Dynamic Gesture Classification for Vietnamese Sign Language Recognition

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

Phuoc-Mien Doan
Tra Vinh University
Tra Vinh, Vietnam

Jean Meunier
DIRO, University of Montreal
Montreal, Canada

***Abstract*—This paper presents an approach of feature extraction and classification for recognizing continuous dynamic gestures corresponding to Vietnamese Sign Language (VSL). Input data are captured by the depth sensor of a Microsoft Kinect, which is almost not affected by the light of environment. In detail, each gesture is represented by a volume corresponding to a sequence of depth images. The feature extraction stage is performed by dividing such volume into a 3D grid of same-size blocks in which each one is then converted into a scalar value. This step is followed by the process of classification. The well-known method Support Vector Machine (SVM) is employed in this work, and the Hidden Markov Model (HMM) technique is also applied in order to provide a comparison on recognition accuracy. Besides, a dataset of 3000 samples corresponding to 30 dynamic gestures in VSL was created by 5 volunteers. The experiments on this dataset to validate the approach and that shows the promising results with average accuracy up to 95%.**

***Keywords*—Dynamic gesture; feature extraction; depth information; Vietnamese Sign Language**

I. Introduction

In recent decades, computer vision algorithms have been employed in many systems such as surveillance, human-computer interaction, robotic, smart home, and communication [1]. Among various vision-related problems, hand gesture recognition is the one which is widely being studied, in which a suitable approach can give supports for hard-of-hearing people in communication, as well as help to perform interacting between human and computer without touching. According to gesture types, researchers separated such problems into two sub-problems. Methods working on static gestures usually describes local and/or global features of hand shape and posture, while the hand motion is mostly estimated to represent dynamic gestures. Some researchers proposed approaches for dynamic gestures based on static ones.

This work focuses on the problem of recognizing dynamic hand gestures, which is considered to be more difficult than the similar objective on static ones. The input of this system is a sequence of depth images captured by a depth camera (Microsoft Kinect in our experiments) via an infrared (IR) sensor.

The remaining content of this article is organized as follows: some related studies as well as existing limitations are described in Section II; the details of this approach is then presented step by step in Section III; Section IV shows experiments and obtained results; and the conclusion is finally given in Section V.

II. Related Work

In recent studies, the stage of data acquisition was usually performed with the support of sensors mounted on gloves or vision-based systems. Therefore they could be separated into two categories with different pros and cons.

A. Sensor-based approach

In recent studies, many approaches focusing on the problem of recognizing hand gestures have been proposed. For example, the work [2] introduced a method for classifying 6 hand gestures of Korean signal language. The study [3] also built a game controller and performed hand symbol recognition based on the collection of a 3D acceleration sensor and electromechanical biological sensors. In [4], researchers developed a system of classifying symbols in Greek sign language using the energy obtained from bodies with biosensors (EMG) and the assembled data from an acceleration sensor mounted on the arm. For VSL, the researchers in [5] used gloves which are combined with sensors to identify 23 character gestures in the Vietnamese alphabet. These methods focused on the medical field as well as controlling, they thus still have a limited capability to identify the actual sign language gestures. Besides, it is inconvenient for users to carry the data acquisition devices on the body. Several other studies use gloves to capture the change of shapes and movements of the hand. In [6], [7], a glove which was equipped with sensors on all fingers and palm was used to detect movement and bending of the fingers. Besides, such glove also helps to retrieve the location, speed and direction of the hand under a predefined reference system. Some other studies used colored gloves combined with a computer vision system instead of using sensors [8]–[11], in which the fingers are marked by different colors. The use of gloves can support such features simplify the preprocessing step, but brings inconvenience to users when they have to wear gloves during performing sign language.

B. Vision-based approach

The work [12] published a database which consists of hand images performing 26 different gestures, in which each one includes 86 images captured from different directions in the 3D space. In total, a collection of 107328 sample images was obtained. In [13], the authors built 249 samples of 49 word symbols in American Sign Language (ASL). In order to detect and distinguish hand movements, they used 2 different colored-gloves to perform gestures. The testing stage was done in the laboratory environment with a dark background

to enhance the ability of segmenting and distinguishing the two hands. The study [14] also built another database of ASL consisting of 2576 videos corresponding to 14 gestures. The data were recorded by a RGB camera, but the performers must wear long-sleeved shirts in which the color is similar to the background. In [15] presented a database of 19 gestures with 4121022 colored image samples. Although the achieved identification accuracy was about 94%, this work strongly depended on the preprocessing since a simple skin color filter was employed to perform the segmentation. Another approach which also focused on skin color pixels was proposed by [16], in which static gestures of VSL are classified by a neural network. Although such mentioned solutions provided promising experimental results, the preprocessing for hand segmentation was significantly affected by the brightness of environment as well as the background texture.

