Graph Spectral Domain Features for Static Hand Gesture Recognition

Basheer Alwaely* and Charith Abhayaratne†
Department of Electronic and Electrical Engineering
The University of Sheffield
Sheffield, S1 3JD, United Kingdom
Email: *b.alwaely@sheffield.ac.uk, †c.abhayaratne@sheffield.ac.uk

***Abstract*—The graph spectral processing is gaining increasing interest in the computer vision society because of its ability to characterize the shape. However, the graph spectral methods are usually high computational cost and one solution to simplify the problem is to automatically divide the graph into several sub-graphs. Therefore, we utilize a graph spectral domain feature representation based on the shape silhouette and we introduce a fully automatic divisive hierarchical clustering method based on the shape skeleton for static hand gesture recognition. In particular, we establish the ability of the Fiedler vector for partitioning 3D shapes. Several rules are applied to achieve a stable graph segmentation. The generated sub-graphs are used for matching purposes. Supporting results based on several datasets demonstrate the performance of the proposed method compared to the state-of-the-art methods by increment 0.3% and 3.8% for two datasets.**

***Index Terms*—Hand gesture recognition, Graph spectral features, Graph partitioning, Fiedler vector, Shape matching.**

I. Introduction

The graph spectral domain provides an appropriate mathematical representation of non-uniform structures such as networks, transportation, map colouring and communications. Therefore, recent years have seen a trend of using the spectral domain to characterize the geometric structure of the data, which helps to match and cluster it. In this paper, we are interested in using the graph spectral domain for static hand gesture recognition for several reasons such as, 1) the global shape of the hand is characterized in the graph eigenvectors and eigenvalues. 2) The ability to reduce the number of nodes and keep the proprieties of the shape intake. 3) Since the graph spectral bases rely on the relative measurements between the nodes, they are invariant to the rotation angle.

A human hand is able to form arbitrary and complex shapes because of the wide range of degree of freedom. A great achievement has been witnessed for hand gesture recognition and the problem has been addressed from different perspectives such as Finger-Earth Mover's Distance (FEMD), dynamic time warping, Superpixel Earth Mover's Distance (SP-EMD) and depth features [1]–[4] respectively. Although extensive research has been carried out based on the silhouette representation, these studies failed to take into account the different directions of the hand especially when the hand is not perpendicular to the camera scene. Such a situation requires interpreting the depth map to understand the skeletal topology of the hand.

In addition, utilizing graphs for hand gesture recognition is limited by using the node domain only to implement point-to-point matching as shown in [5]–[8] respectively. The main issue in the existing studies is the ability to detect the shape in various angles. Also, the purpose of using the depth information in the available literature is only to segment the hand from the background.

In this paper, we use a 3D hand representation to form a hand skeleton and the hand silhouette description for static hand gesture recognition. Our proposed method contains two parts: firstly we use the 2D hand representation mainly to reveal the sequence and locations of the fingers (*i.e.*, the concave and convex along the boundaries). Secondly, we propose a graph spectral partitioning method for matching purposes. We apply a recursive partitioning method using the eigenvector corresponding to the smallest non-zero eigenvalue (Fiedler vector), which divides the graph into two sub-graphs by cutting a small number of edges between components [9], [10]. This vital property can be used to segment and match specific shapes such as the human hands, biological cells, and chemical compositions because these type of shapes have an area of weakness in their structure as in the knuckle where the finger joins the hand. We adapt the conditional connectivity proposed in [11] to strength and weakness the nodes link in the palm and fingers respectively. Therefore, graph can split these parts easily and determine their numbers and locations. In order to overcome the instability of the division, partition procedures must be subject to certain rules. The performance evaluation on three public datasets of static hand gesture demonstrates the strength of our proposed method. The main contributions of this paper are:

- Proposing an automatic hierarchical recursive partitioning method based on the Fiedler vector.
- Utilizing graph spectral domain features for hand gesture recognition.

This paper is structured as follows: the proposed method is presented in Section II. Then, Section III will evaluate and discuss the performance of the proposed method based on different classifiers and parameters. Finally, the work will be concluded in Section IV.

