Consideration of a Selecting Frame of Finger-Spelled Words from Backhand View

Ponlawat Chophuk*, Kanjana Pattanaworapn**, and Kosin Chamnongthai*
* Department of Electronic and Telecommunication Engineering
Faculty of Engineering, King Mongkut's University of Technology Thonburi, Bangkok, Thailand
** Bangkok University, Bangkok, Thailand
Email: ponlawat.baw@mail.kmutt.ac.th, kanjana.pa@bu.ac.th, kosin.cha@kmutt.ac.th

***Abstract*—To understand finger alphabet from backhand sign video, there are many redundant video frames between consecutive alphabets and among video frames of an alphabet. These redundant video frames cause loss in finger alphabet understanding, and should be considered to delete. This paper proposes a method to select significant video frames of sign for finger-spelled words of each letter to make more information from backhand view. In this method, finger-spelled words video is divided into frames, and each frame is converted to a binary image by an automatic threshold, and a binary image change to contour frames. Then, we apply the located centroid as the center of the contour image frame to calculate the distance to all boundaries of image frames. After that, all distances of each frame are presented as signature signals that identify each frame, and these values are used with the selected frame equation to select a significant frame. Finally, 1D Signature signal as their feature is extracted from selected frames. For evaluation of our proposed method, 6 samples of finger-spelled words of the American Sign Language (ASL) are used to select a significant frame, and Hidden Markov Models (HMM) is used to classify the words. The accuracy of the proposed method is evaluated 97.5% approximately.**

I. INTRODUCTION

People who get hearing loss can learn to communicate through development of lip-reading skills, use of written or printed text, and sign language. We focus on sign language because they are popularly used in society for sign language services, such as TTRS cabinet [1], television [2] and so on. In real-life, hearing-impaired persons want the device that can be used to communicate with a normal person. So, sign language device has been developed to be a small size and portable system for society. However, throughout its history, most research used forehand view system which is difficult to move because a system had to set in front of a user, and this system cannot use with backhand view system which the device is set in front of the chest to move to follow a signer. Unfortunately, in backhand view, the system had a problem in similar alphabet such as "M", "N", "S", "T" and "E" because all of them have had similar shape. If these alphabets are used to make word, there might be misunderstanding for meaning.

One of the main aims of this system is to make more information from backhand view which can make more accurate because the signer wants the device that can be communication perfectly with a normal-people.

Recently, there has been growing interest in an American Sign Language Interpreter by using many devices. In 2013, Philip Hays et al. [3] proposed real-time sign to text translation of ASL signs by processing a live video stream using mobile device from forehand view. The experimental results showed that the recognition rate reached 96.21%. However, they also had problem with letter "J", "N", "M", "O" and "S" because they are very similar. In 2015, Edwin Escobedo et al. [4] developed a new method for finger spelling recognition using depth information from Kinect sensor for solving a similar alphabet. They converted the depth data in a 3D point cloud. The point cloud is divided into sub regions, using direction cosines. Their approach had an accuracy rate of 99.37%. Nevertheless, this system could not use with backhand view. So, In 2013[5], they used mobile camera with the pixel-based hierarchical feature to recognize five signs in ASL such as "D', "I", "R", "U", and "X" from backhand view. The mobile is set in front of the chest of user. The result showed the accuracy of 68.4%. Then, 2016 [6] developed the system using discrete wavelet transform and area level run lengths to recognize 23 alphabets based on backhand images. But, they had a problem with fist sign group such as "A", "E", "O", "S", "T", "M", and "N" because they are similar alphabet. The system works well with 23 alphabets which are basically able to recognize by a static frame except the rest three alphabets which need video information. In fact, a signer may spell alphabets consecutively in term of video so that we should set up consecutive alphabet spelling in a video as research problem. Therefore, authors initially started to consider a group of video frames of each alphabet in 26 finger alphabets [7], and classification of an alphabet is proved to be possible. In practice, since finger alphabets in sign language are expressed a word or a phase by the gesture of a hand and its fingers consecutively, the video with consecutive frames has to be segmented in each finger alphabet. Moreover, there exists data redundancy in a segmented finger alphabet which normally consists of many consecutive video frames. This paper therefore tried to find a simple way to segment borders of a finger alphabet, and consider redundant frames for efficient data compression.

II. PROPOSED SEGMENTATION AND KEY FRAME SELECTION

In video frames of finger alphabet spelling, many consecutive frames can be classified as similar sign in a same group as shown in upper rectangle in Fig. 1. If differential values of adjacent frames are calculated, both sides of border with adjacent alphabet can be detected by peaks as shown by a sample at 15th and 30th frames. This means borders of an alphabet can be simply detected by differential values, and one of consecutive frames in the group should be selected as representative frame of the alphabet. Since all of them are similar in shape, this paper proposes to select the middle frame as representative significant frame of the alphabet.