In order to overcome mentioned limitations, recent researchers performed data acquisition using a depth camera. The study [17] built a system supporting hand gesture recognition, in which the data was collected by a Kinect sensor. An obvious drawback of this study is that characteristics, which describe the hand posture, might be significantly affected be the finger detection result. Therefore the system in [17] has to be improved much for applying on alphabetic gestures. This disadvantage also occurred in [18] where the hand gesture was represented based on detected fingers. Another gesture recognition method working on depth images was presented in [19]. The researchers proposed approaches for both static and dynamic gestures. In experiment, the reported error rate was 5%. However, their main disadvantage is that the features, which were based on hand shape, were limited on the orientation, the template matching thus could perform the classification with low accuracy.

This paper proposes a technique for extracting gesture features and classifying them using SVM. An experiment on HMM was also performed in order to provide a discussion. Since input of this approach is depth image, the mentioned drawbacks of previous works, which are related to sensor and color image, could be overcome.

III. Proposed Method

A. Microsoft Kinect

Camera Kinect is a product manufactured by Microsoft. There are two versions of Kinect which used different techniques for estimating depth information. The proposed system uses a Kinect version 2 which contains many components inside, including a color image acquisition device with a high-resolution up to 1920×1080 pixels, the depth sensor consisting of an infrared emitter and an infrared receiver which provide depth images at 512×424 pixels image at real-time rate. The depth map, which is calculated based on the infrared signals, plays an important role in extracting and recognizing objects since the changes in brightness almost have no impact on the received depth information of Kinect. Nowadays, many researches who are working on the field of sign language recognition [18], [20]–[23] use Kinect to extract objects of interest based on color information, depth as well as joint coordinates provided in the SDK. In the proposed approach, the depth image is captured frame by frame, and each dynamic gesture is represented by a sequence of such images.

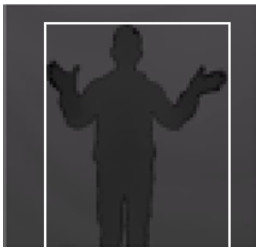

Fig. 1. Bounding box of the object in a depth frame.

B. Preprocessing

As mentioned, a dynamic hand gesture is performed continuously over time, thus the location and movement of the hands and head should be focused. Besides, the beginning and ending time of each gesture is also an important factor. In this study, the position of the hands and the head of performers are focused. The time for executing every gesture is different so that a long-time gesture corresponds to a large number of consecutive images, vice versa. In the preprocessing step, the object of interest, i.e. the performer, in each input image is determined by applying a thresholding technique. In detail, all pixels in the depth image are classified into two groups specifying the object and background (see Fig. 1). In fact, the performer can stand near or far from the camera at an arbitrary distance, as long as in the active area of Kinect. If a predefined threshold is employed to binarize a depth map, the obtained result is not really good since the intensity of object pixels depends on the distance between the object and the Kinect. Therefore, the well-known Otsu thresholding technique [24] is employed to separate the object from background.

After obtaining binary masks corresponding to all depth frame in a sequence, the smallest bounding box that covers all appeared objects in the sequence is estimated. Besides, background pixels in the frame is changed into the intensity of 255, in order to reduce the effect of background on the spatial-temporal volume representing the sequence of depth images. A 5×5 median filter is also used before thresholding for smoothing the depth image as well as noise removal. Although this filter may slightly change the information on each pixel, the thresholding could be performed more effective.

C. Feature extraction

Feature extraction is an important step, which has a great influence on the effectiveness of the process of automatic model training. The featured values are extracted from the sequences of depth image to effectively distinguish between a gesture and another one. An overview of the feature extraction is shown in Fig. 2.

Let d is the number of frames in the sequence, $h \times w$ is size of the smallest bounding box that was mentioned in the previous section, these frames are combined together along the time axis in order to form a three-dimensional array $\mathbf{A}$ with the size of $h \times w \times d$. The goal of this step is to normalize the size of such arrays to a same resolution of $n \times n \times n$ where n is a predefined whole number.