978-9-0827-9703-9/19/$31.00

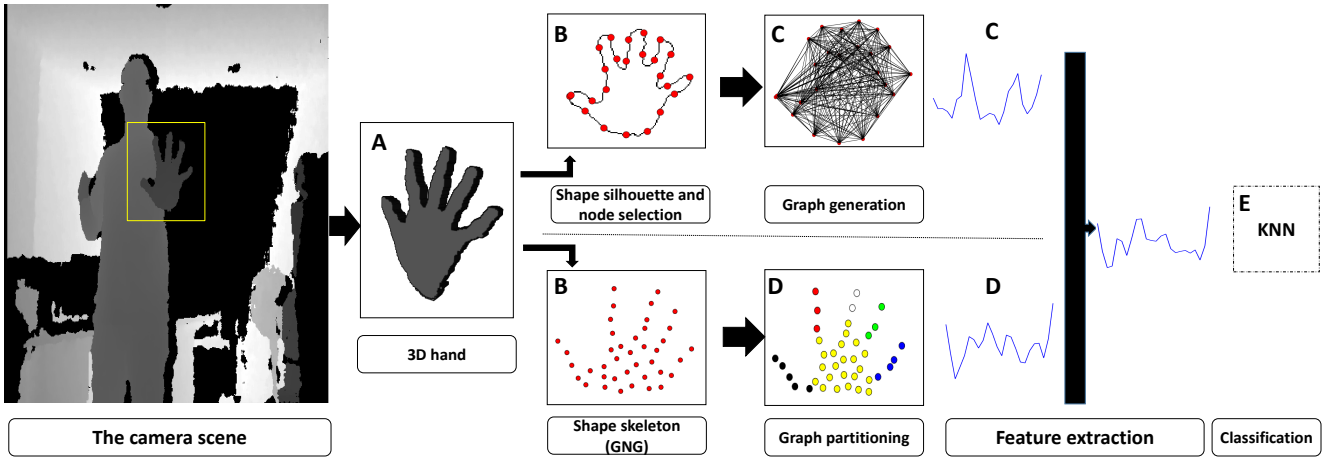

Fig. 1. After extracting the region of interest, the proposed method **consist of** two branches. First, graph generation over the hand silhouette to understand the topology of the hand. Second, extracting the skeleton from the 3D hand representation to be partitioned by the graph spectral domain. A combination of silhouette and skeleton features are classified using KNN. The character in each box refer to the corresponding subsection in Section II.

II. The Proposed Method

The overview of the proposed method is shown in Fig. 1. First, we split the hand's depth map by assuming the hand is the closest object to the camera scene. Second, we formulate a 3D skeleton representation and its 2D silhouette map. Next, we adapt a combination of skeleton and silhouette features using graph spectral domain. At the end, these features are classified using machine learning technique. Details of each step are provided in the next subsections.

A. Depth map

We use a Kinect sensor to capture the depth map at resolution 480×640. First, we search for the closest point to the camera scene at depth $(x, y, d_{nearest})$. Using this point as a centre, we segment the region of interest, which takes the form of cubic of (20, 20, τ). Experimentally, we set $\tau = 80mm$ as a depth range to segment the hand as shown in Fig. 1.

B. Skeleton and silhouette representation

The 2D silhouette representation is formed using an edge detector filter on top of the segmented area as shown in Fig. 1. The resulting 2D closed path consists of random pixels, which are used to generate a new set of pixels (N_1) to form a new down-sampled path, $\hat{P}$, as follows (as in the top sub figure of Fig. 1):

$$\hat{P}(k) = P\left(\left\{\frac{nk}{N_1}\right\}\right), \tag{1}$$

where $k = 0, 1, ..., N_1 - 1$ is the new node index and $\{\}$ is the rounding to the nearest integer.

In order to generate a 3D skeleton with a fixed number of nodes, we use an unsupervised algorithm known us Growing Neural Gas (GNG) [12]. The input data of the GNG is the segmented depth region. Based on the Euclidean distance between pixels, GNG gradually produces new nodes (N_2) inside the shape. By the end of the training, GNG sufficiently covers the area inside the shape as shown in Fig. 1.

