Deep Learning for Vietnamese Sign Language Recognition in Video Sequence

Anh H. Vo, Van-Huy. Pham, and Bao T. Nguyen

***Abstract*—With most of Vietnamese hearing impaired individuals, Vietnamese Sign Language (VSL) is the only choice for communication. Thus, there are more and more study about the automatic translation of VSL to make a bridge between hearing impaired people and normal ones. However, automatic VSL recognition in video brings many challenges due to the orientation of camera, hand position and movement, inter hand relation, etc. In this paper, we present some feature extraction approaches for VSL recognition including spatial and scene-based features. Instead of relying on a static image, we specifically capture motion information between frames in a video sequence. For the recognition task, beside the traditional method of sign language recognition such as SVM, we additionally propose to use deep learning technique for VSL recognition for finding the dependence of each frame in video sequences. We collected two VSL datasets of the relative family topic (VSL-WRF) like father, mother, uncle, aunt.... The first one includes 12 words in Vietnamese language which only have a little change between frames. While the second one contains 15 with gestures involving the relative position of the body parts and orientation of the motion. Moreover, the data augmentation technique is proposed to gain more information of hand movement and hand position. The experiments achieved the satisfactory results with accuracy of 88.5% (traditional SVM) and 95.83% (deep learning). It indicates that deep learning combining with data augmentation technique provides more information about the orientation or movement of hand, and it would be able to improve the performance of VSL recognition system.**

***Index Terms*—Vietnamese sign language (VSL), VSL recognition, local descriptors, spatial feature, scene-based feature, Motion-based feature, deep learning.**

I. Introduction

Sign language is the natural language of deaf people, known as grammatically complete and copious language as spoken language. Recognition of sign language not only brings a connection between deaf people to normal one, but also plays a noteworthy role in many applications such as intelligent control systems, robotic, human computer interaction, smart home, etc. Four main components describing a language sign are *hand shape, location in relation to the body, movement of hands,* and *the orientation of palms*. However, they are not international language because of the different expressions of signs among countries, regions. Vietnamese Sign Language (VSL) is based on the established American Sign Language (ASL), but VSL also has some special signs which are not involved in ASL dictionary.

In this paper, we try to deal with VSL recognition task for real world applications. We proposed to use two types of feature for VSL recognition *(spatial features and scene-based feature)* to capture the important information of language signs. For the recognition task, beside the traditional classification method of sign language recognition such as SVM, in this work, one state-of-the-art technique, deep learning was additionally used to recognize VSL with the aim of finding out the dependence of each frame in video sequence. The using of deep learning technique comes from the prior-knowledge that: in VSL or any sign language system, one sign expression is not only represented by gesture itself, by also depends on the movements of hand between each frame in video sequence. And deep learning has the capability of capturing information from many frames within a video sequence, thus it would able to discriminate signs accurately.

Two VSL datasets of the relative family topic like father, mother, uncle, aunt, etc. (VSL-WRF) were collected. The first one includes 12 words in Vietnamese sign language which only have small changes between frames. While the second dataset was acquired with the corpus of continuous VSL datasets, and consists of 15 words which have the relative position and orientation of motion gestures and the body parts. Moreover, the data augmentation technique is proposed to gain more information of hand movement and hand position.

This paper is organized as follows. The related work is presented in Section II. In Section III, we describe the overview of methodology including preprocessing step, feature extraction and then recognition stage. Experiments and discussion can be found in Section V. Finally, we conclude by mentioning our contributions and some future works in Section VI.

II. Related Works

Two main approaches of sign language recognition are sensor-based and vision-based one.

Sensor-based approach [1]: Signers wear gloves with sensors when describing words. By using sensors, this approach can remove some outliers from the complex environment, therefore it simplifies the preprocessing stage in sign language recognition system. In addition, the use of gloves makes it able to capture not only the change of shapes, but also the hand movements in video sequence. However, it is not convenient to signers when he/she need to wear gloves

Manuscript received October 14, 2018; revised June 1, 2019.

