J. Vis. Commun. Image R. 38 (2016) 658–677

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

journal homepage: www.elsevier.com/locate/jvci

Signer-independence finger alphabet recognition using discrete wavelet transform and area level run lengths ☆

Kanjana Pattanaworapan [a], Kosin Chamnongthai [a,*], Jing-Ming Guo [b]

[a] Department of Electronic and Telecommunication Engineering, Faculty of Engineering, King Mongkut's University of Technology Thonburi, 126 Pracha Uthit Rd., Bang Mod, Thung Khru, Bangkok 10140, Thailand
[b] Department of Electrical Engineering, National Taiwan University of Science and Technology, No. 43, Section 4, Jilong Road, Da'an District, Taipei City 10607, Taiwan, ROC

ARTICLE INFO

Article history:
Received 11 February 2015
Revised 23 February 2016
Accepted 18 April 2016
Available online 19 April 2016

Keywords:
Sign language recognition
Finger alphabet recognition
Discrete wavelet transform
Sign grouping
Signer independence
Run length algorithm
Sign speaking system

ABSTRACT

This paper proposes a method for finger alphabet recognition from backhand images with signer-independence. Input images that are divided into fist sign and non-fist sign groups should be analyzed and processed in different ways. Finger alphabets in the fist group are represented by a one-dimensional signal that represents the external hand boundaries. Its low and high frequency components are then extracted by discrete wavelet transform, which are key features for recognition. The non-fist sign images, which are radically digitized into a 20×20 block mask in terms of the hand geometry, due to the hand's physical structure, can be recognized by the patterns of the occupied blocks. The experimental results show that the proposed method has a high likelihood of differentiating twenty-three static finger alphabets of backhand images. The proposed method reaches an improvement of 27.86% in recognition accuracy on a significant dataset of fist signs that includes multiple users, while the statistical distribution of the area level run length algorithm outperforms previous forehand approaches by 89.38% in recognition accuracy.

1. Introduction

There are approximately seventy million hearing impaired people around the world [1]. In communication between normal-hearing people and hearing impaired people, the normal-hearing people must understand sign language. In fact, normal-hearing people use sound to communicate with other people. Therefore, there should be a sign language recognition system that can convert the sign appearance into an audio solution.

Although car-driving records show that hearing impaired drivers are as good as normal-hearing drivers, hearing impaired people have limitations in their driving ability. Integrating sign language recognition into Advanced Driver Assistance Systems (ADAS) applications could not only improve the active safety aspects of the vehicle [2] but also increase the demands on the automotive industry.

In developing a sign language recognition system and ADAS, users who are hearing impaired people (signers) need a mobile system that has a light weight and tiny size to have a system that is convenient for daily life. The camera, which is an important sensor for receiving pictures of sign language, must be set up in front of the signers in the same direction as the audience. It might not be convenient to always move this camera system following the signer movement during his daily life. If the camera and sign language recognition system are attached to the signer's body, then the camera and system would automatically move together with the signer. The sign language pictures in this system can be interpreted into natural language whenever the signer wants to communicate with normal-hearing people. Therefore, the camera is considered to be fixed at the chest of a signer, and its lens could receive a picture of the sign language from the backhand view. In this case, the system developers must create algorithms for understanding sign language by the signer's hand pictures using the backhand instead of the forehand viewpoint.

Sign language is a special language that consists of hand postures and hand gestures. A hand posture is a static sign, while a hand gesture includes hand movement. Sign language includes the alphabet, word and sentence levels. The sign language dictionary contains only approximately 10,000 vocabulary words, while the English dictionary has approximately 180,000 vocabulary words. Many vocabulary words cannot be expressed in sign language [3]. Only the alphabet level can perform undefined vocabulary words, by using the finger spelling alphabet, and as a

☆ This paper has been recommended for acceptance by M.T. Sun.

* Corresponding author.
E-mail addresses: kanjana.pa@bu.ac.th (K. Pattanaworapan), kosin.cha@kmutt.ac.th (K. Chamnongthai), jmguo@seed.net.tw (J.-M. Guo).

http://dx.doi.org/10.1016/j.jvcir.2016.04.015
1047-3203/

result, many researchers are interested in finger alphabet recognition.

To develop a sign language recognition system using the backhand approach, unfortunately, research on sign language recognition systems based on backhand images has not yet appeared. We therefore review the studies that are based on forehand images as related work.

Sign language recognition (SLR) using forehand [30–32] can be divided into two main categories, which are based on different sensing devices: touch-sensor-based systems [4] and vision-based systems [5]. Touch-sensor-based systems acquire finger information such as the finger position, joint angle and relative distance between the fingers, to provide good results [6–11]. These methods can be applied to backhand-view processes because the finger information that is acquired by the touch sensor does not depend on whether backhand or forehand is used. However, the users could feel uncomfortable from the many cables that are required, even when there is higher accuracy. Vision-based systems provide an extensive area to overcome these limitations when using mobile systems.