C. Silhouette based graph generation and feature extraction

Undirected graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}, \mathbf{A}\}$, where $\mathcal{V}$ is the set of N_1 vertices (defined by the nodes in $\hat{P}$), $\mathcal{E}$ is the set of edges and $\mathbf{A}$ is the adjacency matrix with edge weights. We consider $\mathcal{G}$ as a fully connected graph, which means each vertex has $(N_1 - 1)$ connected edges. We define the weight, $\mathbf{A}_{i,j}$ corresponding to an edge, $\mathcal{E}_{i,j}$ connecting vertices i and j is as follows:

$$\mathbf{A}_{i,j} = \frac{|\mathcal{E}_{i,j}|}{\frac{1}{N_1}\sum_{i=0}^{N_1-1}\sum_{j=0}^{N_1-1}|e_{(i,j)}|}, \tag{2}$$

which is the Euclidean distance $\mathcal{E}_{(i,j)}$ between the vertices, i and j, normalized with the average edge length for a node.

The non-normalized graph Laplacian matrix, $\mathbf{L}$, is defined as

$$\mathbf{L} = \mathbf{D} - \mathbf{A}, \tag{3}$$

where $\mathbf{D}$ is the diagonal matrix of vertex degrees, whose diagonal components are computed as follows:

$$\mathbf{D}_{(i,i)} = \sum_{j=0}^{N-1} \mathbf{A}_{(i,j)}, \qquad i = 0, 1, ..., N_1 - 1. \tag{4}$$

As the shapes form non-regular graphs, we consider the symmetric normalized Laplacian matrix, ($\mathcal{L}$), computed as follows:

$$\mathcal{L} = \mathbf{D}^{-\frac{1}{2}}\mathbf{L}\mathbf{D}^{-\frac{1}{2}}. \tag{5}$$

A complete set of orthonormal eigenvectors $\mathbf{U}$ of $\mathcal{L}$ and their associated real eigenvalues λ_ℓ for $\ell = 0, ..., N_1 - 1$ are calculated.

The graph eigenvectors carry a notion of frequency: low frequencies and high frequencies according to their associated eigenvalues and it is proportional to the degree of the vertices [13]. This concept allows to detect the concave and convex of the fingers along the boundary can be detected by the graph bases. Therefore, we test individual eigenvector for the

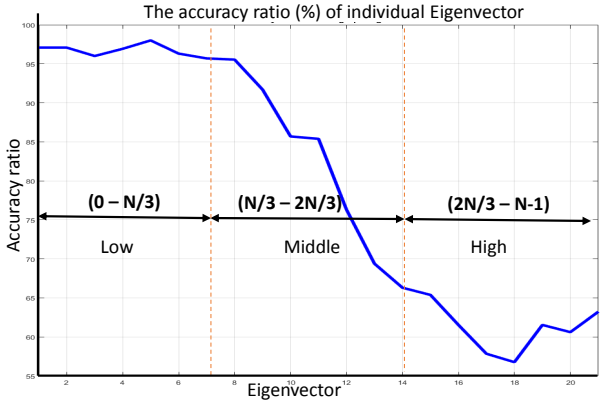

Fig. 2. The accuracy rate of individual eigenvector.

purpose of matching. Fig. 2 shows the accuracy rate of each eigenvector and it is clear that the matching score is high in the low frequencies. We therefore divide the bases into three areas: low, middle and high frequencies.

In order to provide robust features to detect different hand gestures, we propose using the graph frequency where the input signal $\mathcal{S}_{(N_1,3)}$ is a matrix containing (x, y, r), and $r = \sqrt{x_i^2 + y_i^2} \quad i = 0, \ldots, N_1 - 1$. Therefore, the graph frequency response is calculated as follows:

$$\mathrm{F}_{(N_1,3)} = \mathbf{U}_{(N_1,N_1)}\mathcal{S}_{(N_1,3)}. \tag{6}$$

Since we are only interested in the first third of the eigenvectors as shown in Fig. 2, a threshold is applied on F to remove the noise (*i.e.*, middle and high frequencies). The total length of the features will be:
$\mathrm{F}_{(N_1,3)} \xrightarrow{\text{threshold}} \mathrm{F}_{(N_1/3,3)} \xrightarrow{\text{concatenating}} \mathrm{F}_{(1,N_1)}$.