Anh H. Vo and Van-Huy Pham are with the Faculty of Information Technology, Ton Duc Thang University, Ho Chi Minh City, Vietnam (Corresponding author: Van-Huy Pham; e-mail: vohoanganh@tdtu.edu.vn, phamvanhuy@tdtu.edu.vn).

Bao T. Nguyen is with the Faculty of Information Technology, University of Education and Technology, Ho Chi Minh City, Vietnam (e-mail: baont@hcmute.edu.vn).

during the making gesture process. Moreover, it is not suitable in real world because it requires to bring gloves all time. Duy *et al*. [1] used sensor-based approach to identify 23 characters in Vietnamese Sign Language.

Vision-based approach [2]-[4]: This approach captures the hand gesture or body part gesture in form of static images or sequence images by camera without gloves or sensor devices. The vision-based approach is convenient with signers, thus it more suitable for real world than the sensor-based one. Even although, it also has many challenges like complex background problems, variation in lighting condition, changes of skin color, and the properties and configurations of camera (such as variations of scale, translation, rotation, view and occlusion). In [2]-[4], Microsoft Kinect camera are used to capture RGB-D image instead of RGB image to improve the results. However, Microsoft Kinect camera is hard to integrated into real world applications. Due to that reason, in this study we propose to use RGB image in a video sequence to capture motion features of isolated sign for VSL recognition.

III. Approach

We propose and compare two methods for Vietnamese Sign Language recognition in video consequence: the first one used local descriptors (spatial, scene features) with traditional classification SVM; while the second one based on Deep Learning technique.

A. Preprocessing

From the input video sequence of each word, first some key frames were extracted manually. Then, the original images are converted to HSV-color space to segment skin regions. Due to the difference among signers when collecting data, we also applied the data augmentation technique (rotation and noise adding) to get more images with the aim of gaining more information of hand position. More specifically, the original images are rotated ±5, ±10 degree and added salt/pepper noise with probability of 0.05. After that, all hand skin regions are normalized to eliminate face region and background region. Fig. 1 presents all steps of the pre-processing stage.

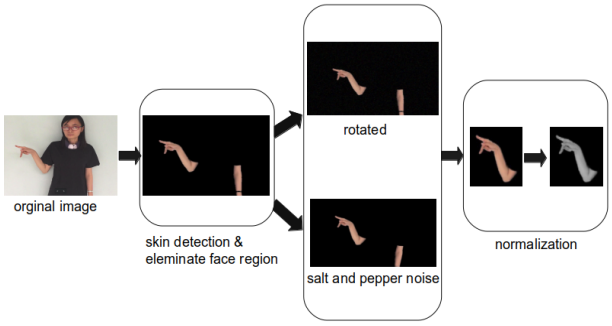

Fig. 1. The preprocessing stage of VSL recognition system. After extracting some key frames, the original images are converted to HSV-color space to segment skin regions. Then data augmentation technique (rotation and noise adding) is used in order to get more information about the hand position.

B. Local Descriptor-Based VSL Recognition

In this approach, we used two local descriptors (spatial feature and scene-based) to capture different information among each word of VSL in a video sequence for recognition.

1) Spatial Features: Spatial features are also called appearance-based ones. There are three appearance-based features as following:

Local Binary Pattern (LBP): LBP operator was initially proposed by Ojala *et al*. [5] to express the texture of the image patches. LBP is tolerant to illumination variations, and simply computing. It calculates a code for each pixel by thresholding its value with that of its neighbor and converts the code into a decimal. Given a pixel i_c in in the image, i_c is the neighbors of i_n

$$LBP_P(x_c, y_c) = \sum_{n=0}^{P-1} s(i_n - i_c)2^n$$

$$s(x) = \begin{cases} 1 & x \geq 0 \\ 0 & x < 0 \end{cases}$$

In this framework, we split hand region into 49 subregions and then applied LBP operator in each subregion. Instead of 256-bin histogram of traditional LBP, we only use 59-bin histogram to reduce dimension for feature vector. Our experiment shows that there is not much difference in recognition result between 256-bin vs 59-bins. As a result, we obtain a 2891dimensional feature vector for each hand region.

