Hybrid Deep Network and Polar Transformation Features for Static Hand Gesture Recognition in Depth Data

Vo Hoai Viet, Tran Thai Son, Ly Quoc Ngoc
Department of Computer Vision and Robotics, University Of Science, VNU-HCM, Viet Nam

***Abstract*—Static hand gesture recognition is an interesting and challenging problem in computer vision. It is considered a significant component of Human Computer Interaction and it has attracted many research efforts from the computer vision community in recent decades for its high potential applications, such as game interaction and sign language recognition. With the recent advent of the cost-effective Kinect, depth cameras have received a great deal of attention from researchers. It promoted interest within the vision and robotics community for its broad applications. In this paper, we propose the effective hand segmentation from the full depth image that is important step before extracting the features to represent for hand gesture. We also represent the novel hand descriptor explicitly encodes the shape and appearance information from depth maps that are significant characteristics for static hand gestures. We propose hand descriptor based on Polar Transformation coordinate is called Histogram of Polar Transformation (HPT) in order to capture both shape and appearance. Beside a robust hand descriptor, a robust classification model also plays a very important role in the hand recognition model. In order to have a high performance in recognition rate, we propose hybrid model for classification based on Sparse Auto-encoder and Deep Neural Network. We demonstrate large improvements over the state-of-the-art methods on two challenging benchmark datasets are NTU Hand Digits and ASL Finger Spelling and achieve the overall accuracy as 97.7% and 84.58%, respectively. Our experiments show that the proposed method significantly outperforms state-of-the-art techniques.**

Keywords*—Hand Gesture Recognition; Deep Network; Polar Transformation; Depth Data

I. INTRODUCTION

Static hand gesture recognition which is an important component of Human Computer Interaction, has appealed many efforts invested from the research field of computer vision in recent decades for its strong potential in numerous applications, such as game interaction and sign language recognition. Hand gesture is a distinct and significant component of human action and hand gesture recognition since the information hand gestures convey is more sophisticated and linguistic than others. The goal of hand gesture recognition is to automatically analyze ongoing gesture from image. Generally speaking, hand gesture framework contains four main steps namely hand segmentation, feature extraction, gesture representation (gesture descriptor, dimension reduction …) and pattern classification. Though much progress has been made [5, 9, 16, 22, 24, 26], recognizing gesture with a high accuracy remains a challenging task due to the wide range of poses and considerable intra-class variations, e.g., rotation, scaling, viewpoint change and hand articulations. In the previous works, the authors have shown that deriving an effective gesture descriptor from image is a vital step for success of hand gesture recognition. There are two common approaches to extract gesture features [24]: appearance feature-based methods and shape feature-based methods.

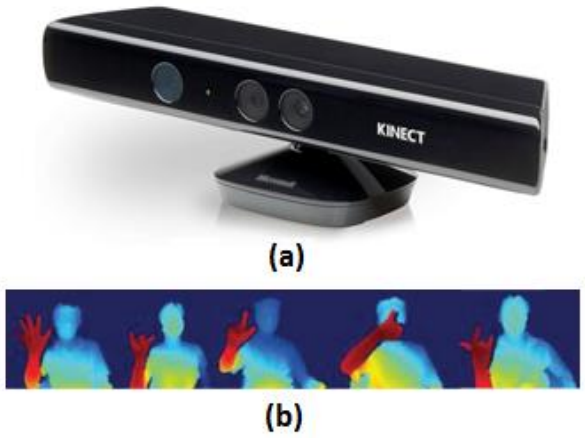

Fig. 1. Microsoft Kinect; b) Some depth images are captured by Kinect

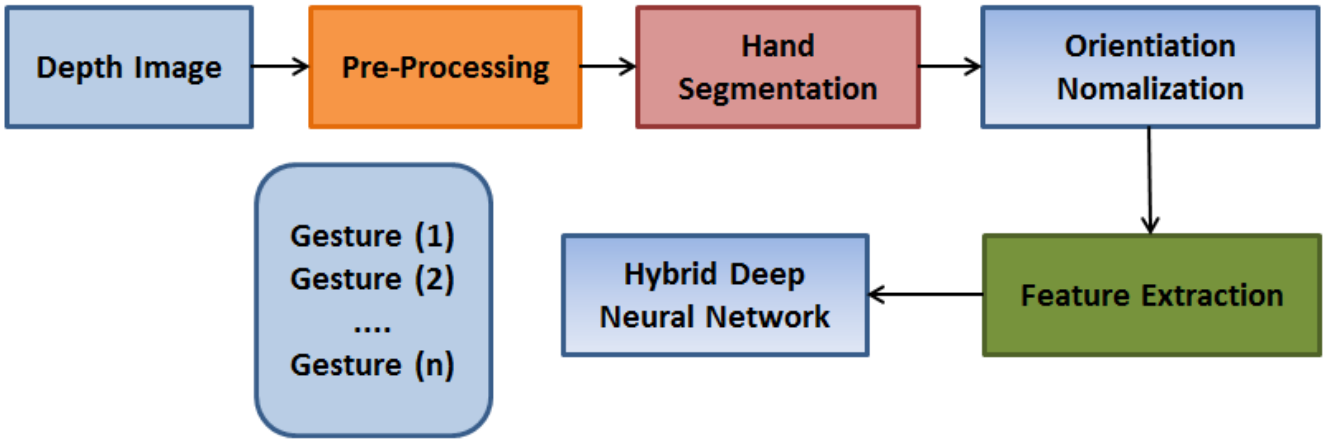

Fig. 2. Our framework of hand gesture recognition system

In all cases, moreover, it is commonly believed that in order to obtain high recognition rate, it is important to select an appropriate set of visual features that usually have to capture the particular properties of a specific domain and the distinctive characteristics of each object class. The most important aspect for any gesture recognition system is to seek an efficient feature representation. An extracted gesture feature can be considered an efficient representation if it could fulfill three criteria: firstly, it minimizes within-class variations while maximizes between-class variations; secondly, it can be easily extracted from the raw video; and thirdly, it can be described in a low-dimensional feature space to ensure computational speed during the classification step. The target of the feature extraction is to find an efficient and effective representation of the gesture which would provide robustness during recognition process.

