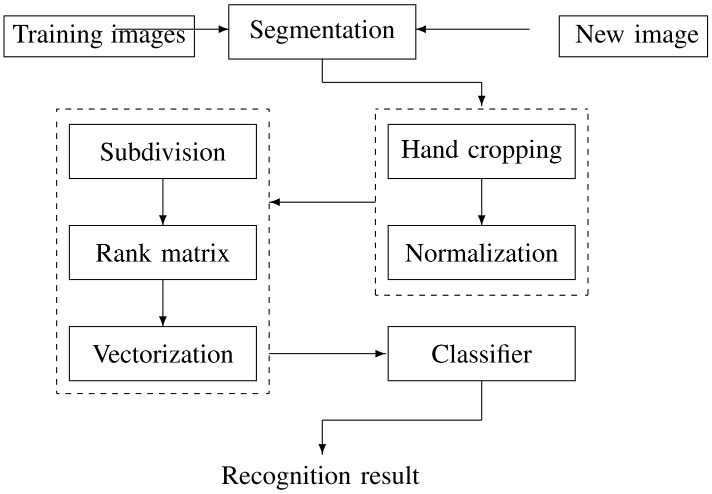

Fig. 1. System overview of the study [1]. The input of the system is a depth map captured by a Kinect 1.

II. LITERATURE REVIEW

Since mentioned studies in this survey employed state-of-the-art classification techniques, the features describing the hand are emphasized. However, the overall processing of each method is also presented.

A. Rank-order Correlation Matrix

The flowchart of the approach [1] is shown in Fig. 1. Since the input is a depth map which directly captures the scene, the hand segmentation is necessary. The authors introduced two methods for isolating the hand from the background and other body parts according to the fact that hand is the object which is nearest to the camera. The difference between these two methods is the definition of the depth range of interest. The first one sets this range at a fixed distance from the camera, thus a user has to remember it and stand in a suitable region when performing the gestures. The second method is more flexible since it defines the range based on the nearest part. This is also the general segmentation for the object of interest in a depth map in some other studies. The segmentation results a depth map of only the hand. A step of normalization is then applied to synchronize the image size of segmented hands. Instead of using a regular resizing method, the researchers pad black regions to some borders in order to keep the dimensional ratio of the hand while the image has a square shape. These padded depth maps are finally resized into a fixed size (e.g. 100×100 pixels).

In order to extract characteristics describing the considering hand, the normalized depth image is divided into square blocks. The hand of interest is thus represented as a square matrix, in which each element is estimated from a block at the corresponding position. Two matrices are proposed, i.e. there are two values that are extracted from each block. They are statistical descriptions including the mean

$$\mu = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{1}$$

and standard deviation

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \mu)^2}{n}} \tag{2}$$

of n pixel's depths x_i inside the considering block. In summary, a normalized depth hand gives two matrices, in which the first one contains d^2 values of μ and the other one has d^2 elements of σ where d is the number of blocks corresponding to each image dimension. The values of these two matrices are not directly employed in the task of classification. Instead, each one is converted into a rank matrix, i.e. each element is replaced by its order after sorting all values. The order is in the range $[0, d-1]$.

Such rank matrices are then transformed into vectors according to a specific rule, in which two adjacent vector's elements correspond to two adjacent blocks in the hand image. Each vector is finally converted into correlation vector that describes the relation between two continuous elements. The combination of two correlation vectors is used as the feature of the considering hand. This feature is called Rank-order correlation matrix (ROCM) though the final representation is a vector. The stage of classification is performed by the support vector machine with the one-vs-one strategy [5].

B. Sorted Block-Histogram

Differently from the work [1], a normal color camera was employed in the study [2]. The overall of this work is presented in Fig. 2.