First, the array $\mathbf{A}$ is considered along the depth, i.e. time, direction. Each two-dimensional array $h \times w$ is resized to $n \times n$.

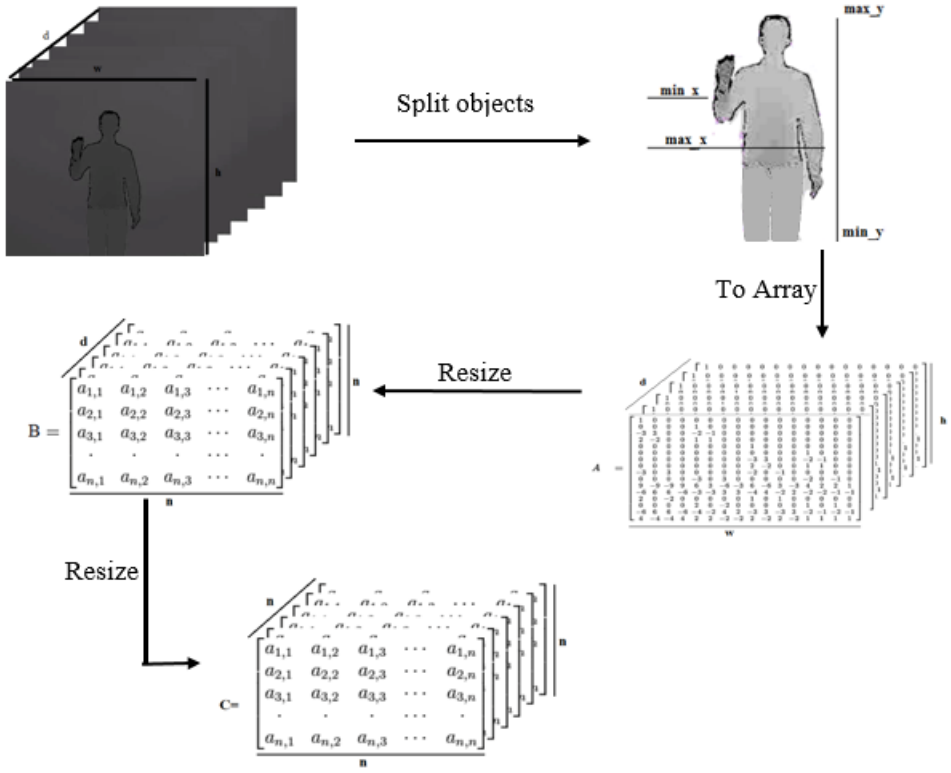

Fig. 2. The diagram of extracting feature for a sequence of depth images corresponding to a gesture.

The results after performing on d arrays are a 3D array **B** with the size of $n \times n \times d$.

$$\mathbf{A}(h, w, d) \rightarrow \mathbf{B}(n, n, d) \tag{1}$$

Next, array **B** is then processed from column 1st to *n*th, in which every 2D array of $n \times d$ is resized to $n \times n$ to obtain a 3D array with the size of $n \times n \times n$.

$$\mathbf{B}(n, n, d) \rightarrow \mathbf{C}(n, n, n) \tag{2}$$

In this work, the process of resizing array is executed by using bicubic interpolation [25] because this method could give smooth results and is used by most of image processing software, digital cameras and printers. In this technique, a new pixel value is calculated based on the mean value of 16 nearest original pixel, i.e. a neighborhood with size of 4×4.

The obtained array of size $n \times n \times n$ is not directly used to represent the feature vector of the corresponding gesture because of the large number of data dimensions and possible noise pixels inside. Therefore, this 3D array is divided into blocks in order to reduce the dimensionality of data as well as the impact of noise since noise level in each block, i.e. the ratio of noise pixels over all elements, is expected to be low. Each block is then converted into a scalar value.

First of all, the elements in the whole array **C** are aligned based on the mean value m. The new value of each element in the array is then recalculated by performing a subtraction on m. The sum of obtained elements in the new array is thus normalized to be zero in order to reduce the impact of the deviation of distance between the performer and Kinect in different sequences of depth images. Next, the new array **C** is divided into a 3D grid with the size of $z \times z \times z$ in which each cell is a block of normalized elements. Finally, each block is represented by a single value corresponding with the average of elements. The result is an array **Z** with the size of z^3, which is much smaller than the original array **C**. Figure 3 illustrates a 1D and 2D representations of array **Z**, in which z is assigned to 4. Each array shape corresponds to an input of the used machine learning models, which support classifying gestures.