D. Skeleton based partitioning and graph features

Using Fiedler vector for partitioning is not a new concept in graph partitioning because it provides the minimum cutting ratio according to the optimization formulae (7) [9].

$$\lambda_k = min\frac{u^T\mathbf{L}u}{u^Tu}. \tag{7}$$

However, we adapt new rules to achieve a fully stable recursive hierarchical partitioning, which leads to automatically identify the meaningful parts of the structure. In other words, there is no need for human intention to determine the number of required clusters.

For graph partitioning, the main differences in terms of graph generation is that we use the non-normalized Laplacian version (3) and conditional connectivity [11]. Conditional connectivity means using the smallest distance to link all nodes as one group. As a result, the limbs node always has one connected node and the other nodes in fingers have weak connectivities compared to the nodes in the palm area. This type of connectivity makes finger segmentation an easy task to be implemented. Repeating the procedure will end up with efficient segmentation quality. In this paper, the process is repeated five times because we assume that the human hand has maximum five fingers. The segmentation process is subject to certain conditions, which are:

1) The minimum number of nodes in each group = 2.

$$n_{g_1} \geq 2 \quad \& \quad n_{g_2} \geq 2 \tag{8}$$

2) In order to split two graphs (g_1, g_2), the difference in number of nodes between them $> N_2/3$, where N_2 the total number of nodes in the skeleton representation.

$$|n_{g_1} - n_{g_2}| > N_2/3 \tag{9}$$

where n_{g_1} and n_{g_2} are the number of nodes in the new generated sub-graphs. These rules are applied to avoid fragmentation in the palm area, where there are no extended fingers. Also, the segmentation process will be neglected inside the fingers. Fig. 3 shows five levels of segmentation with its corresponding number of nodes at each level. From the number of nodes at the bottom of Fig. 3, we can see the failure of the partitioning process at the sixth level and stop at the fifth level. This is because the proposed rules stop the division process. At the end, we compute a set of features including:

1) Number of clusters
2) Number of nodes that connected to only one node.
3) Fiedler value at each node by the end of segmentation.

The feature length = $N_2 + 2$.

The total feature length = $N_1 + N_2 + 2$.

E. Machine learning

Several experiments are contacted to select the optimal classifier including testing Nearest Neighbour (KNN), Multi class support vector machine with cubic kernal (CSVM), Classification tree (CT), Discriminative Analysis (DA), Neural network (NN) as will be shown in Section III.

III. Performance Evaluation

All the experiments were implemented using MATLAB R2018b on a PC with Intel processor, CPU@3.6GHz and RAM 16GB. 10-fold cross validation scheme is used to train and test all the datasets. In order to reduce the complexity, few numbers of nodes are used to generate the silhouette $N_1 = 24$ and $N_2 = 50$ to form the skeleton. The datasets, which were used to evaluate the proposed method, includes:

1) **d1**: NTU dataset [1] contains 10 subjects $\times$ 10 hand gestures $\times$ 10 different orientations = 1000 colour and its corresponding depth images. NTU dataset includes the subject poses with various hand orientation, scale and articulation.
2) **d2**: The second dataset [14] contains 120 samples for 11 classes, which are implemented by 4 people. The dataset provides RGB images and its corresponding depth image. It also provides a confidence depth map for each sample.
3) **d3**: The third dataset [3] contains 100 samples for 10 classes, which are implemented by 5 people. The dataset provides RGB images and its corresponding

