Dynamic Hand Gesture Recognition With Leap Motion Controller

Wei Lu, *Member, IEEE*, Zheng Tong, and Jinghui Chu

***Abstract*—Dynamic hand gesture recognition is a crucial but challenging task in the pattern recognition and computer vision communities. In this paper, we propose a novel feature vector which is suitable for representing dynamic hand gestures, and presents a satisfactory solution to recognizing dynamic hand gestures with a Leap Motion controller (LMC) only. These have not been reported in other papers. The feature vector with depth information is computed and fed into the Hidden Conditional Neural Field (HCNF) classifier to recognize dynamic hand gestures. The systematic framework of the proposed method includes two main steps: feature extraction and classification with the HCNF classifier. The proposed method is evaluated on two dynamic hand gesture datasets with frames acquired with a LMC. The recognition accuracy is 89.5% for the LeapMotion-Gesture3D dataset and 95.0% for the Handicraft-Gesture dataset. Experimental results show that the proposed method is suitable for certain dynamic hand gesture recognition tasks.**

***Index Terms*—Depth data, dynamic hand gesture recognition, hidden conditional neural field (HCNF), leap motion controller (LMC).**

I. Introduction

With the development of many interactive applications in human computer interaction, human action recognition has obtained an increasing amount of attention in the pattern recognition and computer vision communities. Dynamic hand gesture recognition is a crucial part of human action recognition. However, the task is challenging because of the high variability of shape and the serious occlusion between fingers.

It is hard to capture such abundant dynamic hand gestures with a monocular video sensor, and this disadvantage limits the performance of video-based hand gesture recognition. In recent years, innovative depth sensors, such as the Leap Motion controller (LMC) [1] and Microsoft Kinect sensor [2], which provide three-dimensional (3-D) depth data of the scene, have contributed much to object segmentation and 3-D hand gesture recognition [3]. Moreover, Potter *et al*. proved the potential to recognize hand gestures with the LMC in [4]. Therefore, in this paper, we recognize dynamic hand gestures with a LMC. Unlike the Kinect sensor and other depth sensors, the output of the LMC is the depth data which consists of palm direction, fingertips positions, palm center position, and other relevant points. Therefore, no extra computational work is needed to get these information. Moreover, the localization precision of LMC is higher than other depth sensors (which is about 0.2 mm [5]). Recently, LMC was applied to hand gesture recognition by researchers. For example, Marin *et al*. used a LMC and Kinect sensor to recognize American Sign Language (ASL) [1], and Xu *et al*. used a LMC to recognize ten simple dynamic hand gestures [6].

Manuscript received May 30, 2016; revised July 03, 2016; accepted July 10, 2016. Date of publication July 12, 2016; date of current version July 28, 2016. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Antoni B. Chan.

The authors are with the School of Electronic Information Engineering, Tianjin University, Tianjin 300000, China (e-mail: luwei@tju.edu.cn; tongzheng2010@126.com; cjh@tju.edu.cn).

Color versions of one or more of the figures in this letter are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LSP.2016.2590470

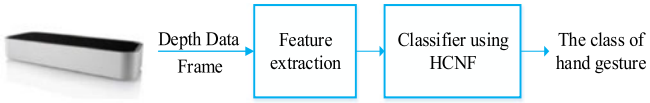

Fig. 1. The systematic framework.

There are already several hand gesture recognition systems [3], [7]. In [2], the cell occupancy feature and silhouette feature from depth data were extracted and fed into a classifier based on action graphs. In [8], the Local Occupancy Patterns feature was proposed and fed into an actionlet ensemble model. In [9], the histogram of oriented gradients feature vector was extracted and fed into a SVM-based classifier.

Additionally, several classifiers are commonly used for gesture recognition [3], [10], such as the Hidden Markov Model (HMM) [11], [12], Conditional Random Field (CRF) [13], Hidden Conditional Random Field (HCRF) [14], [15], Dynamic Time Warping [16], etc. In [17], Wang *et al*. introduced HCRF to recognize human gestures (e.g., head gestures and arm gestures). In [18], the Hidden Conditional Neural Field (HCNF) model was proposed, which extends HCRF by incorporating gate function used in neural networks. Previously, HCNF was only applied to speech recognition. However, our paper attempts to use HCNF in dynamic hand gesture recognition for the first time.