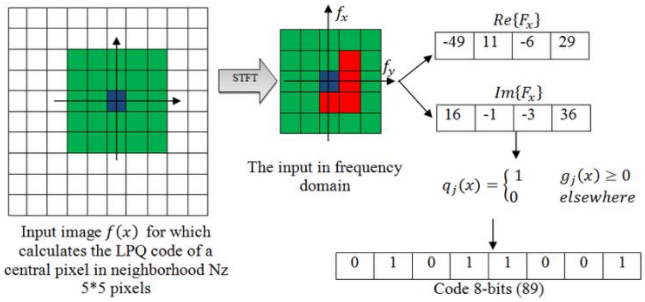

Fig. 2. Local Phase Quantization (LPQ) feature.

Local Phase Quantization (LPQ): LPQ descriptor has been firstly appointed for use in the classification of textures blur [6]. LPQ is constructed to retain an image in the local invariant information to artifacts generated by different forms of blur. This descriptor uses local phase information extracted from a short-term Fourier transform (STFT) over a rectangular $M \times M$ neighborhoods N_x at each pixel position x of the image $f(x)$. Inspired by this idea, we propose the LPQ as an effective method to solve the problem of expressions variations. We applied LPQ operator in 25 sub-regions of hand region, and finally, a histogram of values from all sub-regions are combined into only one feature vector with 6400 dimensions (Fig. 2).

Histogram of Oriented Gradients (HOG): HOG descriptor [7] is one of the appearance descriptors which is able to captures edge or gradient structure of local shape. HOG is invariant to local geometric and photo transformations (translation, rotation). The main idea of HOG is to calculate the distribution of local intensity gradient orientation of a detection window. In this study, we set the size of each block 16 × 16 pixel and 2×2 for each cell.

Finally, we obtain a final feature vector with 8100 dimensions.

2) *Scene-based Feature*: One of the most common scene-based feature is GIST. The GIST descriptor was initially proposed in [8] with the aim of developing a low dimensional representation of the scene, which does not require any form of segmentation. The authors propose a set of perceptual dimensions (naturalness, openness, roughness, expansion, ruggedness) that represent the dominant spatial structure of a scene. Due to that, GIST would able to present the shape of scene image by the relationship between the outlines of the surfaces and their properties and perceptual properties: naturalness, openness, roughness, expansion and ruggedness. It is computed by convolving the image with 32 Gabor filters at 4 scales, 8 orientations, and obtaining 32 feature maps of the same size of the input image. After that each feature map is divided into 16 regions (by a 4×4 grid), and then it averages the feature values within each region. The 16 averaged values of all 32 feature maps are concatenated in a 16×32 = 512 GIST descriptor. The GIST descriptor presents the gradient information (scales and orientations) for different parts of an image, and due to that, it would able to take over many essential information of hand shape for each alphabet or sign word in our case.

After extracting two type of local features *(spatial and scene-based features)*, the classification was inspired by the static image-based approach [9], [10]. Each frame I_i was classified individually and recognition result s was a probability vector $\{p_{i,c_1}, p_{i,c_2}, ..., p_{i,c_n}\}$ where p_{i,c_j}, $(1 \le j \le n)$ is the probability when frame I_i corresponds to one of n classes of Vietnamese signs. Specifically, we used Support Vector Machine to classify sign words of VSL on each frame individually and the recognition result was the normalized sum of probability of all frames to predict sign label in one sequence. Given a training set of labeled examples $\{(x_i, y_i), i = 1, ..., k\}$ where $x \in R^n$ and $y \in \{1, -1\}$, SVM classifies a new test example x basing the following functions:

$$f(x) = \text{sgn}(\sum_{i=1}^{l} \alpha_i \gamma_i K(x_i, x) + b)$$

where α_i are Lagrange multipliers of a dual optimization problem that describes the separating hyperplane; $K(.,.)$ is a kernel function; and b is threshold parameter of hyper-plane. The training sample x_i (with $\alpha_i > 0$) is called support vectors, and SVM would find the hyper-lane that maximizes the distance between the support vectors and the hyper-plane.