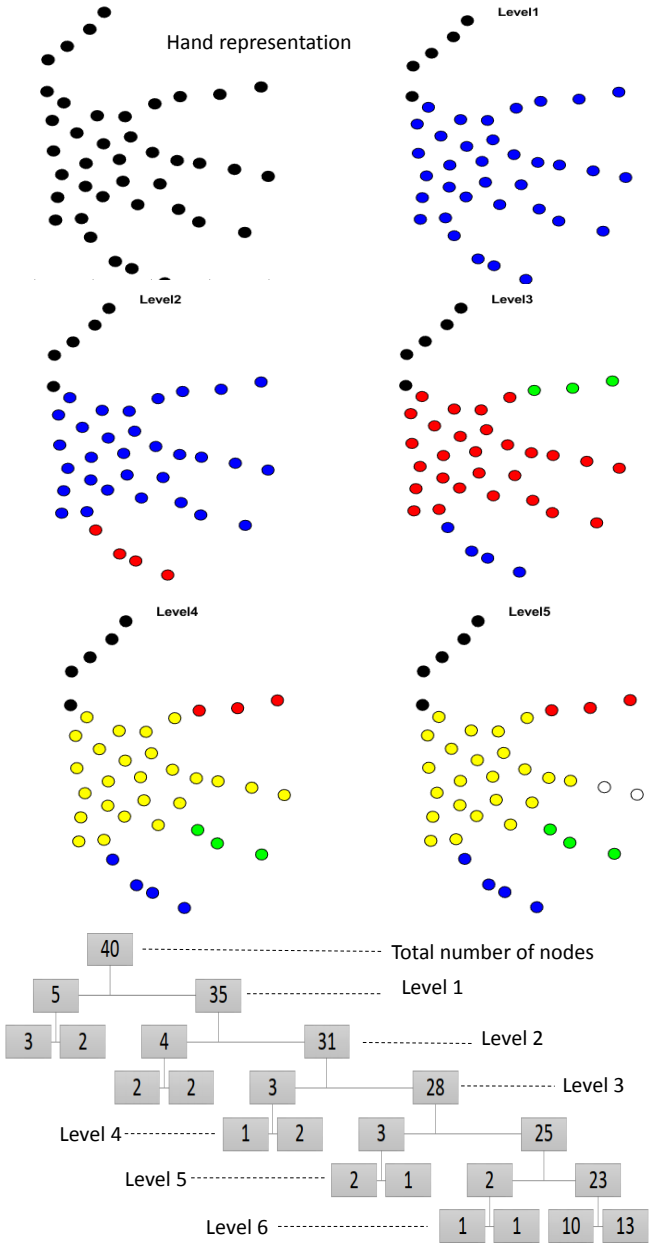

Fig. 3. Segmentation procedure of our proposed graph partitioning with its corresponding number of nodes at each level.

depth image. Only the depth information of each gesture is used for evaluation in this paper.

Initially, different classifiers are tested as shown in Table I. KNN, NN and CSVM show the highest accuracy rate compared to other classifiers. We select KNN with K=1 to evaluate our method because it is fast and accurate. The proposed features demonstrate a high recognition rate for **d1**, **d2** and **d3** by mean accuracy rate achieved to 99.7%, 93.7 and 99.4% respectively as shown in the confusion matrices Fig. 4. From the confusion matrices, we note that the error usually occurs with gestures, which have the same number