In this paper, we present a novel dynamic hand gesture recognition approach with an LMC. The basic framework is illustrated in Fig. 1. A feature vector is extracted from the depth data and fed into an HCNF-based classifier in order to recognize dynamic hand gestures.

II. Systematic Framework

Dynamic hand gesture recognition is considered to be the problem of sequential modeling and classification. This paper specifically offers a solution to depth data frame sequence classification with the corresponding hand gesture model in hand gesture recognition. The systematic framework of the proposed method (shown in Fig. 1) includes two steps: 1. Feature extraction, 2. Classification with the HCNF classifier.

A. Feature Extraction

Unlike the Kinect sensor, the LMC outputs depth data frames that consist of finger position, hand position, scaling data, frame timestamp, rotation, and so on. Therefore, the feature extraction

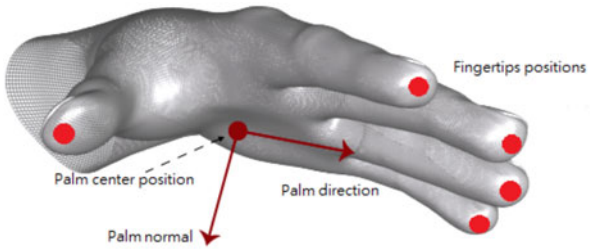

Fig. 2. Illustrations of Palm direction, Palm normal, Fingertips positions, and Palm center position.

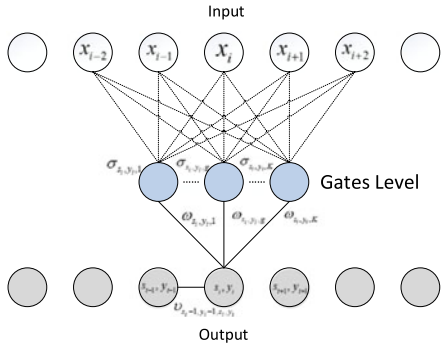

Fig. 3. The graph structure of HCNF.

time of the LMC is less than the Kinect sensor. The features used in this paper are based on palm direction, palm normal, fingertips positions, and palm center position data in depth data frames (shown in Fig. 2), including

1) Palm direction $\vec{D}$ represents a unit direction vector that points from the palm position toward the fingers.
2) Palm normal $\vec{N}$ is a normal vector to the palm.
3) Fingertips positions $F_i,\ i = 1, \ldots, 5$, represent the 3-D positions of individual fingertips.
4) Palm center position C represents the palm center position in 3-D space.

The proposed feature vector consists of single-finger features and double-finger features. The single-finger features derive from the work in [1], and to describe the interaction between adjacent fingertips, we present double-finger features. All the features values are normalized to the interval [0, 1].

The two types of features are described as follows:

1. Single-Finger Features: a) Fingertip-distances $\mathrm{Df}_i = \|F_i - C\|/M,\ i = 1, \ldots, 5$, are the Euclidean distances between the fingertips and the palm center. M is the Euclidean distance between the palm center and the middle fingertip. Note that, dividing by M normalizes the fingertip-distances to the interval [0, 1], and at the same time makes the method suitable for hands in different sizes. The scale factor M can be computed with the palm opened completely when the user starts to use the system.

b) Fingertip-angles $\mathrm{Af}_i = \angle(F_i^p - C,\ \vec{D})/\pi\ ,\ i = 1, \ldots, 5$, where F_i^p is the projection of F_i on the plane identified by $\vec{N}$, are the angles corresponding to the orientation of the projected fingertips with respect to the palm direction $\vec{D}$. The fingertip-angles are normalized with π.

c) Fingertip-elevations $\mathrm{Ef}_i = \mathrm{sgn}((F_i - F_i^P) \cdot \vec{N})\|F_i - F_i^P\|/M, i = 1, \ldots, 5$, are the distances of the fingertips from the plane corresponding to the palm region.