C. Deep VSL Recognition (DVSL)

Beside using SVM with different two types of local descriptors for VSL recognition *(spatial and scene-based features)*, we also proposed a Deep Vietnamese Sign Language (DVSL) to recognize a word using the image caption generation models, which inspired from previous works [11], [12]. DVSL system firstly extracts CNN features from f_7 layer of pre-trained VGG16 model [14] with 4096 feature dimensions, then decodes this vector into a sequence learning to predict the sign language of each image by Long Short Term Memory (LSTM) models. Long Short Term Memory (LSTM) [13] is used for activity classification, which predicts class for each image or flow frame of activity. In addition, LSTM model has been also exploited as sequence decoders to learning long-range dependencies. Finally, we use majority voting to compute the final sign for each input video. The details of VSL recognition model based on deep learning (DVSL) is presented in Fig. 3.

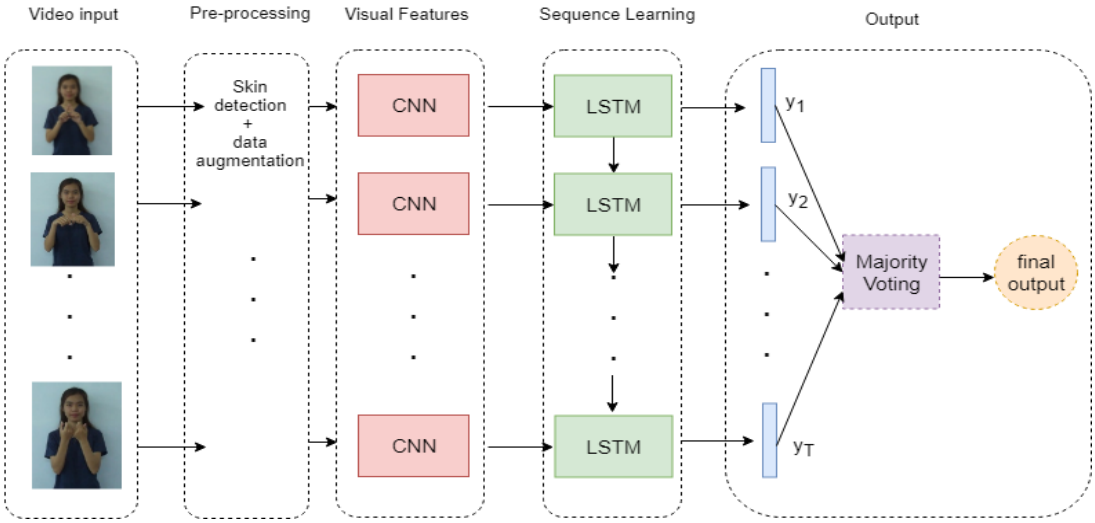

Fig. 3. The pipeline of Deep Vietnamese sign language recognition (DVSL) system including four main steps: pre-processing, visual feature extraction, sequence learning and majority voting. The CNN features from f_7 layer of pre-trained VGG16 model [14] with 4096 feature dimensions were extracted, then fed into a sequence learning to predict the sign language of each image by Long Short Term Memory (LSTM) models. Finally, majority voting was used to compute the final sign for each input video.

IV. Vietnamese Sign Language Dataset

The dataset was acquired in the laboratory environment with white background and the signers wear black and short sleeved shirt. We captured the videos of words with one 2D camera in front of signers, thus it does not contain deep

information as the previous Microsoft Kinect datasets [3], [4]. Moreover, the 2D camera which makes more reasonable and popular in daily life than using gloves with MEMS sensors as in [1]. And finally, we asked human annotators who are teachers of deaf people[1] to evaluate each video to make sure all the information be correct.

The first dataset consists of 12 words which present about the vocabulary of the relative family topic (VSL-WRF) which contains the words which only have small changes between frames. Meanwhile, the second one consists of 15 words, which have the relative position and orientation of motion gestures and the body parts.

After being evaluated from human annotators, who are teachers of deaf people, in total, 480 videos are kept to construct VSL-WRF-01 and 600 videos for VSL-WRF-02 (detail in Table I and Table II). Note that there are some cases the words in Vietnamese language system are different but they are represented by only one word in English vocabulary system. Fig. 4 shows two examples of this case.