The methods in vision-based systems can be divided into two groups: color glove usage and bare hands. In the group of color glove usage, Lamar [12] used color gloves with markers to make Japanese and American finger alphabets. PCA is applied to the pixel distributions of all of the color regions, and a neural network is used for classification, with an 89.1% and 93.3% average recognition rate, respectively. Akmeliawati [13] developed an automatic sign language translator using gloves with specific color patches to determine the fingertip positions. The colored markers in [12,13] depicted the geometric characteristics of the hand that are important for sign language recognition. Although users in this group do not need to wear cumbersome equipment, such as touch sensors, wearing color gloves still places them in unnatural conditions, especially in mobile cases.

In the bare hands group, the methods are divided into four subgroups: spatial geometric information on the bare hand, frequency domain transform, graph-like hand model, and hybrid system between spatial and frequency domain information.

In the spatial geometric information, Krishnaveni and Radha [14] extracted a set of statistical descriptors from the exterior boundaries of the hands to clearly distinguish the signs of Indian sign language. Sharma et al. [15] used a contour tracing descriptor to recognize 11 ASL finger alphabets with a 76.82% recognition accuracy. This method is robust to rotation and size invariance. Munib et al. [16] applied Canny's edge detection algorithm to extract both the exterior and interior boundaries of the hands and determined the feature descriptors of 14 alphabets, 3 numbers and 3 vocabularies by using the Hough transform and neural network. Information from the interior boundaries enhanced the sign recognition accuracy performance. Uras and Verri [17] proposed sign language recognition using size functions that represent apparent hand outlines of various signs with seventy feature vectors. Handouyahia et al. [18] used a moment-based size function to decrease the number of feature vectors of the size functions in previous work [17]. A size function is robust to scale, translation and orientation invariance and can recognize signs with a good percentage, but it is not efficient for the user-independent condition. Menotti et al. [19] compared the Hu and Zernike moments for sign language recognition; these moments are also extracted from the hand boundaries. The experiments show that the Zernike moment features provided better results. Kelly et al. [20] combined the size functions and Hu moments to recognize 10 ASL alphabets that were signed by independent users. Dahmani and Larabi [21] applied the Tchebichef moment on the internal and external hand outlines, the Hu moments and a set of geometric feature descriptors. The experiments indicated that using both internal and external edges leads to better hand posture recognition. The mentioned methods in the spatial geometric information sub-group can recognize unambiguous signs with high accuracy in both forehand and backhand views. However, these methods have limitations with some of the similar signs, which are differentiated by rotation and statistical distributions, such as "H and U and R", "G and V and L", "I and D and R", "F and W" and "V and Y and G and L".

In the frequency domain transform sub-group, Karami et al. [22,23] selected the Haar wavelet transform to derive the feature vectors from Persian sign language, to train a neural network. The experimental results are 83.03% recognition accuracy in the testing data. However, this method involves a large number of feature descriptors. Ashrafulamin and Yan [24] worked on a method of phase-invariant Gabor filters at five scales and eight different orientations and PCA to classify 26 finger alphabets with a 93.23% accuracy. Gabor filters require a high computational cost, and thus, they cannot be used for classification in real time. The methods in this sub-group theoretically improve the accuracy compared with the spatial geometric information sub-group. They might be able to address backhand as well. However, a large amount of 2D data of hand images is employed as input data to the wavelets in such a way that differences were obtained among especially similar signs in the fist group, which are sometimes too small to differentiate.

In the sub-group of the graph-like hand model, the hand's physical structure has been used to understand the meaning in the signs by using different approaches. Triesch and Von Der Malsburg [25] proposed Elastic Graph Matching (EGM) to represent a hand model of 10 ASL finger alphabets against complex backgrounds. Marnik [26] also employed EGM to distinguish 23 Polish finger alphabets. Stergiopoulou and Papamarkos [27] applied a Self-Growing and Self-Organized Neural Gas Network (SGONG) on a hand area to approach its shape. The network started with only two neurons, and new neurons are inserted during a growing stage to achieve better data clustering. The grid of neurons should successfully approximate the anatomical characteristics of the hand. These characteristics accomplish the identification of gestures with separated extended-fingers. Flasinski and Myslinski [28] recognized Polish finger alphabets by using a graph grammar parsing model. Although the graph-like hand model achieves a description of the hand geometries, which are very useful information for providing a high recognition rate, it has a high computational complexity when analyzing a single image. These methods are not yet suitable for a mobile system that requires a real-time process.

To develop a real-time processing system, some of the researchers considered a hybrid system between the spatial and frequency domain information. Pugeault [29] employs Microsoft Kinect to collect depth data on all of the static alphabets in ASL and used a multiclass random forest classifier. The depth information is robust to the illumination and skin color differences, which affect the detection process. However, the depth in the ASL finger alphabet is too small to differentiate each sign. Moreover, useful depth information is located in the front of the hand, which cannot be seen from the backhand perspective. The depth information is ultimately useless for the system when using a backhand view.

The existing methods mentioned above in the spatial geometric information, frequency domain transform and graph-like hand model group can be concluded to perform well with forehand images, which are obviously those seen by audiences. Some information, such as finger's boundaries and fingertip positions, is obviously seen in the forehand scenarios, but they are mostly occluded by a signer's palm when considered from the backhand view. Moreover, some existing systems that provide high recognition rates are limited by the time consumed, user dependences or number of recognized signs in such a way that they are not implementable in a real-time application.