2. Double-Finger Features: a) Adjacent fingertip-distances $\mathrm{Daf}_i = \|F_i - F_{i+1}\|/M,\ i = 1, \ldots, 4$, are the Euclidean distances between adjacent fingertips.

b) Adjacent fingertip-angles $\mathrm{Aaf}_i = \angle(F_i - F_{i+1})/\pi$, $i = 1, \ldots, 4$, are the absolute angles between adjacent fingertips.

The proposed feature vector has two main benefits. First, single-finger features help to solve the problem of mislabeling which is usually caused by performing the dynamic hand gesture in different positions. Second, double-finger features help in distinguishing the different types of interactions between adjacent fingertips.

B. Classification with the HCNF Classifier

HCNF is used for the classification of temporal sequences (e.g., speech recognition). In this paper, a HCNF-based classifier is used to recognize dynamic hand gestures. HCNF not only has the advantages of HCRF, but can also consider different kinds of features.

1. Graph Structure of HCNF: The graph structure of HCNF is shown in Fig. 3. HCNF is an extension of HCRF by introducing gate function into it. From HCNF, we learn the mapping of observations $\boldsymbol{x} = \{x_1,\ x_2,\ \ldots x_m\}$ to class labels $y \in Y$, where $x_i \in \boldsymbol{x}$ is the feature vector proposed in the previous section.

A HCNF computes the conditional probability of a class label given a set of observations by

$$P(y|\ \boldsymbol{x},\ \theta) = \sum_{\boldsymbol{s}} P(y, \boldsymbol{s} \mid \boldsymbol{x}, \theta) = \frac{\sum_{\boldsymbol{s}} \exp(\Psi(y,\ \boldsymbol{s}, \boldsymbol{x}; \theta))}{\sum_{y' \in Y, s \in S^m} \exp(\Psi(y',\ \boldsymbol{s}, \boldsymbol{x}; \theta))} \quad (1)$$

where $\boldsymbol{s} = \{s_1,\ s_2, \ldots, s_m\}$, each $s_i \in S$ captures a certain underlying structure of each class and S is the set of hidden states in the model. Each observation x_i has a vector of hidden variables $s_i = \{g_i,\ m_i\}$ which are not observed on feature vector sequences. Further, g_i corresponds to the subgesture structure in an observation and m_i corresponds to the mixture component for each subgesture. The two kinds of hidden variables can capture surface variation. The potential function $\psi(y,\ \boldsymbol{s},\ \boldsymbol{x};\ \theta) \in \mathcal{R}$, parameterized by θ, measures the compatibility between a label, a set of observations, and a configuration of hidden states. $\psi(y,\ \boldsymbol{s},\ \boldsymbol{x};\ \theta)$ is shown below

$$\psi(y,\ \boldsymbol{s},\ \boldsymbol{x};\ \theta) = F_n^1(y,\ \boldsymbol{s},\ \boldsymbol{x}) + F_n^2(y,\ \boldsymbol{s},\ \boldsymbol{x}) \quad (2)$$

where $F_n^1(y,\ \boldsymbol{s},\ \boldsymbol{x})$ and $F_n^2(y,\ \boldsymbol{s},\ \boldsymbol{x})$ are functions defining the features in the model. $F_n^1(y,\ \boldsymbol{s},\ \boldsymbol{x})$ is a raw observation feature function extracted from frame t, $F_n^2(y,\ \boldsymbol{s},\ \boldsymbol{x})$ is a transition feature function extracted from frame t and $t-1$. $F_n^1(y,\ \boldsymbol{s},\ \boldsymbol{x})$ and $F_n^2(y,\ \boldsymbol{s},\ \boldsymbol{x})$ defined as

$$F_n^1(y,\ \boldsymbol{s},\ \boldsymbol{x}) = \sum_t \sum_g^K \omega_{y_t, s_t, g} G\left(\sigma_{y_t, s_t, g}^T f^1(y,\ \boldsymbol{s},\ \boldsymbol{x}, t)\right) \quad (3)$$