D. Support Vector Machine

Support vector machine (SVM) is a supervised learning method which is popularly used for classification and regression analysis. Given a set of training data which was divided into two classes, the SVM algorithm tries to build a binary classification model to separate the input patterns into two defined classes corresponding to the positive class and negative class. Visually, a SVM model builds a super plane to separate input data points in the training set so that the distance from it to nearest points of the two classes is maximized.

In order to create a multi-class support vector machine, one of these two strategies including one-against-all and one-against-one was usually used. In this study, the latter one is selected because of its efficiency and stability. Specifically, a collection of $k(k-1)/2$ binary SVM classifiers, where k is the number of gestures, is built. Each classifier is trained based on data extracted from two classes. In summary, multi-class classification problem is solved by an ensemble of binary SVM classifiers in this work.

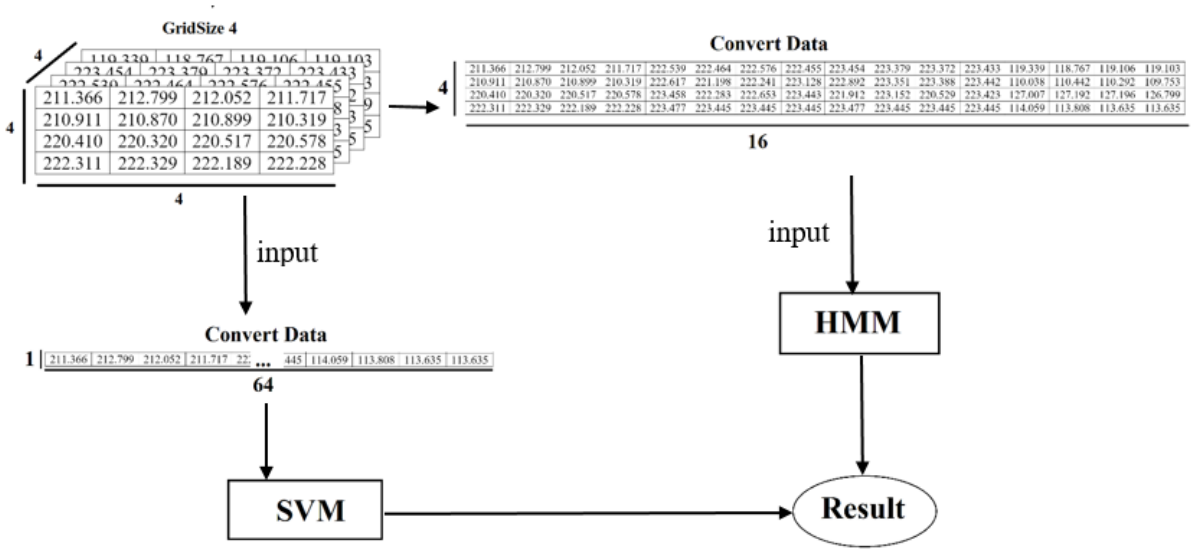

Fig. 3. Low-dimensional array **Z** and different representations for various training models.

E. Hidden Markov Model

Hidden Markov Model (HMM) is a statistical model in which the system being modeled is assumed to be a Markov process with unobserved (hidden) states. This technique was proposed and developed in [26]. With a system involving N states which are numbered in order from 1 to N, HMM is characterized by following elements:

- N is the number of states
- $S = \{s_1, s_2, ..., s_N\}$ is the set of states
- M is the number of distinct observations
- $V = \{v_1, v_2, ..., v_M\}$ is the set of observations
- $A = \{a_{ij}\}$ is the set of transition probabilities
- $a_{ij} = \mathrm{P}(q_{t+1} = s_j | q_t = s_i)$, in which $1 \leq i, j \leq N$, q_t is the actual state at time t
- B is the set of output probability distribution
- π is the initial state distribution, i.e. $\pi_i = \mathrm{P}(q_1 = s_i)$
- $\lambda = (A, B, \pi)$ is the compact notation of a HMM

The objective of HMM related problems includes determining the probability which a sequence of observations is generated from a HMM model in case of testing, and approximating a HMM based on a training set of series of observations in case of training. This technique is selected to perform a comparison because it was employed to solve many problems related to temporal information (e.g. gait assessment [27], [28]).