Scheme overview is shown in Fig. 2. It starts from video input process. The video is first processed by preprocessing and key frame selection which is explained in the next sub-sections. The selected key video frames are then performed feature extraction and classification.

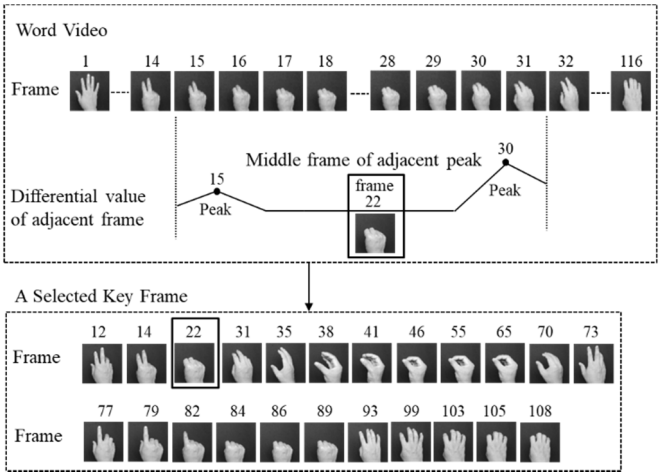


Fig.1 Concept of Key Frame Selection

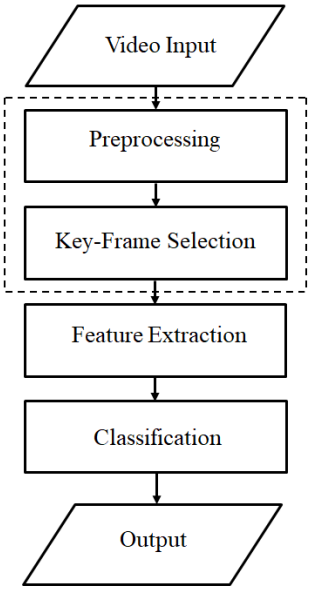


Fig. 2 Scheme overview

Preprocessing is depicted in Fig. 3. A grayscale frame is used to plot a histogram as shown in the first row, and an appropriate threshold value is determined for binarization. In Binarization, as shown in the second row, hand area and background are separated by the threshold value, and hand contour is then detected as shown in the third row. Finally, signature which is 1D information is obtained by a centroid of the hand contour and Euclidian distances from the centroid to all pixels on the contour as shown in the forth row.

A. *Preprocessing*

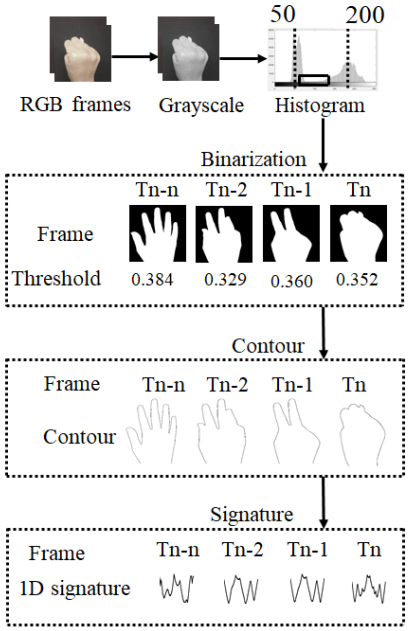
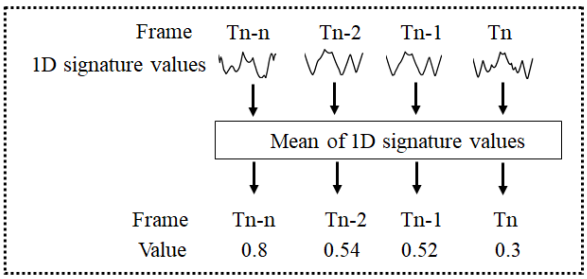