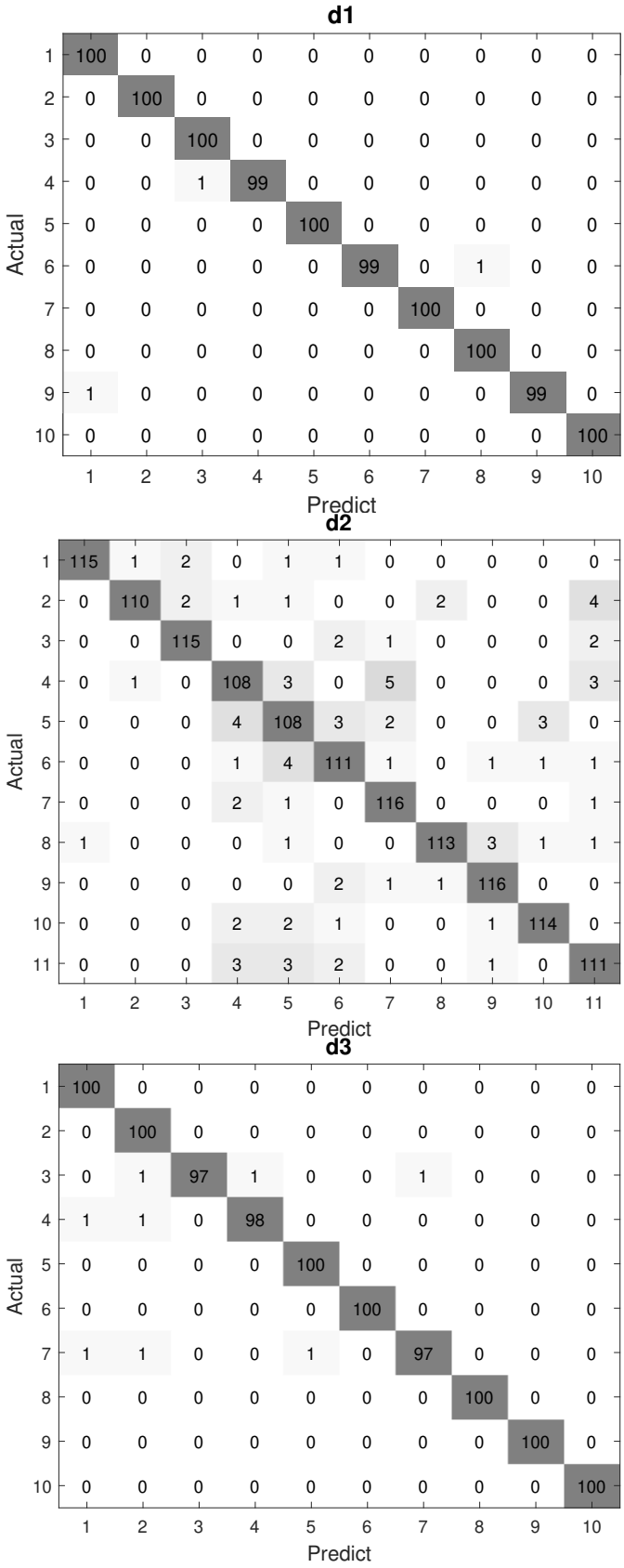

Fig. 4. The confusion matrices of **d1**, **d2** and **d3**.

TABLE I
THE ACCURACY RATES (%) OF DIFFERENT CLASSIFIERS FOR THE THREE DATASETS.

	KNN	CSVM	CT	DA	NN
d1	99.7	99.1	91.3	98.8	99.4
d2	93.7	91.4	86.8	90.1	94.3
d3	99.4	95	88.1	95.1	99.6

TABLE II
THE ACCURACY RATES (%) OF THE SILHOUETTE, SKELETON AND COMBINATION OF SILHOUETTE AND SKELETON REPRESENTATION.

	Silhouette	Skeleton	Both
d1	**99.2**	**75**	**99.7**
d2	**92.9**	**53**	**93.7**
d3	**97.8**	**71**	**99.4**

TABLE III
THE AVERAGE TIME TO PERFORM DIFFERENT STEPS OF THE PROPOSED METHOD.

Step	Performance average time (ms)
Hand segmentation	119.977
GNG	3175.127
Partitioning	177.427
Graph generation (silhouette)	0.548
Feature extraction	0.104
Classification	1.437
Full time system	$\approx$ 3.4 seconds

TABLE IV
COMPARISON OF PROPOSED METHOD WITH OTHER EXISTING METHODS IN TERMS OF THE ACCURACY RATES (%).

	Proposed method	Existing methods
d1	**99.7**	**100** [4]
d2	**93.7**	**89.9** [14]
d3	**99.4**	**99.1** [3]

of extended fingers. However, gestures, which have the same number of extended fingers, can be distinguished by silhouette representation. In order to provide an idea about the accuracy rate of both procedures, Table II shows a detail accuracy of the skeleton and silhouette representation. We can see that the silhouette-based graph features provide an efficient matching score compared to the skeleton-based representation.