IV. Experiment

A. Dataset

The dataset that was used in the experiments was built from 5 volunteers with the average distance between each one and the camera is about 2.5m. Each volunteer performed 30 predefined gestures with 20 times for each one, corresponding to 600 sequences of depth images. Each depth image is created at 30 fps with the resolution of 512×424 pixels. Figure 4 illustrates some gestures in the recorded dataset.

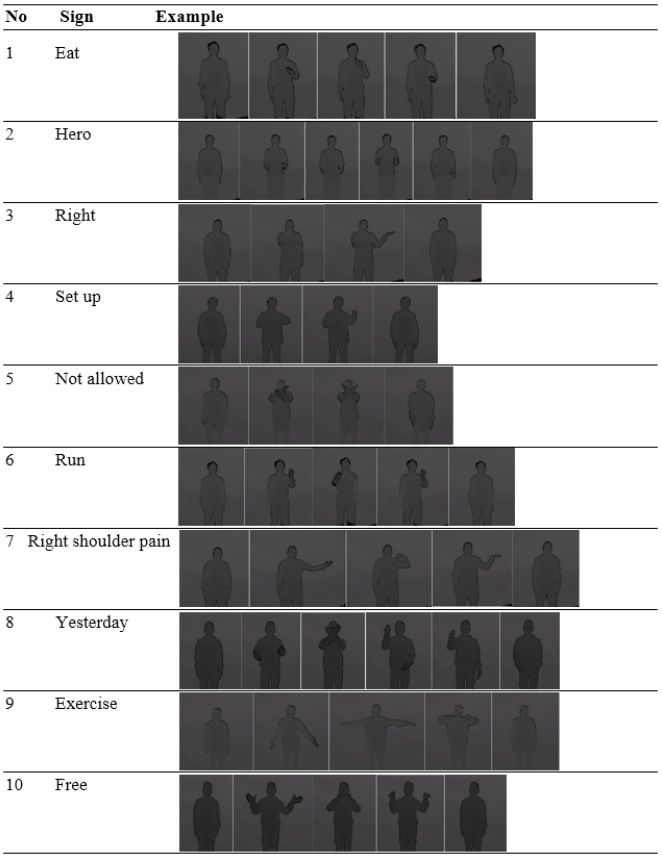

Fig. 4. Examples of some words in our dataset containing depth sequences.

In total 3000 patterns recorded by 5 volunteers, the training

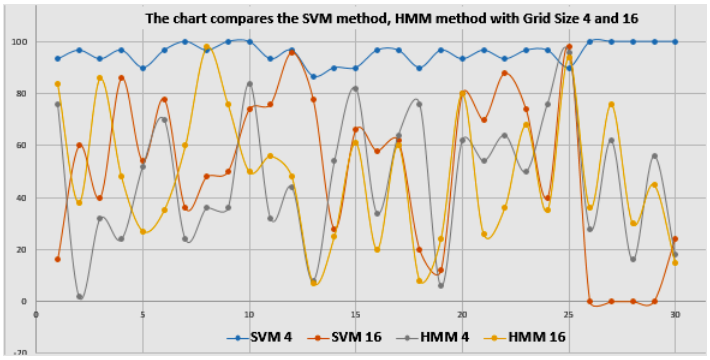

Fig. 5. The result of SVM, HMM when the grid size z is assigned to 4 and 16, respectively.

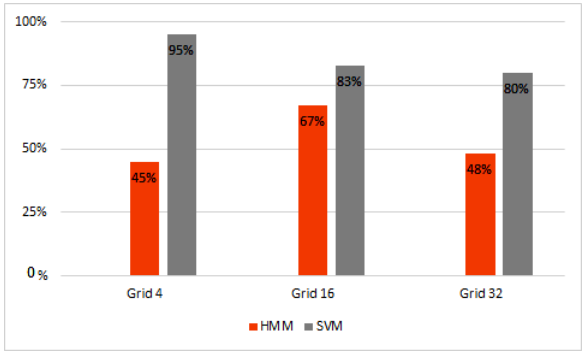

Fig. 6. Accuracy of different machine learning techniques on 30 gestures with different values of z.

set is formed by 1800 samples corresponding to 3 subjects, and the remaining is used for testing stage. This experiment was performed with different grid sizes z as mentioned in section III-C.