$$F_n^2(y,\ \boldsymbol{s},\ \boldsymbol{x}) = \sum_j \nu_i \sum_t f_j^2(y,\ \boldsymbol{s},\ \boldsymbol{x}, t, t-1) \quad (4)$$

where f^1 feature depends on single hidden variable values in the model, f^2 feature depends on pairs of values, ω and ν are components of θ, and $\sigma_{y_t, s_t, g}$ is a corresponding weight vector specific to the triple of y, s, and g, as shown in Fig. 3. In our hand gesture recognition task, we encode spatial consistency between the features $\boldsymbol{x} = \{x_1,\ x_2,\ \ldots, x_m\}$ with an undirected graph structure, where the hidden variables $\{s_1,\ s_2, \ldots, s_m\}$ correspond to vertices in the graph and the graph edges $(s_i,\ s_j) \in E$ correspond to links between variables s_i and s_j. Therefore, f^1 and f^2 correspond to the vertex and the edge in the graph respectively, and thus represent the structure of the graph. f^1 and f^2 are defined as

$$f^1\left(y,\ \boldsymbol{s},\ \boldsymbol{x}, t\right) = \begin{cases} 1, \text{ if } y_t = y_i,\ s_t = s_i, \text{ and } x_t = x_i \\ 0, \text{ otherwise,} \end{cases} \tag{5}$$

$$f^2\left(y,\ \boldsymbol{s},\ \boldsymbol{x}, t, t-1\right) = \begin{cases} 1, \text{ if } y_t = y_i, y_{t-1} = y_j,\ s_t = s_i, s_{t-1} = s_j, \text{ and } x_t = x_i \\ 0, \text{ otherwise.} \end{cases} \tag{6}$$

Further, $G(x)$ is a gate function defined as follows:

$$G\left(x\right) = 1/\left(1 + \ e^{-x}\right) - 0.5. \tag{7}$$

In HCNF, the observation feature function uses K gate functions by which it considers nonlinearity among features. It allows HCNF to be able to consider different kinds of features.

2. Training: We use the following objective function in training the parameters:

$$L\left(\theta\right) = \sum_{i=1}^{N} \log P\left(y \mid \boldsymbol{x}, \theta\right) \tag{8}$$

where N is the total number of training sequences. The $\log P(y \mid \boldsymbol{x},\ \theta)$ is the log-likelihood of the data. For our experiments, we use the quasi-Newton gradient ascent method to search for the optimal parameter values, $\theta^* = \text{argmax}_\theta L(\theta)$.

3. Inference: In [19], Sung *et al.* marginalize the hidden states using an N-best inference algorithm. Although the algorithm gives a reasonable result, it is not suitable for dynamic hand gesture recognition. Consequently, we decided to use the Viterbi algorithm for inference. This means that hidden state s is not marginalized in inference.

III. Experiments

A. Dynamic Hand Gesture Datasets

We built two kinds of dynamic hand gesture datasets, i.e., LeapMotion-Gesture3D dataset and Handicraft-Gesture dataset, with an LMC. All depth data frames of each dataset were acquired with the LMC's specific API.

1. LeapMotion-Gesture3D Dataset: Currently most of the dynamic hand gesture datasets were captured with the Kinect sensor, such as the MSRGesture3D dataset [2]. In order to compare the performance of our hand gesture recognition method with other methods, we built a dataset with an LMC, named LeapMotion-Gesture3D, which imitates the MSRGesture3D dataset. This dataset contains a subset of gestures defined by ASL. There are 12 gestures in the dataset: bathroom, blue, finish, green, hungry, milk, past, pig, and store, where, *j*, z.

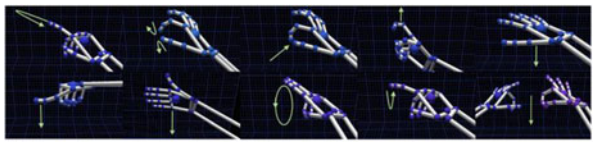

Fig. 4. Illustrations of the ten hand gestures from the Handicraft-Gesture dataset. Left to right, top to bottom: *Poke, Pinch, Pull, Scrape, Slap, Press, Cut, Circle, Key Tap, and Mow*. The green arrows are the motion trajectories of the hand or the fingers.