TABLE I: DESCRIPTION OF VSL DATASET 01 WITH THE WORDS OF THE RELATIVE FAMILY TOPIC (VSL-WRF-01)

Resolution	1280 × 720
# video	480
# signer	27
# word	12
# video for each word	40
Time for each video	2-5 second

TABLE II: DESCRIPTION OF VSL DATASET 02 WITH THE WORDS OF THE RELATIVE FAMILY TOPIC (VSL-WRF-02)

Resolution	1280 × 720
# video	600
# signer	27
# word	15
# video for each word	40
Time for each video	2-5 second

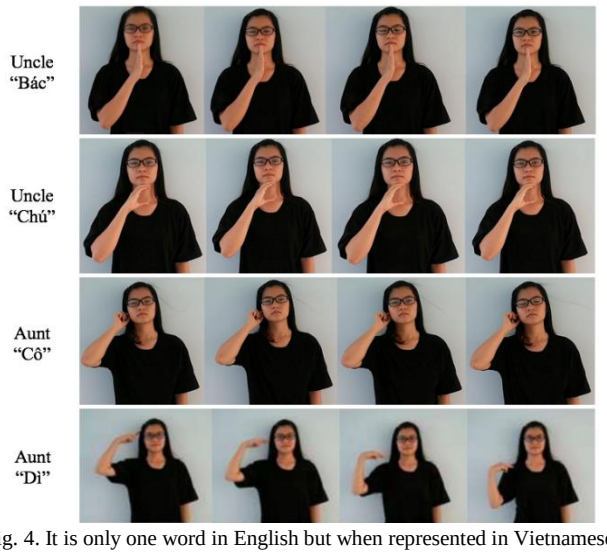

Fig. 4. It is only one word in English but when represented in Vietnamese, there are different meanings and different words.

[1] We would like to thank teachers at The Deaf People Community in Binh Duong province, Vietnam for revising the dataset.

V. EXPERIMENTS

Inspiring from the motion changes between frames in a video sequence of one word, the first dataset (VSL-WRF-01) contains the words which only have a little change between frames (e.g Chú (uncle), Cha (father)), while the second dataset (VSL-WRF-02) includes words, which have the relative position and orientation of the motion gestures and the body parts (e.g Dì (uncle), Trẻ em (children)). Moreover, as describing in the previous section, in order to gain more information about the movement of body parts between frames, we applied the data augmentation consists of rotation transformation (±5, ±10 degree) and the salt/pepper noise (probability of 0.05). As the result we have the original dataset (VSL-WRF-0x) and the extended one (VSL-WRF-0x-EXT).

We divide each VSL dataset into training and testing set with ratio of 9:1 for both local descriptor-based VSL recognition (using SVM), and deep VSL recognition (using deep learning). With local descriptor-based recognition, the cross-validation technique is used to evaluate. The system was performed in Matlab environment with CPU Intel Core (TM) i7, 8GB of RAM.

First of all, we run an experiment to find out the most satisfied kernel for SVM classifier (in consist of linear, polynomial, *rbf* and *sigmoid* kernel) and the highest accuracy with *polynomial* kernel is 85.48% comparing with linear, *rbf* and *sigmod* kernel at 82.26%, 83.87%, 38.71% respectively According to that result, we select *polynomial* kernel for all other experiments based on the local feature descriptor with SVM.

The next experiment is aiming at comparing the performance of local descriptors on the original VSL-WRF dataset and on the extended VSL-WRF-EXT one. Fig. 5 shows the accuracy when running SVM (*polynomial* kernel) with different local descriptors on the extended VSL-WRF-EXT dataset and the original VSL-WRF one. In overall, the accuracy of models on the extended VSL-WRF-EXT dataset by using the traditional data augmentation in preprocessing stage, outperforms than on the original dataset for both VSL-WRF-01 dataset and VSL-WRF-02 dataset. It is proved that data augmentation method is able to improve the accuracy of the model on VSL-WRF dataset because it reduces the variance of signer behaviors. Second, we proposed to use deep learning for VSL recognition (DVSL). More specifically, we extracted CNN features from VGG16 model and considered them as input features for Long Short Term Memory (LSTM) to predict a word for each video.