B. Experimental results

Figure 5 and 6 show the accuracy with different experiments, using HMM and SVM approach, respectively. It is obvious to see that the SVM method with the 3D grid size $4\times4\times4$ gives the highest result, which is about 95%, compared with other ones.

In order to perform a comparison with a state-of-the-art approach, the method proposed in [29] was applied on the dataset in section IV-A. Differently to the preprocessing in [29] where the stage of hand segmentation is performed on color images, the built-in hand detector of the Kinect was employed to determine the hand silhouette in this experiment. The resulting average accuracy was 93%, which is lower than the best one (95%) of the proposed method. This is because the employed dataset focuses on motion trajectory of the hand while the hand geometry is more difficult to be described compared with the dataset in [29]. The classification accuracies of the proposed approach with different sizes of grid and the method [29] are shown in Table I.

TABLE I. CLASSIFICATION ACCURACIES OF PROPOSED APPROACH AND THE METHOD [29]

Method	**HMM**			**SVM**			**[29]**
Grid size	**4**	**16**	**32**	**4**	**16**	**32**	
Accuracy	0.45	0.67	0.48	**0.95**	0.83	0.80	0.93

V. CONCLUSION

This paper propose an approach for performing feature extraction and recognizing of hand gesture in VSL with the data collected from a Kinect camera. The experimental result on 3000 gestures has confirmed the classification ability of this approach on VSL since the highest accuracy is up to 95%.

With such promising results, we intend to expend the experiment with more words as well as complicated gestures, e.g. combining the hands with motion of other body parts such as head and shoulder, and also to build a system supporting communication between hard-of-hearing people, which focus on the deaf in the context of VSL.

ACKNOWLEDGMENT

The authors would like to thank Trong-Nguyen Nguyen, DIRO, University of Montreal, for his helpful comments. This work was supported by the Polytechnic Computer Vision Group (PCVG), University of Science and Technology, The University of Danang.

REFERENCES

[1] Y. Zhu, Z. Yang, and B. Yuan, "Vision based hand gesture recognition," in *Service Sciences (ICSS), 2013 International Conference on*. IEEE, 2013, pp. 260–265.

[2] K. K. Jung, J. W. Kim, H. K. Lee, S. B. Chung, and K. H. Eom, "Emg pattern classification using spectral estimation and neural network," in *SICE Annual Conference 2007*, Sept 2007, pp. 1108–1111.

[3] X. Zhang, X. Chen, W.-h. Wang, J.-h. Yang, V. Lantz, and K.-q. Wang, "Hand gesture recognition and virtual game control based on 3d accelerometer and emg sensors," in *Proceedings of the 14th International Conference on Intelligent User Interfaces*, ser. IUI '09. New York, NY, USA: ACM, 2009, pp. 401–406. [Online]. Available: http://doi.acm.org/10.1145/1502650.1502708

[4] V. E. Kosmidou and L. J. Hadjileontiadis*, "Sign language recognition using intrinsic-mode sample entropy on semg and accelerometer data," *IEEE Transactions on Biomedical Engineering*, vol. 56, no. 12, pp. 2879–2890, Dec 2009.

[5] T. D. Bui and L. T. Nguyen, "Recognizing postures in vietnamese sign language with mems accelerometers," *IEEE Sensors Journal*, vol. 7, no. 5, pp. 707–712, May 2007.

[6] T. Kuroda, Y. Tabata, A. Goto, H. Ikuta, M. Murakami, and T. Limited, "Consumer price data-glove for sign language recognition," in *In: Proc. of 5th Intl Conf. Disability, Virtual Reality Assoc. Tech*, 2004, pp. 253–258.

[7] S. A. Mehdi and Y. N. Khan, "Sign language recognition using sensor gloves," in *Neural Information Processing, 2002. ICONIP '02. Proceedings of the 9th International Conference on*, vol. 5, Nov 2002, pp. 2204–2206 vol.5.

[8] H. Brashear, V. Henderson, K.-H. Park, H. Hamilton, S. Lee, and T. Starner, "American sign language recognition in game development for deaf children," in *Proceedings of the 8th International ACM SIGACCESS Conference on Computers and Accessibility*, ser. Assets '06. New York, NY, USA: ACM, 2006, pp. 79–86. [Online]. Available: http://doi.acm.org/10.1145/1168987.1169002

[9] K. Assaleh and M. Al-Rousan, "Recognition of arabic sign language alphabet using polynomial classifiers," *EURASIP J. Appl. Signal Process.*, vol. 2005, pp. 2136–2145, Jan. 2005. [Online]. Available: http://dx.doi.org/10.1155/ASP.2005.2136

[10] X. Li, "Gesture recognition based on fuzzy c-means clustering algorithm," *Department Of Computer Science The University Of Tennessee Knoxville*, 2003.