Since the input is a color image, the authors first detect body parts by applying a skin detection [6]. The image is converted into YCbCr color space in order to separate color channels from the intensity one. To reduce the computational cost, only the chrominance channel (Cr) is then considered. The binary mask of skin-filter result $S_{i,j}$ is formed by applying a checking on pixel's values of the $Cr_{i,j}$ image as

$$S_{i,j} = \begin{cases} 0, & 10 < Cr_{i,j} < 45 \\ 1, & \text{otherwise} \end{cases} \tag{3}$$

The noise removal is then performed to reduce possible noises in the obtained binary image. In details, the mask is decomposed into blocks of 5×5 pixels. Each block is fully filled by white or black pixels according to the ratio of white pixels inside the block. In normal situations, the filtered body parts usually consist of the face and hand. Therefore, the possible face is detected and discarded using a face detector.

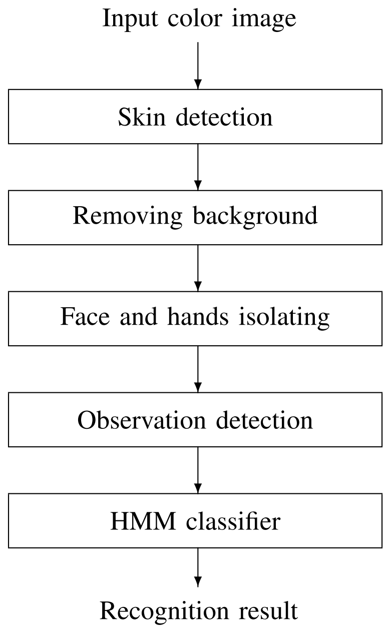

Fig. 2. Main stages of the work [2]. The principal tasks are hand segmentation, feature extraction, and classification.

In the next stage, the segmented hand is processed to provide a feature vector. Similarly to [1], the hand silhouette is also divided into blocks. However, instead of employing statistical descriptions, each block is represented by a simple value which corresponds to the number of white pixels inside it. Every block position is assigned a label, and the feature is formed as a sequence of such labels, in which the label's order is determined based on the sorted pixel-count values.

In order to perform the classification, HMM with discrete observations is employed. The number of observations is the same with the number of blocks (and labels). Many HMMs are built corresponding to the number of gestures. Given the feature vector representing a unknown hand gesture, the returned class is determined as the HMM providing the highest likelihood.

C. Gabor Filtering

The pipeline of the study [3] is shown in Fig. 3. Similarly to the study described in Section II-B, a skin filter is also applied to detect the hand region. Instead of YCbCr, the L.A.B color space is used for representing the input image. The mask corresponding to skin regions is formed by thresholding automatically the channel B. The noise removal is then performed on the resulted binary image using morphological operations such as erosion and dilation, in which the structure element has the size of 5×5 pixels.

In the stage of feature extracting, a collection Gabor filters is employed to emphasize hand characteristics in different orientations. A Gabor filter can be considered as a Gaussian kernel function that is modulated by a sinusoidal wave. It provides two components representing orthogonal directions that include the real and imaginary parts which are respectively

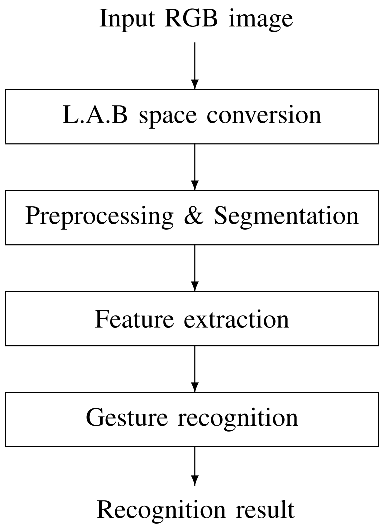

Fig. 3. Pipeline of the work [3]. The stage of feature extraction is performed in frequency domain.

calculated as

$$g_r(x,y;\lambda,\theta,\psi,\sigma,\gamma) = \exp\big(-\frac{x'^2+\gamma^2y'^2}{2\sigma^2}\big)\cos\big(2\pi\frac{x'}{\lambda}+\psi\big) \tag{4}$$

$$g_i(x,y;\lambda,\theta,\psi,\sigma,\gamma) = \exp\big(-\frac{x'^2+\gamma^2y'^2}{2\sigma^2}\big)\sin\big(2\pi\frac{x'}{\lambda}+\psi\big) \tag{5}$$

in which

$$\begin{cases} x' = x\cos(\theta) + y\sin(\theta) \\ y' = y\cos(\theta) - x\sin(\theta) \end{cases} \tag{6}$$

where λ is the sinusoidal factor of the wavelength, θ is the orientation characterizing the Gabor filter, ψ is the phase offset, σ is the standard deviation, and γ is the spatial aspect ratio. These parameters are assigned by the collections of values presented in Table I.