The results are shown in Table III. It is clear that DVSL model outperforms on VSL-WRF-01-EXT dataset comparing with the local descriptor models which used SVM as a classifier. More details, DVSL model achieved the highest accuracy of 95.83% on VSL-WRF-01- EXT dataset. However, DVSL model is not good as the local descriptor models on VSL-WRF-02-EXT dataset because CNN feature does not capture enough information of motion feature while VSL-WRF-02-EXT contains more gesture motions which involves to the relationship between body parts. This is the weakest point of DVSL recognition system, which should be investigated more in future.

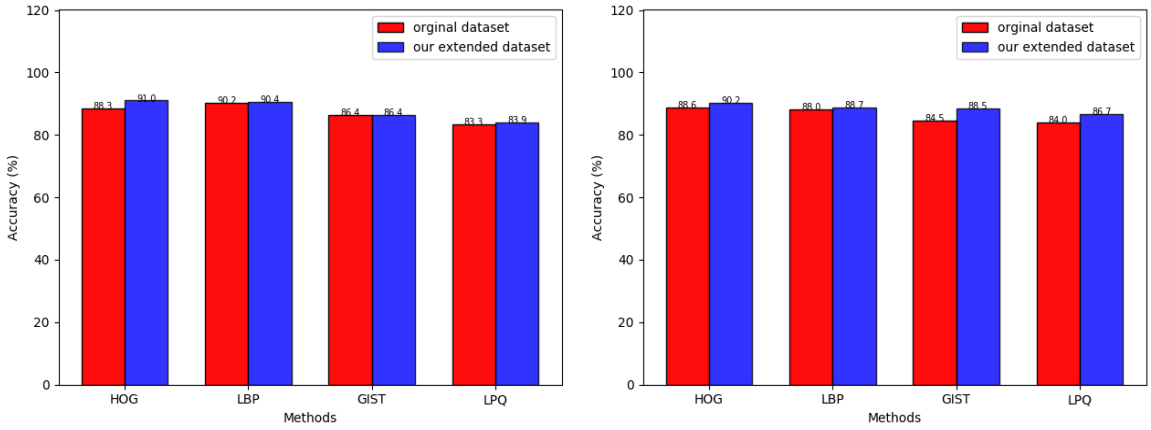

Fig. 5. The accuracy of VSL recognition with different local descriptors on VSL-WRF-01 (left) and VSLWRF-02 dataset (right). In each figure, the red color presents for the original dataset, while the green is marked as the extend one. Different local features (HOG, LBP, GIST and LPQ) was used with polynomial kernel SVM. In overall, the accuracy of models on the extended dataset with the data augmentation technique in preprocessing stage, outperforms than on the original dataset for both VSL-WRF-01 dataset and VSL-WRF-02 dataset. It is proved that data augmentation method is able to improve the accuracy of the model.

TABLE III: VSL RECOGNITION ACCURACY WITH DIFFERENT TYPES OF LOCAL FEATURES FOR SVM; AND DVSL. IN THE FIRST COLUMN, FIVE TOP ROWS ARE LOCAL DESCRIPTORS USED IN SVM, THE LAST ROW IS NEW DVSL METHOD. AND THE NEXT TWO COLUMNS SHOW THE RESULTS ON 02 EXTENDED DATASETS

Method	VSL-WRF-01-EXT	VSL-WRF-01-EXT
HOG	91.04	90.16
LBP	90.42	88.66
LPQ	83.95	88.5
GIST	86.45	88.5
DVSL	95.83	86.67

VI. CONCLUSION

Sign language is one of the tools of communication for those with hearing and vocal disabilities. Sign language recognition hence plays very important role in this regard by capturing the sign language video and then recognizing the sign language accurately. This paper deals with the Vietnamese sign language recognition by using different local features and techniques to recognize the hand gesture.

We acquired and evaluated our proposed methods on two VSL dataset of the relative family topic like father, mother, uncle, aunt, etc. (VSL-WRF) were collected. The first one includes 12 words in Vietnamese sign language contains the words which only have small changes between frames. With the second one, we acquired the corpus of continuous VSL datasets consisting of 15 words, which have the relative position and orientation of motion gestures and the body parts.