Fig. 3 Preprocessing

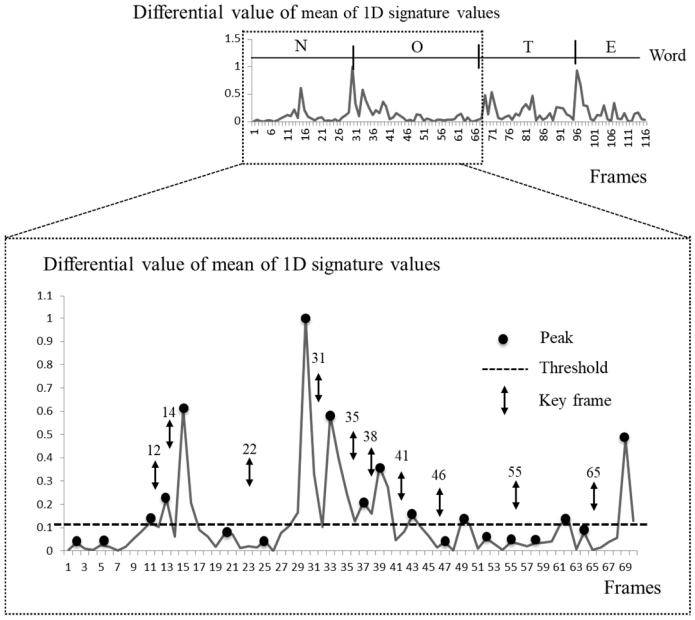
$$D = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \qquad (1)$$

where: "D", $(x_2 - y_2)$, and $(x_1 - y_1)$ stand for Euclidian distances, centroid, and contour point, respectively.

B. *Key Frame Selection*



(a) Mean value of 1D signature values



(b) Key frames of consecutive finger alphabets, "NOTE"

Fig. 4 Key Frame Selection

As the next process, key frame selection is conceptually shown in Fig.4. Amplitudes of a 1-D signal representing signature of hand shape are first used to calculate its mean as shown by a sample in each frame in Fig. 4 (a). The mean values are calculated to obtain the threshold value according to equation (2), and local peaks are simply detected by slope-change pattern among consecutive frames as shown by black dots in Fig. 4 (b). These peaks are regarded as borders of same video frame group so that one of video frames between two neighboring peaks has to be selected as representative video frame or key frame of the group, and the middle frame in the group is selected to represent the key frame of the video frame group in this paper. Equation (3) shows how to select the key frame in the middle of video sequence. By selecting the key frame in the proposed method, redundant frames in the video frame group are considered to reduce.

$$T = \left(\frac{\sum_{i=1}^{n} x_i}{n}\right) + \left(\left(\frac{\sum_{i=1}^{n} x_i}{n}\right) * 0.25\right) \qquad (2)$$

where: "T" is an auto threshold, "x_i" is 1 value per frame, "n" is the number of terms in the sample, $\sum$ is sum of all data values.

$$K = \frac{p_x - p_y}{2}, \quad K = fix(K) \qquad (3)$$

where: "K" is key frame, "P_x" is peak "x", "P_y" is peak "y", fix ( ) is rounds toward zero.

C. Feature Extraction

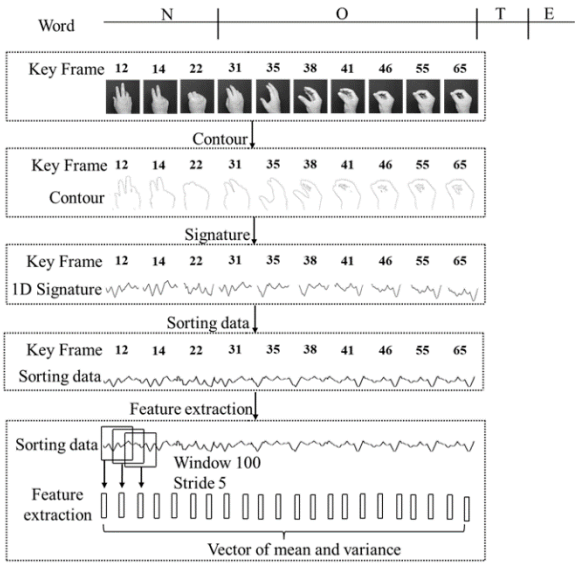


Fig. 5 Feature Extraction

The feature extraction is shown in Fig. 5. A word spelling "N" and "O" is used to demonstrate the process. Several frames spelling "N" and "O" as shown in the first row are converted into contours as shown in the second row, and the contours as 2D signals are transformed into signature as 1D signal shown in the third row. These signatures are subsequently merged as signal sequence, and performed feature extraction by HMM [8].

D. Classification

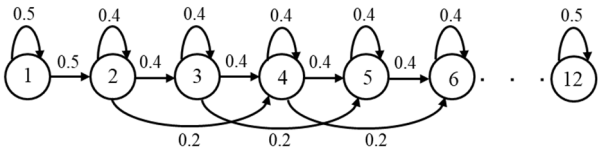


Fig. 6 Hidden Markov Model Toolkit [8]

For classification, Hidden Markov Model is applied in this system to classify six words as shown in Fig. 6.