The average processing time for the individual step is shown in Table III. Except for GNG training time, all other actions are in real time. The total required time is only above three seconds and can be improved with other alternative programs. The training time of the GNG depends on the several parameters such as the required number of nodes, number of iterations, etc.

To evaluate our proposed method performance, we compare the proposed method with the best existing matching score of the three datasets. Table IV shows that our method performs better than the existing works in **d2** and **d3**.

IV. CONCLUSIONS

This paper has presented a novel method for static hand gesture recognition based on the graph spectral feature representation. Both silhouette and skeleton maps have been analysed for matching purposes. The global hand shape details have been detected using the graph based feature of the hand silhouette. The fingers have been segmented from the palm using an automatic hierarchical recursive partitioning method based on the Fiedler vector. A combination of spectral features were classified using KNN. The performance evaluation shows that the proposed method exceed the state-of-the-art performance by 0.3% and 3.8% for two datasets. In the future, we are planning to extend our proposed method for general 3D shape recognition.

REFERENCES

[1] Z. Ren, J. Yuan, J. Meng, and Z. Zhang, "Robust part-based hand gesture recognition using Kinect sensor," *IEEE Transactions on Multimedia*, vol. 15, no. 5, pp. 1110–1120, Aug 2013.

[2] G. Plouffe and A. M. Cretu, "Static and dynamic hand gesture recognition in depth data using dynamic time warping," *Transactions on Instrumentation and Measurement*, vol. 65, no. 2, pp. 305–316, Feb 2016.

[3] C. Wang, Z. Liu, and S. Chan, "Superpixel-based hand gesture recognition with Kinect depth camera," *IEEE Transactions on Multimedia*, vol. 17, no. 1, pp. 29–39, 2015.

[4] B. Feng, F. He, X. Wang, Y. Wu, H. Wang, S. Yi, and W. Liu, "Depth-projection-map-based bag of contour fragments for robust hand gesture recognition," *IEEE Transactions on Human-Machine Systems*, vol. 47, no. 4, pp. 511–523, 2017.

[5] H. Hamer, K. Schindler, E. Koller-Meier, and L. V. Gool, "Tracking a hand manipulating an object," in *Proc. on International Conference on Computer Vision*, Sept 2009, pp. 1475–1482.

[6] Y. Li and J. Wachs, "Recognizing hand gestures using the weighted elastic graph matching (WEGM) method," *Image and Vision Computing*, vol. 31, no. 9, pp. 649–657, 2013.

[7] P. Kumar, P. Vadakkepat, and L. Poh, "Graph matching based hand posture recognition using neuro-biologically inspired features," in *Proc. on International Conference on Control Automation Robotics Vision (ICARCV)*, Dec 2010, pp. 1151–1156.

[8] J. Wan, Q. Ruan, G. An, W. Li, Y. Liang, and R. Zhao, "The dynamic model embed in augmented graph cuts for robust hand tracking and segmentation in videos," *Mathematical Problems in Engineering*, vol. 2014, 2014.

[9] M. Fiedler, "A property of eigenvectors of nonnegative symmetric matrices and its application to graph theory," *Proc. Czechoslovak Mathematical Journal*, vol. 25, no. 4, pp. 619–633, 1975.

[10] B. Alwaely and C. Abhayaratne, "Graph spectral domain feature representation for in-air drawn number recognition," in *Proc. European Signal Processing Conference (EUSIPCO)*, 2017, pp. 370–374.

[11] ——, "Graph spectral domain shape representation," in *Proc. European Signal Processing Conference (EUSIPCO)*, 2018, pp. 603–607.

[12] T. Martinetz and K. Schulten, "A "neural-gas" network learns topologies," *Artificial Neural Networks*, pp. 397–402, 1991.

[13] D. I. Shuman, S. K. Narang, P. Frossard, A. Ortega, and P. Vandergheynst, "The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains," *IEEE Signal Processing Magazine*, vol. 30, no. 3, pp. 83–98, May 2013.

[14] L. Minto and P. Zanuttigh, "Exploiting silhouette descriptors and synthetic data for hand gesture recognition," *The Eurographics Association*, 2015.