Depth sensors have been available in many years. Though, they are used in limitation to many applications because of high cost and complexity of operations. However, the recent popularity of new 3D sensors such as Kinect [15] with low cost has alleviated the hardness of the traditional gesture recognition problem, by exploiting the depth data. With its advanced sensing techniques, this technology opens up an opportunity to significantly increase the capabilities of many automated vision-based recognition tasks [14]. And, it promoted interest within the vision and robotics community for its broad applications [14, 21]. In fact, this is a significant motivation for computer scientists to get deep in this research field to find out effective ways to utilize benefits from both the available depth and color information. Compared with conventional color data, depth maps provide several advantages, such as the ability of reflecting pure geometry and shape cues, or insensitive to changes in lighting conditions. Moreover, the range sensor provides 3D structural information of the scene, which offers more discerning information to recover postures and recognize gestures. These properties of depth data provide more natural and discriminative vision cues than color or texture. Furthermore, depth data has been demonstrated its capability to provide more information of object size, shape, and position.

However, feature extraction is just one of the significant steps to create a robust static hand gesture recognition system. Moreover, the classification is also considered as final step for the system and will determine the success of static hand gesture recognition system. This means that a powerful classifier will help to increase recognition rate of the system.

In this paper, we empirically study gesture descriptor based on Polar Transformation and projected views in depth data for gesture recognition. This descriptor combines both shape and appearance properties for gesture representation. And, hybrid deep neural network is deployed to classify gesture descriptors. The contributions of this paper are three-folds: firstly, we propose a method that is simple but effective for hand segmentation and orientation normalization. This method based on distance by horizontal and vertical region of interest. Secondly, we propose robust descriptors for hand gesture based on polar transformation. In this step, the depth map is projected onto three pre-defined orthogonal Cartesian planes and then normalized. After that, we compute the descriptors for each view and concatenate them into a feature vector. This captures appearance of gestures in creating hand descriptor. Finally, we propose hybrid deep network for gesture classification. In this model, we apply Sparse Auto-encoder (SAE) to pre-training for Deep Neural Network (DNN) so that we improve the performance of system.

The rest of this paper is organized as follows: in section II, we review related works. In section III, we introduce our approach for hand gesture recognition. In section IV, we show some results from our experiments and discussion. We conclude in section V.

II. Related Works

In recent years, sign language has been a popular topic in human behavior recognition. Many works have emerged as the American Sign Language recognition [10], the Portuguese Sign Language [19] and the Indian Sign Language [2]. Many hand gesture recognition methods which are based on visual information analysis have been proposed for hand gesture recognition [20]. The traditional approaches focused on using

RGB data. Sebastiean Marcel [19] proposed the approach based on Input-output Hidden Markov Models [23]. Moreover, the state of the art local features are also used by Chieh-Chih Wang et al. [4] and Y. Yao al. [27] such as SIFT [7] and SURF [3] with Adaboost algorithm.

Vision-based systems have been extensively researched and have been recently complemented with 3D sensors as Kinect [15]. Many research works have already used these popular sensors [5, 9, 16, 22]. One approach of recognizing hand gestures has used static depth frame as in [5, 9, 16, 22]. In [16], the authors treated each static depth frame as a regular gray scale image. They used a bank of Gabor filters to capture gradient information and solved the classification problem by random forests. In comparison to [18], the authors focused on a different type of information: contour [28] and a different application area: hand digits recognition. Without using gradients and contours, Hui Li [12] applied HOG [17] from RGB image to depth image and Zhang et al. [5] proposed a new descriptor to model hand gesture using histogram of 3D normals.

Feature extraction is just one of the significant steps to create a robust static hand gesture recognition system. Classification which is the final stage will play a very important role to the success of static hand gesture recognition system. In the classification stage, the traditional methods are used in many researches such as KNN, ANN, SVM, Adaboost and HMM... Although SVM is considered the state of art method for this stage and are used in many researches but deep learning which is an emerging trend in recent years is used in many researches with promising experimental results [6, 8, 13].

In this work, we capture both shape and appearance information to have hand descriptor. We use Polar coordinate system and depth data are projected onto three pre-defined orthogonal Cartesian planes and then they are normalized. Moreover, instead of using ANN, SVM, KNN… for classification, we apply deep network that are used in the recent years with some improvements. We apply hybrid deep network by deep neural network and pre-training it with Sparse Auto-encoder to improve the performance of system.

III. Proposed Method

The proposed hand gesture recognition system is shown in Fig. 2. Firstly, we use bilateral filter to remove noise. Secondly, we segment hand region from full depth image. Thirdly, we estimate the dominant orientation and achieve in-plane rotation invariance. Next, we extract histogram of Polar transformation to describe hand gesture. Finally, Hybrid Deep Network is used to identify the most likely class for input image.

A. Preprocessing

The 3D sensors such as Kinect based on structured light to estimate depth information, it is prone to be affected by noises due to reflection issues. These effects of noise could significantly decrease the overall performance of depth-based gesture recognition framework. Therefore, we firstly relieve the missing data and oultliers from the depth channel. As a result at [16], we adopted the bilateral filter for smoothing the depth channel. The bilateral filter [16] is a combination of a domain kernel, which gives priority to pixels that are close to the target pixel in the image plane, with a range kernel, which gives priority to the pixels which have similar labels as the target pixel. This filter is often useful to preserve edge information based on the range kernel advantages. The edge is important information to represent shape of gesture. The bilateral filter is defined as follows:

$$I^f(x) = \frac{1}{W_p}\sum_{x_i\in\Omega} I(x_i) f_r(\|I(x_i) - I(x)\|) g_s(\|I(x_i) - I(x)\|)$$

$$W_p = \sum_{x_i\in\Omega} f_r(\|I(x_i) - I(x)\|) g_s(\|I(x_i) - I(x)\|)$$

Where I^f is the filtered image, I is the original input image , x are the coordinates of the current pixel to be filtered, Ω is the window centered in x, f_r is the range kernel for smoothing differences in intensities and g_s is the spatial kernel for smoothing differences in coordinates.In this research, f_r and g_s are supposed as Gaussian functions.