III. Experiment and Results

TABLE I. Specification of experiment

Video	Frame rate: 29 frames/second
	Time: 4-6 second
	Background: Black
Operating System	Dell G3 Gaming-w56691425TH CPU: Intel Core i7-8750H GPU: NVidia GeForce GTX 1050Ti Memory Size: 8 GB DDR4 Hard Disk Drive: 1 TB

Experiment	Sample:	6 words
	Testing:	4 people (20 times/word)
	Training:	4 people (20 times/word)

The specification of experiment is explained in Table I, the properties of video are frame rate of 29 frames/second, time of each word of 4-6 second and background of black. Parts of an Operating System are Dell G3 Gaming-w56691425TH, CPU of Intel Core i7-8750H, GPU of NVidia GeForce GTX 1050Ti, memory size of 8 GB and hard disk drive of 1TB. For experiment, 6 words are used. 20 times/word from 4 people are used for training, and the other half is for testing.

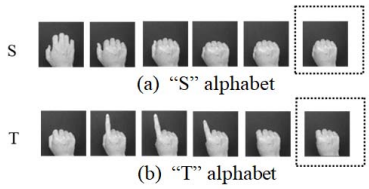


Fig. 7 Key Frame of Word Sample

Fig. 7 depicts key frames of a couple of sample alphabets, "S" and "T" as shown in Fig. 7 (a) and (b), respectively. Although those two alphabets look similar in the last frames (surrounded by circles) which are basically used to recognize, they are able to classify when rewind through previous frames of those alphabets.

TABLE II. The recognition rate of word by proposed method

	Cast	East	Neat	Nest	Nose	Note
Cast	20	0	0	0	0	0
East	0	20	0	0	0	0
Neat	0	0	19	1	0	0
Nest	0	0	2	18	0	0
Nose	0	0	0	0	20	0
Note	0	0	0	0	0	20

In experiments, video frames of six words, "cast", "east", "neat", "nest", "nose", and "note" were used as samples to evaluate performance of our proposed method. Those words are demonstrated by four sign-language experts by signing five times per word. Sign frames demonstrated by two persons out of four are used for training in advance. The evaluation results are shown in Table II where number of correctly classified samples are allocated in the diagonal elements of the table. A sample of "NEST" are misclassified to "NEAT" due to their similarity.

Fig. 8 Errors of alphabet classification

Fig. 8 shows errors of finger alphabet classification. Since the video sequence of "S" looks similar to "A", classifier may confuse and produce an incorrect result. Due to less amount of samples for training and testing processes in the experiments, more data should be considered to add in the training and evaluation as future works.

IV. Conclusion

This paper basically considered video frames of finger alphabets in English sign language in backhand views, tried to reduce redundancy in term of classification, and selected significant video frames for classification. A video-frame sequence of a word consisting of several finger alphabets were considered significance in this paper. Some consecutive frames which looked similar were grouped in the same meaning signs, and the middle of the frame sequence was selected as the key frame. The performance evaluated by some sign-language experts showed significant improvement of our proposed method.

Acknowledgment

The research project is financially supported by Petchra Pra Jom Klao Research Scholarship of King Mongkut's University of Technology Thonburi. The authors are thankful for dataset of hand video collected from the Setsatian School and Sot Sueksa Thung Mahamek School for the deaf in Bangkok, Thailand.

References

[1] "Thailand introduces videophone booths for people with hearing disabilities,"http://globalaccessibilitynews.com/2015/04/21/thail and-introduces-videophone-booths-for-people-with-hearing-disabilities/ (last access: May 2019).
[2] "ASL sign for: television (TV)," https://www.handspeak.com /word/search/index.php?id=2639.
[3] H. Philip, P. Raymond and M. Roy, "Mobile device to cloud co-processing of ASL finger spelling to text conversion," Image Processing Workshop (WNYIPW), Western New York, 2013.
[4] E. Edwin and C. Guillermo. "Finger Spelling Recognition from Depth Data Using Direction Cosines and Histogram of Cumulative Magnitudes," 28th SIBGRAPI Conference on Graphics, Patterns and Images, 2015.
[5] P. Kanjana, C. Kosin, and G. Jing-Ming. "Hand gesture recognition using codebook model and Pixel-Based Hierarchical-Feature Adaboosting," Communications and Information Technologies (ISCIT), 2013 13th International Symposium on. IEEE, 2013.
[6] P. Kanjana, C. Kosin, and G. Jing-Ming. "Signer-independence finger alphabet recognition using discrete wavelet transform and area level run lengths," Journal of Visual Communication and Image Representation 38 (2016): 658-677.
[7] C. Ponlawat, P. Kanjana and C. Kosin. "Backhand-based video frame selection for finger alphabet recognition," International Symposium on Multimedia and Communication Technology.
[8] "HTK," http://htk.eng.cam.ac.uk/ (last access: May 2019).