TABLE I
Recognition Accuracy of Single-finger or Double-finger Features for the LeapMotion-Gesture3D Dataset

Single-finger Features		Double-finger Features	
Feature Set	Accuracy	Feature Set	Accuracy
Df	**0.819**	Daf	**0.812**
Af	0.803	Aaf	0.749
Ef	0.787		
Df + Af	**0.829**	Daf + Aaf	**0.859**
Df + Ef	0.825		
Af + Ef	0.817		
Df + Af + Ef	**0.848**	/	/

TABLE II
Recognition Accuracy of Other Feature Sets for the LeapMotion-Gesture3D Dataset

Feature Set	Accuracy
Velocity of fingertip [16]	0.692
Acceleration of fingertip [16]	0.641
Chain code of fingertip [16]	0.560
Position of fingertip [1]	0.736
Hand pitch, roll, and yaw [20]	0.648

2. Handicraft-Gesture Dataset: In order to evaluate our method with more practical gestures, we built a dataset named Handicraft-Gesture. This dataset comprises of ten gestures which originate from pottery skills, i.e., poke, pinch, pull, scrape, slap, press, cut, circle, key tap, mow. The ten hand gestures are shown in Fig. 4.

In both datasets, the depth data is captured with 60 frames per second. There were 10 subjects helping to build the datasets and each one performed every gesture three times. Therefore, the LeapMotion-Gesture3D and Handicraft-Gesture dataset contain 360 and 300 sequences of depth data, respectively.

B. Experimental Results

The evaluation metric used in our experiments is the average recognition accuracy of six tests. In each test, the gestures belonging to seven randomly selected subjects are used for training, and the gestures belonging to the remaining three subjects are used for testing. Therefore, no subject appears in both the training set and test set.

The first group of experiments evaluated the prominence of different features for the LeapMotion-Gesture3D dataset with the HCNF-based classifier. Experimental results are summarized in Table I and Table II. Table II shows the recognition accuracies with the feature sets in [1], [16], and [20], respectively.

TABLE III
RECOGNITION ACCURACY OF HCNF OR HCRF FOR THE LEAPMOTION-GESTURE3D DATASET

Classifier	Feature Set	Accuracy
HCRF	$\mathrm{Df} + \mathrm{Af} + \mathrm{Ef} + \mathrm{Daf} + \mathrm{Aaf}$	0.878
HCNF	$\mathrm{Df} + \mathrm{Af} + \mathrm{Ef} + \mathrm{Daf} + \mathrm{Aaf}$	**0.895**

TABLE IV
RECOGNITION ACCURACY OF HCRF UNDER DIFFERENT PARAMETERS FOR THE LEAPMOTION-GESTURE3D DATASET

H	K	Accuracy	H	K	Accuracy
7	10	0.877	5	12	0.848
7	11	0.859	6	12	0.848
7	**12**	**0.895**	8	12	0.837
7	13	0.739	9	12	0.763

H corresponds to hidden state number, and K corresponds to gate number.

Due to low accuracy, the velocity, acceleration, chain code, hand pitch, roll, and yaw features are not suitable in describing the characteristics of dynamic hand gestures with abundant relative motions among fingers. Additionally, we did not use the feature of the positions of fingertips because the features of Df and Daf which are already used in our method are based on the information of fingertips positions.

The single-finger features were proposed in [1]. Therefore, the results of single-finger features are similar to the results in [1]. In order to describe the interaction between adjacent fingertips, we propose double-finger features. Table I also shows that double-finger features receive higher scores than single-finger features. Moreover, the combination of single-finger features with double-finger features receives the highest score, as shown in Table III. Further, we have performed a series of experiments to choose the best combination of hidden state number and gate number, and the different settings with the corresponding recognition accuracies are listed in Table IV. The best accuracy of 0.895 for the LeapMotion-Gesture3D dataset occurs at 7 hidden states and 12 gates.