[11] S. Mitra and T. Acharya, "Gesture recognition: A survey," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 37, no. 3, pp. 311–324, May 2007.

[12] V. Athitsos and S. Sclaroff, *Database Indexing Methods for 3D Hand Pose Estimation*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 288–299.

[13] R. Bowden, D. Windridge, T. Kadir, A. Zisserman, and M. Brady, *A Linguistic Feature Vector for the Visual Interpretation of Sign Language*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 390–401.

[14] A. C. Kak, "Purdue rvl-slll asl database for automatic recognition of american sign language," in *Proceedings of the 4th IEEE International Conference on Multimodal Interfaces*, ser. ICMI '02. Washington, DC, USA: IEEE Computer Society, 2002, pp. 167–. [Online]. Available: http://dx.doi.org/10.1109/ICMI.2002.1166987

[15] D. V. Hieu and S. Nitsuwat, "Image preprocessing and trajectory feature extraction based on hidden markov models for sign language recognition," in *2008 Ninth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing*, Aug 2008, pp. 501–506.

[16] T.-N. Nguyen, H.-H. Huynh, and J. Meunier, "Static hand gesture recognition using artificial neural network," *Journal of Image and Graphics*, vol. 1, no. 1, pp. 34–38, 2013.

[17] Z. Ren, J. Yuan, and Z. Zhang, "Robust hand gesture recognition based on finger-earth mover's distance with a commodity depth camera," in *Proceedings of the 19th ACM International Conference on Multimedia*, ser. MM '11. New York, NY, USA: ACM, 2011, pp. 1093–1096. [Online]. Available: http://doi.acm.org/10.1145/2072298.2071946

[18] Y. Li, "Hand gesture recognition using kinect," in *2012 IEEE International Conference on Computer Science and Automation Engineering*, June 2012, pp. 196–199.

[19] X. Liu and K. Fujimura, "Hand gesture recognition using depth data," in *Sixth IEEE International Conference on Automatic Face and Gesture Recognition, 2004. Proceedings.*, May 2004, pp. 529–534.

[20] 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.

[21] Z. Zafrulla, H. Brashear, T. Starner, H. Hamilton, and P. Presti, "American sign language recognition with the kinect," in *Proceedings of the 13th International Conference on Multimodal Interfaces*, ser. ICMI '11. New York, NY, USA: ACM, 2011, pp. 279–286. [Online]. Available: http://doi.acm.org/10.1145/2070481.2070532

[22] S. Lang, M. Block, and R. Rojas, *Sign Language Recognition Using Kinect*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 394–402.

[23] X. Chai, G. Li, Y. Lin, Z. Xu, Y. Tang, X. Chen, and M. Zhou, "Sign language recognition and translation with kinect," in *IEEE Conf. on AFGR*, 2013.

[24] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62–66, Jan 1979.

[25] R. Keys, "Cubic convolution interpolation for digital image processing," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 29, no. 6, pp. 1153–1160, Dec 1981.

[26] A. A. Markov, "An example of statistical investigation in the text of 'Eugene Onyegin' illustrating coupling of 'tests' in chains," in *Proceedings of the Academy of Sciences*, vol. 7 of VI, St. Petersburg, 1913, pp. 153–162.

[27] T.-N. Nguyen, H.-H. Huynh, and J. Meunier, "Skeleton-based abnormal gait detection," *Sensors*, vol. 16, no. 11, 2016. [Online]. Available: http://www.mdpi.com/1424-8220/16/11/1792

[28] L. Tao, A. Paiement, D. Damen, M. Mirmehdi, S. Hannuna, M. Camplani, T. Burghardt, and I. Craddock, "A comparative study of pose representation and dynamics modelling for online motion quality assessment," *Computer vision and image understanding*, vol. 148, pp. 136–152, 2016.

[29] D.-H. Vo, H.-H. Huynh, and T.-N. Nguyen, "Modeling dynamic hand gesture based on geometric features," in *2014 International Conference on Advanced Technologies for Communications (ATC 2014)*, Oct 2014, pp. 471–476.