TABLE I. Parameter Values Defining Gabor Filters

Parameter	**Notation**	**Values**
Orientation	θ	$\{0, \frac{\pi}{8}, \frac{2\pi}{8}, \frac{3\pi}{8}, \frac{4\pi}{8}, \frac{5\pi}{8}, \frac{6\pi}{8}, \frac{7\pi}{8}\}$
Wavelength	λ	$\{4, 2^{1/4}, 8, 2^{1/8}, 16\}$
Phase	ψ	$\{0, \frac{\pi}{2}\}$
Aspect ratio	γ	$\gamma = \lambda$

After applying the defined Gabor filters on the segmented hand, a classifier is used to determine the corresponding class of the input gesture. The Bayesian techniques was employed in this work.

D. Centroid Distance Signature

In the final considered approach [4], the spatial and frequency domains are continuously employed for feature extraction. Similarly to the study [1], the input is a depth map captured by a Kinect 1 at 30 fps with resolution of 640×480 pixels. The overall of the system is presented in Fig. 4. In order to segment the hand from input depth image, the researchers first employ a functionality in the Kinect SDK to isolate the human body from the background. Possible hands are then determined and separated by applying a threshold on depth layers representing the whole body. The k-mean clustering technique is then performed to check the number of hands. The number of clusters is thus assigned to 2. In the case that there is only one hand, two clustered regions are combined. This case is detected by comparing the distance between two centroids with a predefined threshold. The next stages are to recognize each hand gesture.

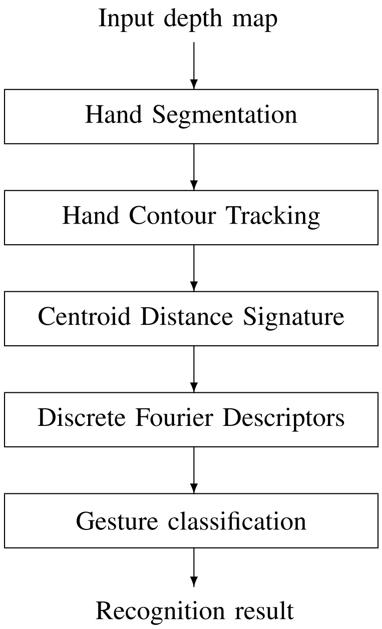

Fig. 4. The gesture recognition approach proposed in [4].

Instead of considering the entire pixels representing the hand, only the contour is focused. The 8-connectivity algorithm [7] is applied to perform this task. The centroid of these determined pixels is also estimated. In the next step, the feature vector is formed as a sequence of distances between contour pixels and the centroid. This vector is finally converted into frequency domain using the Fast Fourier Transform (FFT) [8]. However, the FFT requires that the input must have a power-of-two elements. Besides, the lengths of sequences of contour points corresponding to different hands also need to be normalized for the classification stage. Therefore the researchers converted feature vectors to 128-element ones based on equal angle sampling. The recognition is finally performed based on the nearest neighbor with Euclidean distance.

These described studies are summarized in Table II.

III. Experiments

The survey is performed on four different methods for static hand gesture recognition. The selection of dataset is thus important since it must be appropriate for all of these approaches. We employ the dataset provided in the study [1]. This is the collection of segmented depth hands corresponding to 23 static gestures. The dataset consists of 4637 images that were captured by the depth sensor of a Microsoft Kinect 1 and segmented based on depth thresholds. A visualization of these gestures is presented in Fig. 6. This dataset is suitable for methods processing on depth images as well as hand silhouettes. Therefore the stage of hand segmentation in considering