In [17], Wang *et al*. had proven with experiments that HCRF outperformed both CRF and HMM for certain gesture recognition tasks. Furthermore, HCNF extends HCRF by incorporating gate function used in neural networks. Since HCRF computes the score of a hypothesis by summing up the linearly weighted features, it cannot consider nonlinearity among features that will be crucial for hand gesture recognition. Compared with HCRF, HCNF can incorporate any kind of features and is more suitable for our task. Table III compares the recognition accuracy using an HCRF-based classifier and using an HCNF-based classifier individually, with the same feature vector. The HCNF-based classifier outperforms the HCRF-based classifier in our experiments.

Table V compares the performance of the proposed method with the methods proposed in [2] and [21]. On the MSRGesture3D dataset, the accuracy of our method is higher than the other two methods. It is worth pointing out that our method only uses an LMC, while the other two methods use a Kinect sensor. That is to say, our method can well solve the dynamic hand gesture recognition task just with a low-cost LMC.

TABLE V
RECOGNITION ACCURACY COMPARISON FOR THE GESTURE3D DATASET

Method	Dataset	Accuracy
Proposed Method	LeapMotion-Gesture3D	**0.895**
Proposed Method	MSRGesture3D	0.890
Action Graph on Silhouette Features [2]	MSRGesture3D	0.877
SVM on random occupancy pattern features [21]	MSRGesture3D	0.885

TABLE VI
RECOGNITION ACCURACY FOR THE HANDICRAFT-GESTURE DATASET WITH HCRF

Feature Set	Accuracy
$\mathrm{Dcf} + \mathrm{Acf}$	0.939
$\mathrm{Df} + \mathrm{Af} + \mathrm{Ef}$	0.930
$\mathrm{Df} + \mathrm{Af} + \mathrm{Ef} + \mathrm{Dcf} + \mathrm{Acf}$	**0.950**

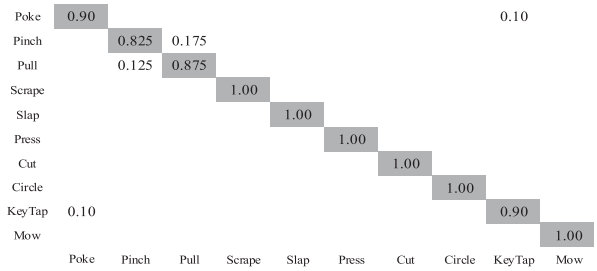

Fig. 5. The confusion matrix for the Handicraft-Gesture dataset. Gray cells represent true positive. Overall accuracy achieved is 95.0%.

Further, we tested our proposed method with the Handicraft-Gesture dataset. The best accuracy of 0.950 for the Handicraft-Gesture dataset occurs at 4 hidden states and 12 gates. The results are similar to the results on the LeapMotion-Gesture3D dataset. Table VI shows that double-finger features receive higher scores than single-finger features. Moreover, the combination of single-finger features with double-finger features receives the highest score. Fig. 5 shows the confusion matrix for recognizing hand gestures in the Handicraft-Gesture dataset. Since *Pitch* and *Pull* are very close, they were misclassified in some cases.

IV. CONCLUSION

In this work, we propose a novel feature vector which is suitable for representing dynamic hand gestures. The proposed feature vector that consists of single-finger features and double-finger features has two main benefits. First, single-finger features solve the problem of mislabeling which is caused by executing dynamic hand gesture in different positions. Second, double-finger features can help in distinguishing the different types of interactions between adjacent fingertips. The HCNF-based classifier considers the two main factors for dynamic hand gesture recognition: different kinds of features and complex underlying structure of dynamic hand gesture sequences. The experimental results show that our method achieved 95.0% recognition accuracy for the Handicraft-Gesture dataset and 89.5% for the LeapMotion-Gesture3D dataset. We have presented a satisfactory solution for recognizing dynamic hand gestures with the LMC alone, which has not been reported in other papers.

REFERENCES

[1] G. Marin, F. Dominio, and P. Zanuttigh, "Hand gesture recognition with jointly calibrated leap motion and depth sensor," *Multimedia Tools Appl.*, pp. 1–25, 2015, doi: 10.1007/s11042-015-2451-6.