According to the results of experiments, we proved the effective of model when applied data augmentation technique in preprocessing step before extracting features. Generally, with words having less motion gesture, the DVSL method tends to outperform SVM in recognizing them. DVSL seems to capture not enough information of the relative position and orientation of the motion gestures and the body parts. Besides, motion-feature also contributes to the accuracy of model, due to its capability of capturing the changes from frames in sequence images. In our future work, we plan to improve DVSL to handle words which are showed by complicated gestures.

ACKNOWLEDGMENT

The authors would like to thank teachers at The Deaf People Community in Binh Duong province, Vietnam. We acknowledgment the support of students in Ton Duc Thang University for collecting the dataset.

REFERENCES

[1] T. D. Bui and L. T. Nguyen, "Recognizing postures in Vietnamese sign language with mems accelerometers," *IEEE Sensors Journal*, vol. 7, issue 5, pp. 707–712, 2005.

[2] D.-H. Vo, H.-H. Huynh, P.-M. Doan and J. Meunier, "Dynamic gesture classification for Vietnamese sign language recognition," *International Journal of Advanced Computer Science and Applications*, vol. 8, no. 3, 2017.

[3] P. Lionel, D. Sander, J. K. Pieter, and S. Benjamin, "Sign language recognition using convolutional neural networks," *Linear Networks and Systems*, Belmont, CA: Wadsworth, pp. 123-135, 1993.

[4] V. D. Hoang, H. H. Hung, N. T. Nguyen, and M. Jean, "Recognizing Vietnamese sign language based on rank matrix and alphabetic rules," in *Proc. 2015 International Conference on Advanced Technologies for Communications (ATC)*, 2015.

[5] T. Ojala, M. Pietikäinen, and D. Harwood, "A comparative study of texture measures with classification based on feature distribution," *Pattern Recognition*, vo. 29, issue 1, pp. 971–987, 1996.

[6] V. Ojansivu and J. Heikkila, "Blur insensitive texture classification using local phase quantization," in *Proc. International Conference on Image and Signal Processing*, pp. 236–243, 2008.

[7] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, pp. 886–893, 2005.

[8] A. Oliva and A. Torralba, "Modeling the shape of the scene: a holistic representation of the spatial envelope," *International Journal of Computer Vision*, vol. 42, no. 3, p. 145–175, 2001.

[9] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on local binary pattern: A comprehensive study," *Image and Vision Computing*, vol. 27, issue 6, pp. 803–816, 2009.

[10] M. Bartlett, G. Littlewort, and F. J. Movellan, "Real time face detection and facial expression recognition: Development and applications to human computer interaction," in *Proc. 2003 Conference on Computer Vision and Pattern Recognition Workshop*, 2003.

[11] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," *Biota Neotrop*, vol. 9, no. 3, 2015.

[12] J. Y. Ng, M. J. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Deep networks for video classification," in *Proc. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[13] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with LSTM," *Neural Computation*, vol. 12, pp. 2451–2471, 1999

[14] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR ABS*, 2014.

Anh H. Vo received the M.S. degree in computer science from University of Sciences, Ho Chi Minh City, Vietnam in 2015, and is currently a PhD candidate. Since 2012, she has been a lecturer and researcher at Information Technology Faculty, Ton Duc Thang University, Vietnam. Her main research interests include image processing, pattern recognition, computer vision, data mining.

Huy V. Pham received the Ph.D in computer science from Ulsan University, South Korea, in 2015, and M.S. degree in Computer Science from University of Sciences, Ho Chi Minh City, Vietnam in 2007. Since 2015, he has been a lecturer and researcher at Information Technology Faculty, Ton Duc Thang University, Vietnam. His main research interests include artificial intelligence, image processing, computer vision.

Bao T. Nguyen currently serves as a lecturer at University of Technology and Education, HCM Vietnam. Before that, he worked at a laboratory for mathematics in imaging at Harvard University. He received his PhD from the Trento-FBK ICT program at Trento University, Italy as part of the Neuroinformatics Lab at the FBK Institution. His research interests are image processing, computer vision, medical imaging and neuroinformatics.