TABLE II. SUMMARY OF FOUR STUDIES CONSIDERED IN THIS PAPER

Method	Input	Hand segmentation	Feature extraction	Classification
[1]	Depth map	Depth thresholding	Block dividing & Statistical values	Support vector machine
[2]	Color image	Skin filtering in YCbCr	Block dividing & Sorted indices	Hidden Markov model
[3]	Color image	Skin filtering in L.A.B	Gabor filtering	Bayesian technique
[4]	Depth map	Body detection & Depth thresholding	Centroid distance signature & FFT	Nearest neighbor

TABLE III. GESTURE RECOGNITION ACCURACIES CORRESPONDING TO FOUR CONSIDERING METHODS

Approach	Vo *et al.* [1]	Abdo *et al.* [2]	Ashfaq *et al.* [3]	Gani *et al.* [4]
Recognition accuracy	94.22%	87.71%	86.98%	98.32%

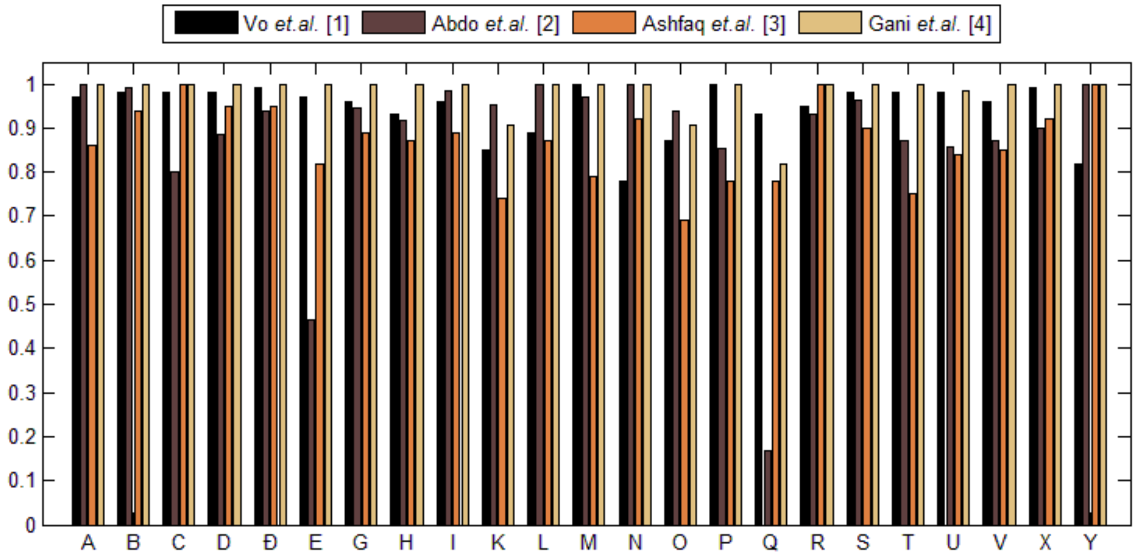

Fig. 5. Accuracies when recognizing each gesture of the benchmark dataset.

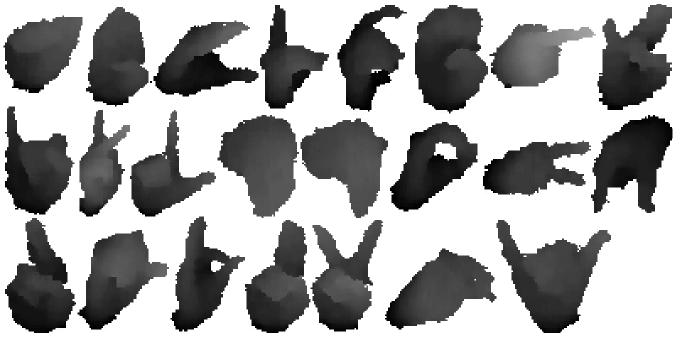

Fig. 6. Benchmark dataset including 23 static hand gestures which represent 23 letters in Vietnamese alphabet.

studies is not necessary to perform in these experiments. In other words, this paper accesses only the combination of feature extraction and classification. The overall accuracies of considered methods in recognizing gestures corresponding to 23 classes are shown in Table III. The details are also presented in Fig. 5.