[2] A. Kurakin, Z. Zhang, and Z. Liu, "A real time system for dynamic hand gesture recognition with a depth sensor," in *Proc. 20th Eur. Conf. Signal Process.* , 2012, pp. 1975–1979.

[3] H. Cheng, L. Yang, and Z. Liu, "A survey on 3D hand gesture recognitions," *IEEE Trans. Circuits Syst. Video Technol.*, 2015, doi: 10.1109/TCSVT.2015.2469551.

[4] L. E. Potter, J. Araullo, and L. Carter, "The leap motion controller: A view on sign language," in *Proc. Australian Comput.-Human Int. Conf. Augmentation Appl. Innovation Collaboration*, 2013, pp. 175–178.

[5] F. Weichert, D. Bachmann, B. Rudak, and D. Fisseler, "Analysis of the accuracy and robustness of the leap motion controller," *Sensors*, vol. 13, no. 5, pp. 6380–6393, 2013.

[6] Y. Xu, Q. Wang, X. Bai, and Y. L. Chen, "A novel feature extracting method for dynamic gesture recognition based on support vector machine," in *Proc. IEEE Int. Conf. Inf. Autom.*, 2014, pp. 437–441.

[7] R. Yang and S. Sarkar, "Coupled grouping and matching for sign and gesture recognition," *Comp. Vis. Image Underst.*, vol. 113, no. 6, pp. 663–681, 2009.

[8] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *Proc. Comput. Vis. Pattern Recogn.*, 2012, pp. 1290–1297.

[9] E. Ohn-Bar and M. M. Trivedi, "The power is in your hands: 3D analysis of hand gestures in naturalistic video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn. Workshops* , 2013, pp. 912–917.

[10] R. Poppe, "A survey on vision-based human action recognition," *Image Vis. Comput.*, vol. 28, no. 6, pp. 976–990, 2010.

[11] L. R. Rabiner and B.H. Juang, "An introduction to hidden markov models," *IEEE ASSP Magazine*, vol. 3, no. 1, pp. 4–16, Jan. 1986.

[12] H. M. Zhu and C. M. Pun, "Real-time hand gesture recognition from depth image sequences," in *Proc. 9th Int. Conf. Comput. Graphics, Imaging Vis.*, 2012, pp. 49–52.

[13] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. 18th Int. Conf. Mach. Learn.*, 2001, pp. 282–289.

[14] M. Chikkanna and R.M.R. Guddeti, "Kinect based real-time gesture spotting using HCRF," in *Proc. Int. Conf. Adv. Comput., Commun. Inform.*, 2013, pp. 925–928.

[15] A. Quattoni, M. Collins, and T. Darrell, "Conditional random fields for object recognition," in *Proc. NIPS*, 2004, pp. 1097–1104.

[16] M. K. Sohn, S. H. Lee, D. J. Kim, B. Kim, and H. Kim, "A comparison of 3D hand gesture recognition using dynamic time warping," in *Proc. 27th Conf. Image Vis. Comput. New Zealand*, 2012, pp. 418–422.

[17] S. B. Wang, A. Quattoni, L. P. Morency, D. Demirdjian, and T. Darrell, "Hidden conditional random fields for gesture recognition," in *Proc. Comput. Vis. Pattern Recogn.*, 2006, pp. 1521–1527.

[18] Y. Fujii, K. Yamamoto, and S. Nakagawa "Automatic speech recognition using hidden conditional neural fields," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2011, pp. 5036–5039.

[19] Y. H. Sung and D. Jurafsky, "Hidden conditional random fields for phone recognition," in *Proc. IEEE Workshop Autom. Speech Recogn. Understading*, 2009, pp. 107–112.

[20] M. Mohandes, S. Aliyu, and M. Deriche, "Arabic sign language recognition using the leap motion controller," in *Proc. IEEE 23rd Int. Symp. Ind. Electron.*, 2014, pp. 960–965.

[21] J. Wang, Z. Liu, J. Chorowski, Z. Chen, and Y. Wu, "Robust 3D action recognition with random occupancy patterns," in *Proc. 12th Eur. conf. Comput. Vis.*, 2012, pp. 872–885.