B. Hand Segmentation

The hand region extraction can be done in several ways, such as to retrieve a hand joint using a pose estimator or to filter a hand using skin color [21]. In hand detection phase, we do not use color-markers as the traditional methods. This is very important step in hand gesture recognition system. If we failed in this step, the following steps would be negatively affected and the system performance would be decreased. In this paper, the depth image generated by the camera is scaled to the range 0-255. Otsu's thresholding algorithm is applied to the depth histogram to segment the hand from the rest of the image. After thresholding the image, the pixel co-ordinates (x, y) and the corresponding un-scaled depth values (d) of the segmented hand region are extracted.

Because the depth of hand and neighbor region does not have large difference, region of interest will contain noise and unimportant region are captured. So, we propose the method to choose the interest region of hand to extract gesture descriptor and remove unimportant region. S is the interest region after segmentation with Otsu's thresholding algorithm. We calculate two distances and center of S as follows:

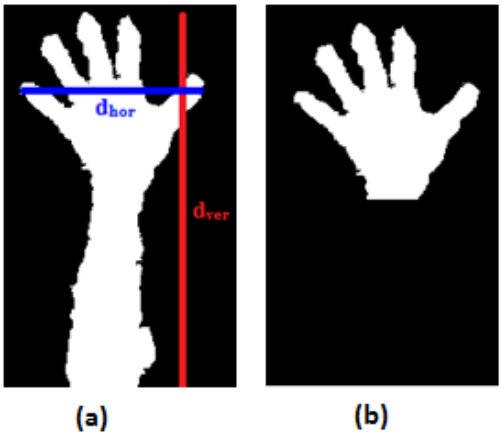

Fig. 3. Illustrate for hand region of interest segmentation: a) Hand region of Otsu's thresholding algorithm; b) Hand region of interest of our proposal

$$d_{ver} = \max(y) - \min(y)$$
$$d_{hor} = \max(x) - \min(x)$$

$$(\mathrm{x_{mean}}, \mathrm{y_{mean}}) = \frac{1}{|S|} \sum_{(\mathrm{x_i}, \mathrm{y_i}) \in S} (\mathrm{x_i}, \mathrm{y_i})$$

Hand region is segmented as follows: If d_{ver} greater than d_{hor} then we remove all the points that have y greater than y_{mean} . Otherwise, we remove all the points that have x less than x_{mean}.

C. Orientation Normalization

A big problem for static hand gesture recognition is the large intra-class variation incurred by hand rotations. The depth maps of the same gesture can significantly vary due to the in-plane rotation. Based on SIFT descriptor [7], we will assign dominant orientation for hand region. In order to estimate the dominant orientation and achieve in-plane rotation invariance, we compute the dominant depth gradient orientation as the normalization employed by SIFT descriptors [7] in 2D images. And we achieve the feature which is robust to the variety of rotation angle, scale, light conditions, viewpoints and noise.

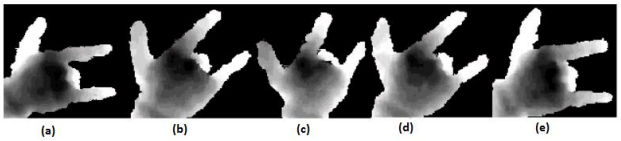

Fig. 4. Examples of one gesture have different orientations

D. Polar Feature Extraction

The shape is important property for hand gesture descriptor. A good descriptor has to contain the shape that can be described by gradient, edge… In this paper, in order to effectively describe the gesture shape, we utilize the polar coordinate system to capture the relative angles and distances between the salient points and the reference point for each gesture. This reference point is defined as the geometric center of the hand gesture and the relative distances are normalized by maximum distance on the support region, which makes the gesture descriptor insensitive to changes in scale of the hand gestures. The Cartesian coordinate system is transformed into the Polar coordinate system through the following equations:

$$r_i = \sqrt{(x_i - x_c)^2 + (y_i - y_c)^2}$$

$$\theta_i = \tan^{-1}\left(\frac{y_i - y_c}{x_i - x_c}\right)$$

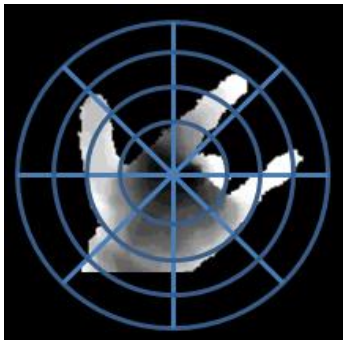

Fig. 5. Illustration of hand gesture descriptor is mapped onto polar coordinate system

where (x_i, y_i) is the coordinate of pixels in the Cartesian coordinate system. (r_i, θ_i) is the radius and the angle in the Polar coordinate system. (x_c, y_c) is the centre of the hand region. The center of the hand region can be calculated by:

$$x_c = \frac{1}{N}\sum_{i=1}^{N} x_i$$
$$y_c = \frac{1}{N}\sum_{i=1}^{N} y_i$$

where N is the total number of pixels.

In this paper, we compute the histogram of Polar coordinate system based on partition polar coordinate space into K cells by uniformly dividing each radius into R parts, and angles into A orientations such that K=A×R. Therefore, feature vector for hand gesture descriptor is K dimensions.