According to Table III, the method proposed in [4] has best ability in classifying 23 gestures in the benchmark dataset. Frequency domain is thus an appropriate choice for improving current works on hand gesture recognition. This method also provides absolute accuracies when identifying many gestures (see Fig. 5). Therefore the processing flow in [4] should be applied at the beginning of a study on gesture recognition. The approach gives the second highest accuracy is the study [1]. The partial recognition accuracies corresponding to 23 classes are mostly similar. Therefore the statistical information extracted from regions and/or blocks of depth maps should also be focused when dealing with the problem of gesture recognition. However, the work [1] requires a depth camera for data acquisition while the study [4] does not. The accuracies of two other methods are slightly different. However, the experimental values corresponding to several gestures in [2] are quite low (e.g. E and Q). It means that the simple feature proposed in [2] may be significantly affected by the gesture selection. In other words, the ability of gesture recognition may be very different when applying the method on various collections of gestures. A validation for each particular dataset is thus necessary. The remaining study [3] employs a collection of Gabor filters for feature extraction. This method thus requires a large number of operations. In our experiments, the computational cost of [3] is very high compared with the others. Therefore this approach is only suitable for (1) systems with high strength of computation and/or (2) low-resolution input images.

IV. CONCLUSION

This paper presents a survey on some recent methods which perform the static hand gesture recognition. These studies employ different input data types, extracted features, and classification techniques. Details of each processing stage corresponding to each method are described. The experiments are performed on a common dataset in order to provide a comparison on recognition ability of these approaches. According to the obtained results, advantages and limitations of them are also given.

In future works, a combination of these approaches and some recent ones will be focused to improve the ability of recognizing more complicated static gestures. In detail, we attempt to estimate gesture scores based on a weighted sum of likelihoods computed from each individual method. Such classification fashion is expected to take advantage of each approach. In addition, each considered method can be extended, by employing more features and/or combining with different weak classifiers, to be appropriate with a specific practical gesture collection.

ACKNOWLEDGMENT

The authors would like to thank Trong-Nguyen Nguyen, DIRO, University of Montreal, for his technical support.

REFERENCES

[1] D. H. Vo, T. N. Nguyen, H. H. Huynh, and J. Meunier, “Recognizing vietnamese sign language based on rank matrix and alphabetic rules,” in *2015 International Conference on Advanced Technologies for Communications (ATC)*, Oct 2015, pp. 279–284.

[2] M. Z. Abdo, A. M. Hamdy, S. A. E.-R. Salem, and E. M. Saad, “Arabic alphabet and numbers sign language recognition,” *International Journal of Advanced Computer Science and Applications*, vol. 6, no. 11, pp. 209–214, 2015.

[3] T. Ashfaq and K. Khurshid, “Classification of hand gestures using gabor filter with bayesian and naïve bayes classifier,” *International Journal Of Advanced Computer Science And Applications*, vol. 7, no. 3, pp. 276–279, 2016.

[4] E. Gani and A. Kika, “Albanian sign language (albsl) number recognition from both hand’s gestures acquired by kinect sensors,” *International Journal Of Advanced Computer Science And Applications*, vol. 7, no. 7, pp. 216–220, 2016.

[5] J. Milgram, M. Cheriet, and R. Sabourin, ““One Against One” or “One Against All”: Which One is Better for Handwriting Recognition with SVMs?” in *Tenth International Workshop on Frontiers in Handwriting Recognition*, G. Lorette, Ed., Université de Rennes 1. La Baule (France): Suvisoft, Oct. 2006.

[6] Y. t. Pai, S. j. Ruan, M. c. Shie, and Y. c. Liu, “A simple and accurate color face detection algorithm in complex background,” in *2006 IEEE International Conference on Multimedia and Expo*, July 2006, pp. 1545–1548.

[7] T. Pavlidis, *Algorithms for graphics and image processing*. Springer Science & Business Media, 2012.

[8] J. W. Cooley and J. W. Tukey, “An algorithm for the machine calculation of complex fourier series,” *Mathematics of computation*, vol. 19, no. 90, pp. 297–301, 1965.