Moreover, in order to increase the discriminative descriptors, the depth map is projected onto three pre-defined orthogonal Cartesian planes and then normalized. After that, we transform each view into Polar coordinate and each Polar coordinate view is quantized by partitioning it into several cells with different radius and angles. This process will help capture the appearance information of hand gestures. The appearance is also importance properties to describe for hand gestures. So, this hand descriptor is extracted containing both shape and appearance information.

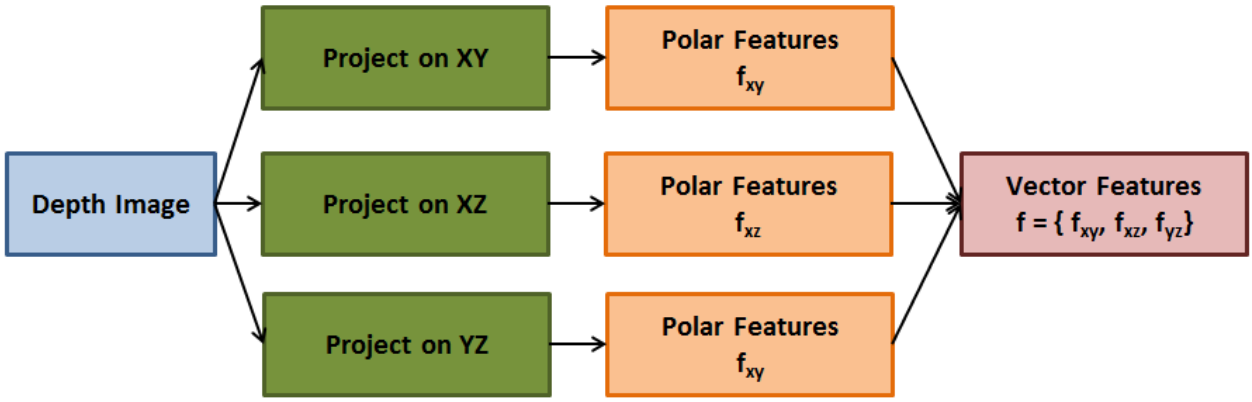

Fig. 6. Process of hand descriptor extraction on three views from hand region

E. Gesture Classification

The classification is the final step for the static hand gesture recognition system. To perform reliable recognition, there is first important problem that the features extracted from the training pattern are detectable that should have more descriptive and distinctive information. Besides, we need a good model for classifying between gestures to have a good recognition rate that accepted. The state of art method for classification is SVM have been used in many researches. However, deep learning which is an emerging trend is used in many researches with promising results in recent years. In this paper, we adopted deep neural network that is a kind of deep learning. Deep Neural Network is a neural network which has three or more hidden layers. In order to train deep neural network, a traditional way to train a deep neural network is an optimization problem by specifying a supervised cost function on the output layer with respect to the desired target. Neural Network is used to a gradient-based optimization algorithm in order to adjust the weights and biases of the network so that its output has low cost on samples in the training set. Unfortunately, deep networks trained in that manner have generally been found to perform worse than neural networks with one or two hidden layers [8, 13]. To overcome this problem, Dumitru Erhan et al. [6], answered the question "Why Does Unsupervised Pre-training Help Deep Learning?". The research indicates that pre-training is a kind of regularization mechanism, by minimizing variance and introducing a bias towards configurations of the parameter space that are useful for unsupervised learning [6, 13]. The greedy layer wise unsupervised strategy provides an initialization procedure, after which the neural network is fine-tuned to the global supervised objective.

The algorithm of the deep network training is decomposed in two steps:

- Step 1: greedily train subsets of the parameters of the network using a layer wise and unsupervised learning criterion, by repeating for each layer.
- Step 2: fine-tune all the parameters of the network using back-propagation and stochastic gradient descent.

In this paper, we adopted Sparse Auto-encoder [1] is unsupervised learning criterion to build deep neural network with 5 layers (3 hidden layers) as Fig. 7.

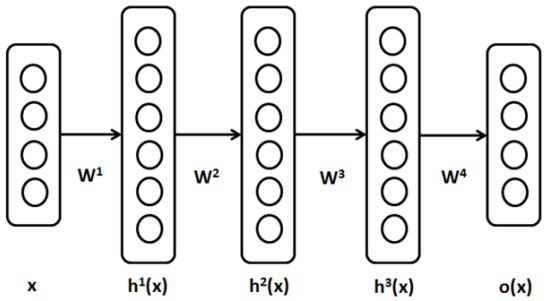

Fig. 7. Illustration of a deep neural network with 5 layers

IV. Exprimental Results

A. Data Set

We evaluate our approach on two benchmark datasets (NTU Hand Digits, ASL Finger Spelling) that we gather from the author's websites.

NTU Hand Digits dataset is the hand gesture dataset with a Kinect sensor. The dataset is collected from 10 subjects, and it contains 10 gestures. Each subject performs 10 different poses for the same gesture. Thus in total our dataset has 10 subject 10 gestures/subject 10 cases/gesture = 1000 cases, each of which consists of a color image and a depth map. Our dataset which is a very challenging real-life dataset is collected in uncontrolled environments. Besides, for each gesture, the subject poses with variations, namely the hand changes in orientation, scale, articulation, etc.

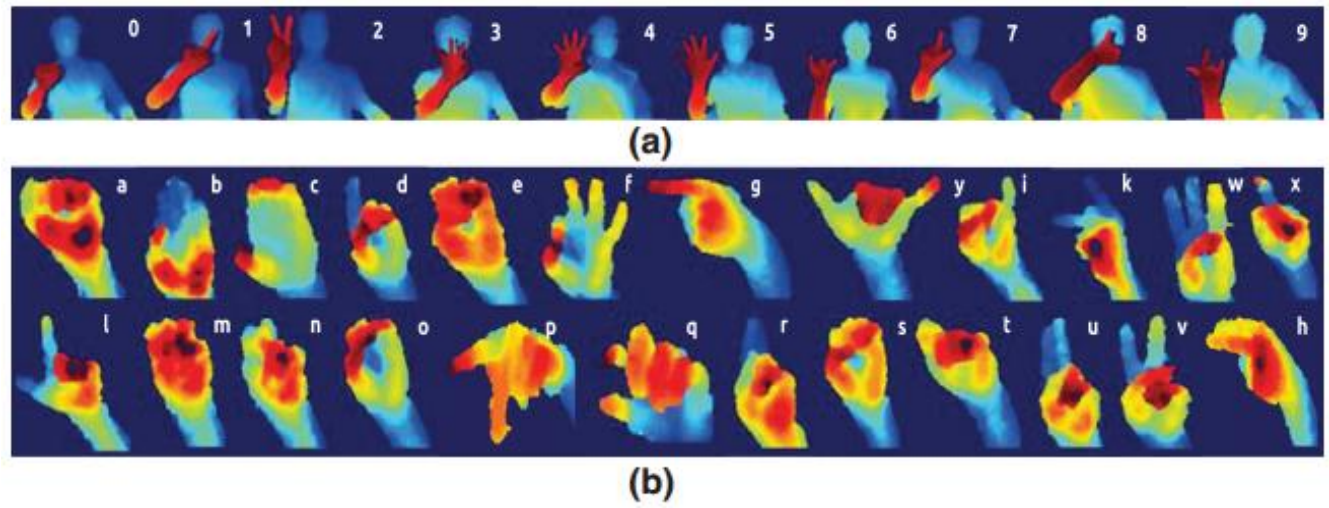

Fig. 8. (a) Some depth images from the NTU hand digits dataset. (b) Some depth images from the ASL finger spelling dataset

The ASL Finger Spelling dataset captures 60,000 hand gestures from 5 subjects. It includes 24 English letters from a to z, but with j and z discarded as these two letters in ASL are dynamic. And the dataset only provides the hand regions after segmentation. So, for the ASL Finger Spelling dataset, we skip the preprocessing step of hand segmentation as describe in section III. The dataset focus on for estimating generalization.

B. Evaluation Framework

In order to have the fair comparison with the other works, we use two experiments to compare with others. Firstly, subject independent test which uses the leave-one-out strategy, i.e., for a dataset with N subjects, N – 1 subject are used for training and the rest one for testing. This process is repeated for every subject and the average accuracy is reported. This test focuses on ability of generalization of approaches. Secondly, subject dependent test where all subjects are used in both training and testing, where the whole dataset is evenly split 50%-50% for training and testing. This test focuses the performance of approaches in the standard test in real-world the same human ability test are the things are learned then they will be tested.

For feature extraction, we adopted the number of orientations (A) and radius parts (R) for polar transformation of hand descriptor based on experiments, and a histogram with A×R bins is obtained for each polar coordinate system. So, the hand descriptor which is extracted for gesture representation from depth image is 3×A×R dimensions from three views are projected.

For classification, we use deep neural network with 5 layers (input layer, 3 hidden layers, and output layer) has the parameters such as the input layer is the number of feature vector, each hidden layer is 200 nodes, the number of output layer is the number of gesture classes in dataset (NTU Hand Digits is 10 nodes and ASL Finger Spelling is 24 nodes), the learning rate is 0.2, and the number of loop is 1000. In order to improve performance, we adopted Auto Sparse-encoder to pre-train deep neural network is proven that will have better than traditional methods without pre-training.

C. Experiment Results

We firstly evaluate our proposed approach on the two benchmark hand gesture datasets. Then we compare our experimental results to the-state-of-the-art methods to prove the effectiveness and robust of the proposed method.

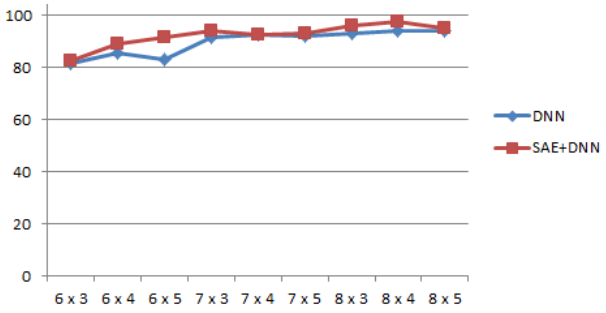

Fig. 9. Accuracies (%) of hand gesture recognition on the NTU Hand Digits dataset under different AxR of gesture descriptors, from 6x3 to 8x5

In this research, we present static hand gesture descriptor as a mixture of these two properties: 1) shape of the hand; 2) appearance of the hand. These properties are extracted frompolar coordinate system and depth image is projected on three views. The relative importance of these elements is based on the nature of the gestures that we aim to recognize. From this experimental results, we argue that no one single category of feature can deal with all kinds of static hand gesture datasets equally well. So, it is quite necessary and useful to combine different categories of features to improve the static hand gesture recognition performance. Moreover, we need a robust classification model to have a good performances. Tables I, and II give our experimental results on NTU Hand Digits, and ASL Finger Spellingdataset on both dependent and independent test. However, the same approach has the different result on the different dataset. This is the different characteristics of these datasets.The ASL Finger Spelling dataset has a larger data scale than NTU Hand Digits about the number of classes and samples.

To study the effect of the two parameters A and R in polar features, we choose the parameters from 6×3 to 8×5 on both NTU and ASL Finger Spelling datasets (see Fig. 9 and 10). The experimental results show that A=8 and R=4 are the best parameters on both datasets.

TABLE I. EXPERIMENTAL RESULTS OF OUR METHODS ON NTU DATASET

Method	**Accuracy**	
	Subj. Indep.	**Subj. Dep.**
Polar + DNN	94.2%	99.33%
Polar + SAE + DNN	97.7%	100%

TABLE II. EXPERIMENTAL RESULTS OF OUR METHODS ON ASL FINGER SPELLING DATASET

Method	**Accuracy**	
	Subj. Indep.	**Subj. Dep.**
Polar + DNN	78.17%	99.8%
Polar + SAE + DNN	84.58%	100%

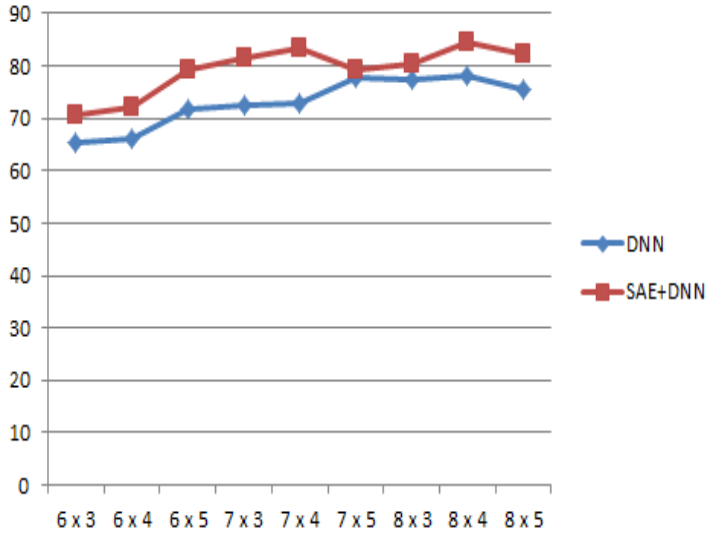

Fig. 10. Accuracies (%) of hand gesture recognition on the ASL Finger Spelling dataset under different AxR of gesture descriptors, from 6x3 to 8x5

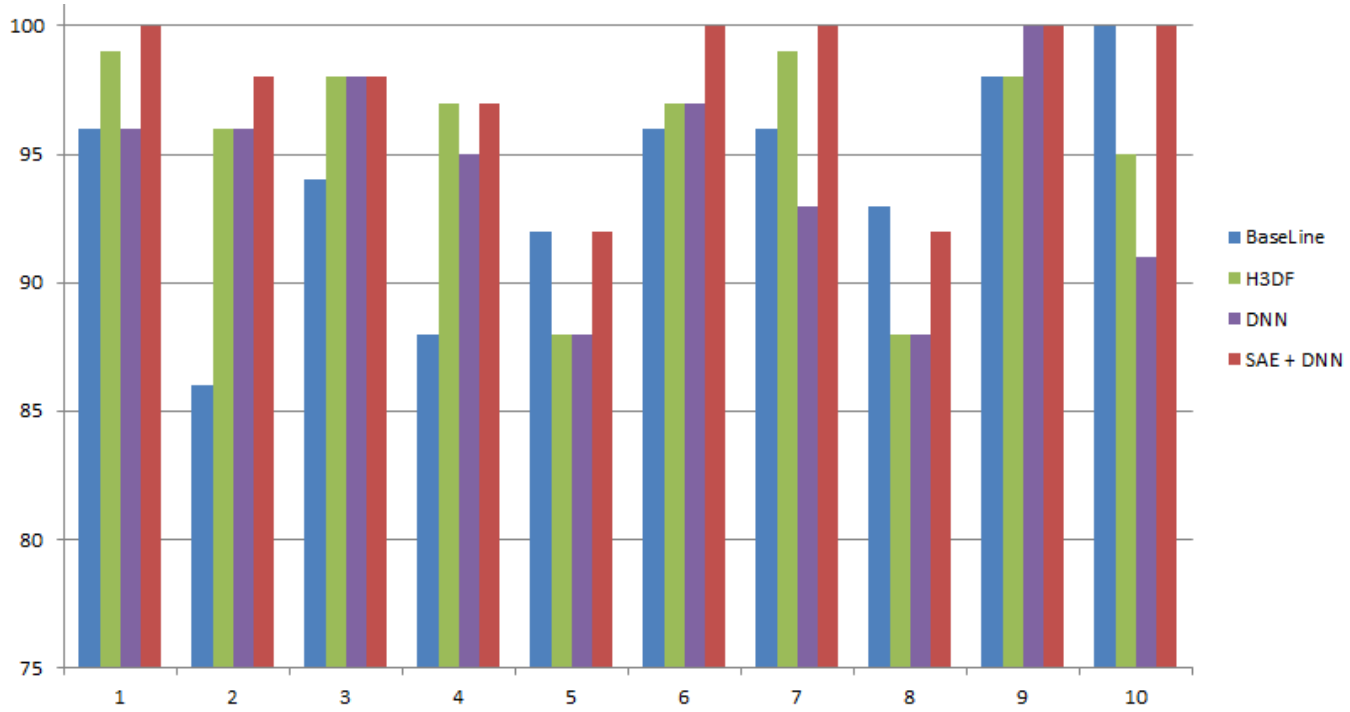

Fig. 11. Comparison of our proposed method with the baseline method in [28] and H3DF [5] on NTU Dataset

Tables III, and IV compare our experimental results with state-of-the-art results on NTU Hand Digits and ASL Finger Spelling dataset respectively. On NTU Hand Digits, our recognition rate is 97.7% on subject independent test and 100% on subject dependent test, more than the current best rate by 2.2% and 0.8%. On ASL Finger Spelling, however, our recognition rate is 84.58% and 100% more than the current best rate by 11.28% and 1.1%. Recognition rate has been improved significantly accuracy in independent test on both datasets, this shows that our approach is effective and stable on cross-dataset with the same configuration. In addition, our approach extracts gesture features based on these algorithms that is rapidly implementation and low computational cost with compact feature vector compare in comparison to existing techniques. From above experimental results, we argue that a successful gesture recognition system not only extract a robust descriptor contains both shape and appearance information but also has a robust classification model. Furthermore, the experimental results at [5, 28] and our approach also show that subject dependent tests significantly outperforms subject-independent tests and is more stable to the changes of locality. This shows a nature of training prolem in real-word that we have to get a base knowlegde about subjects in order to have a good performance when apply into complex and various subjects in real-word.

TABLE III. EXPERIMENTAL RESULTS ON NTU DATASET

Method	**Accuracy**	
	Subj. Indep.	**Subj. Dep.**
Contour Matching [28]	93.9%	N/A
HOG [5]	93.1%	96.4%
H3DF [5]	95.5%	99.2%
Polar + DNN	**94.2%**	**99.33%**
Polar + SAE + DNN	**97.7%**	**100%**

TABLE IV. EXPERIMENTAL RESULTS ON ASL FINGER SPELLING DATASET

Method	**Accuracy**	
	Subj. Indep.	**Subj. Dep.**
Contour Matching [28]	49.0%	N/A
HOG [5]	65.4%	96.0%
H3DF [5]	73.3%	98.9%
Polar + DNN	**78.17%**	**99.8%**
Polar + SAE + DNN	**84.58%**	**100%**

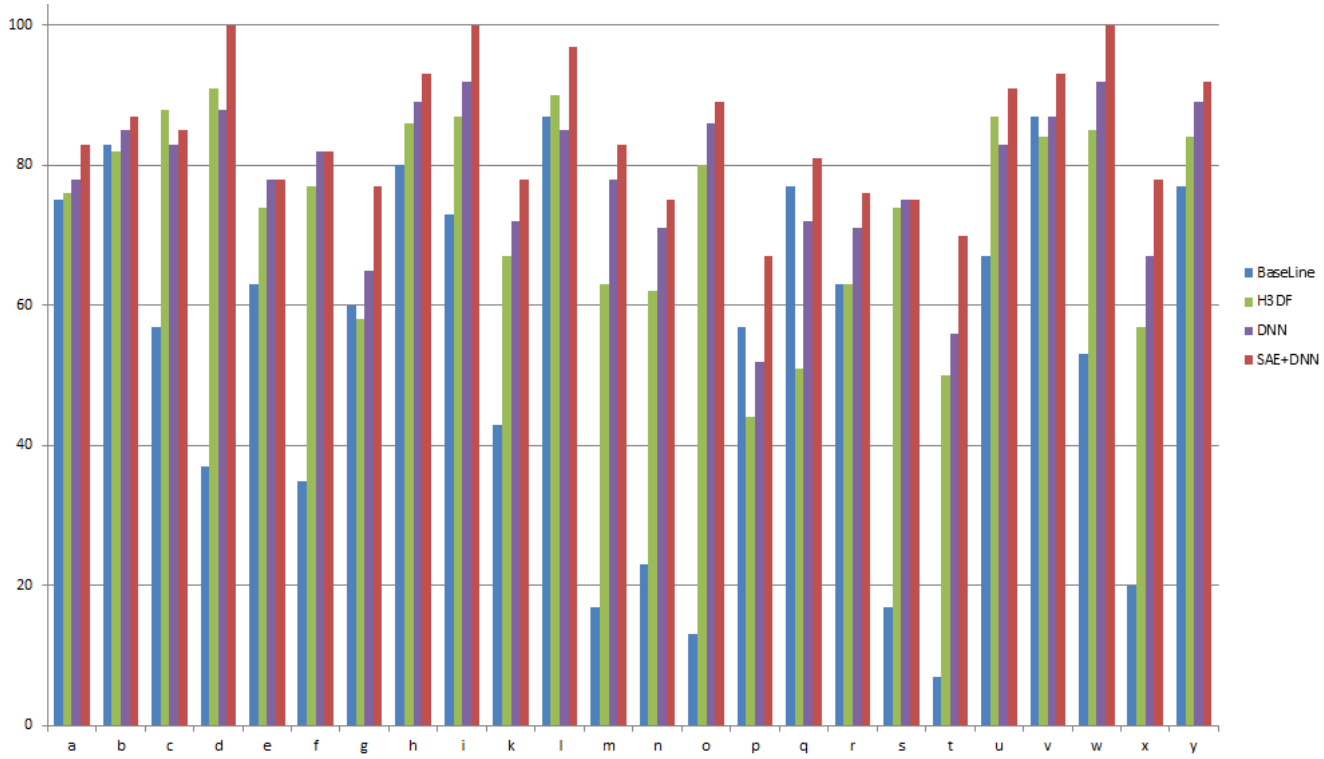

Fig. 12. Comparison of our proposed method with the baseline method in [28] and H3DF [5] on ASL Finger Spelling Dataset

V. Conclusion

In this paper, we represent a novel approach for recognizing static hand gestures based on polar transformation and deep neural network in depth data. Our proposed method consisted of steps as follows: firstly, we use bilateral filter to smooth depth data. Secondly, we segment hand region from full depth and orientation normalization based on depth gradient the same SIFT's idea. Thirdly, we represent a gesture based on using polar transformation for three views are projected from hand region and concatenate them into a feature vector. Our descriptor captures shape and appearance information that are a robust characteristics to distinguish between gestures. Finally, hybrid model is applied for gestures classification to improve the performance of system. Specify, we use deep network with sparse auto encoder for pre-training stage. In this framework, we have exploited the powerfulness of polar transformation in gesture descriptor and effectiveness of deep learning in classification stage. We have evaluated the effectiveness of our proposed on two public hand gesture recognition datasets. Our experimental results achieve superior performance to the state-of-the-art algorithm on NTU Hand Digits and ASL Finger Spelling datasets on overall accuracy as 97.7% and 84.58%, respectively. In addition, our approach is fast and compact in feature descriptor thus it is suitable for real-time hand recognition.

In the future, we will fuse with RGB features to improve the performance of system and extend this descriptor to the temporal domain to capture motion properties in recognizing dynamic hand gesture from depth videos. In addition, we also consider applying feature learning into the system.

Acknowledgment

This research is funded by Vietnam National University, Ho Chi Minh City (VNU-HCM) under grant number *B2014-18-02.*

References

[1] Andrew Ng, CS294A Lecture notes: Sparse Autoencoder, 2011.

[2] A. S. Ghotkar, R. Khatal, S. Khupase, S. Asati, and M. Hadap. Hand gesture recognition for indian sign language. In Proc. Int Computer Communication and Informatics (ICCCI) Conf, pages 1–4, 2012.

[3] Bay, H., Tuytelaars, T., and Van Gool, L .SURF:Speeded Up Robust Features", In Proceedings of the Ninth European Conference on Computer Vision, May, 2006.

[4] Chieh-Chih Wang and Ko-Chih Wang, Hand Posture Recognition Using Adaboost with SIFT for Human Robot Interaction, Recent Progress in Robotics: Viable Robotic Service to Human, pp 317-329, 2008.

[5] C. Zhang, X. Yang and Y. Tian. Histogram of 3D Facet: A Characteristic Descriptor for Hand Gesture Recognition, IEEE International Conference on Automatic Face and Gesture Recognition, 2013.

[6] Dimitru Erchan, Yoshua Bengio, Aaron Courville, Pierre-Antoine Manzagol and Pascal Vincent, Why Does Unsupervised Pre-Training Help Deep Learning?, Journal of Machine Learning Research, 2010.

[7] D. G. Lowe. Distinctive image features from scale-in variant keypoints. International Journal of Computer V ision (IJCV), 60(2):91–110, 2004.

[8] Erhan, D., Manzagol, P.-A., Bengio, Y., Bengio, S., & Vincent, P.(2009b). The difficulty of training deep architectures and the effect of unsupervised pre-training. AISTATS'2009 (pp. 153–160), 2009.

[9] F. Domino, M. Donadeo, P. Zanuttigh, Combining multiple depth based descriptors for hand gesture recognition, Pattern Recognit. Lett. 50 (2014) 101–111, 2014.

[10] F. Ullah. American sign language recognition system for hearing impaired people using cartesian genetic programming. In Proc. 5th Int Automation, Robotics and Applications (ICARA) Conf, pages 96–99, 2011.

[11] H. Liang, J. Yuan, D. Thalmann, Parsing the hand in depth images, in: IEEE Trans. on Multimedia (T-MM), 2014.

[12] Hui Li, Lei Yang, Xiaoyu Wu, Shengmiao Xu, Youwen Wang, "Static Hand Gesture Recognition Based on HOG with Kinect", International Conference on Intelligent Human- Machine Systems and Cybernetics, 2012.

[13] Hugo Larochelle, Yoshua Bengio, Jerome Louradour and Pascal Lamblin, Exploring Strategies for Training Deep Neural Networks, Journal of Machine Learning Research, 2009.

[14] Leandro Cruz, Djalma Lucio, Luiz Velho: Kinect and RGBD Images: Challenges and Applications. SIBGRAPI Tutorials, pp 36-49, 2012.

[15] Microsoft Kinect. http://www.xbox.com/kinect, 2012.

[16] M. Camplani and L. Salgado. Efficient spatio-temporal hole filling strategy for kinect depth maps, A. M. Baskurt and R. Sitnik, Eds., vol. 8290, no. 1. SPIE, p. 82900E, 2012.

[17] N. Dalal, and B. Triggs. Histogram of Orientated Gradients for Human Detection, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 886-893, 2005.

[18] N. Pugeault and R. Bowden. Spelling it out: Real-time asl finger spelling recognition, In ICCV Workshops, 2011.

[19] P. Trindade and J. Lobo. Distributed accelerometers for gesture recognition and visualization. In DoCEIS'11 - Doctoral Conference on Computing, Electrical and Industrial Systems, pages 215–223, Lisbon, Portugal, February, 2011.

[20] Pham Thanh Tung, Ly Quoc Ngoc. Elliptical Density Shape Model for Hand Gesture Recognition. The Fithth International Sysmposium on Information and Communication Technology (SoICT 2014), Hanoi, December 4th -5th , pp.186-191, 2014.

[21] Quang D Tran, Ngoc Q Ly. Sparse Spatio-Temporal Representation of Joint Shape-Motion Cues for Human Action Recognition in Depth Sequences. IEEE-RIVF International Conference on Computing and Communication Technologies (RIVF 2013), Ha noi, Vietnam, November 10th-13th, 2013 (Best Student Running-Up Paper Award), pp.253-258, 2013.

[22] R. Munoz-Salinas, R. Medina-Carnicer, F. Madrid-Cuevas, and A. Carmona-Poyato. Depth silhouettes for gesture recognition, Pattern Recognition Letters, vol. 29, no. 3, pp. 319–329, February, 2008.

[23] Sebastian Marcel, Oliver Bernier, Jean Emmanuel Viallet and Daniel Collobert. Hand Gesture Recognition using Input – Output Hidden Markov Models, Proc. of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition, pp. 456 – 461, 2000.

[24] V. Pavlovic, R. Sharma, and T. Huang. Visual interpretation of hand gestures for human-computer interaction: a review, IEEE Transaction on Pattern Analysis and Machine Intelligence, 19(7), July, 1997.

[25] Vladimir Vezhnevets, Vassili Sazonov and Alla Andreeva. A Survey on Pixel-Based Skin Color Detection Techniques, In Proceedings of the GraphiCon, 2003.

[26] Y. Wu, T. Huang, Vision-based gesture recognition: a review, Gesture-based commun. in hum. comput. interact. 103–115, 1999.

[27] Y, Yao, C.-T. Li and Y. Hu. Hand Posture Recognition Using SURF with Adaptive Boosting, British Machine Vision Conference, Guildford, UK, 3-7 September, 2012.

[28] Z. Ren, J. Yuan, and Z. Zhang. Robust hand gesture recognition based on finger-earth mover's distance with commodity depth camera. In International Conference on ACM Multimedia, 2011.