HAND GESTURE RECOGNITION BASED ON BAYESIAN SENSING HIDDEN MARKOV MODELS AND BHATTACHARYYA DIVERGENCE

Sih-Huei Chen, Ari Hernawan*, Yuan-Shan Lee*, and Jia-Ching Wang*

Dept. of Computer Science and Information Engineering, National Central University, Taoyuan, Taiwan

ABSTRACT

This work develops a system for recognizing common hand gestures. The main idea that underlies the developed system is the incorporation of Bhattacharyya divergence into Bayesian sensing hidden Markov models (BS-HMM). The system consists of two stages. First, a sequence of depth images is captured by Microsoft Kinect. The hand region is identified from the depth images by tracking the position of the hand using information about the skeleton, yielding the segmented depth images. A histogram of the oriented normal 4D (HON4D) and a histogram of oriented gradient (HOG) are then extracted from the segmented depth images to represent the motion patterns. Second, all training feature vectors are transformed by combining every k consecutive feature vectors into a sequence of distributions. The proposed Bhattacharyya divergence based BS-HMM (BDBS-HMM) is trained using the sequence of distributions. The proposed system is compared to the standard HMM and the BS-HMM using MSRGesture3D database and our database. Experimental results indicated that the proposed method outperforms the baseline methods.

Index Terms— Hand gesture recognition, Bayesian sensing hidden Markov model, Bhattacharyya divergence

1. INTRODUCTION

The natural interaction between computers and humans is one of the main focuses of human computer interaction (HCI) research. One natural way for humans to interact with computers is based on hand gestures. Several methods of hand gesture recognition (HGR) have been developed, and these methods can be classified by the use of different sensors [1, 2, 3, 4]. This work develops an HGR system, which uses a depth sensor, and does not require the user to wear any sensor on his or her body.

HGR systems must adopt suitable features. Kaâniche *et al.* [5, 6] utilized the HOG descriptor [7], which characterizes the local structure of an image captured by traditional RGB camera, as a hand gesture representation. Kläser *et al.* [8] developed a descriptor for RGB video sequence, which generalizes the idea of HOG to 3D. Rather than using RGB camera [5, 6, 8], Yang *et al.* [9] applied the HOG descriptor to depth images. Oreifej *et al.* [10] proposed an HON4D feature to describe the sequence of depth images, which jointly captures shape and motion information. Wang *et al.* [11] presented a random occupancy pattern (ROP) feature for depth sequence.

Previous works have used various classifiers for HGR, including k-nearest neighbors (k-NNs) [5], the support vector machine (SVM) [10], neural networks [12], and the finite state machine (FSM) [13]. In particular, hidden Markov models (HMMs), in which each observation can be considered to be a mixture model, have been utilized to provide a powerful probabilistic framework for capturing the temporal structure of data. They have been straightforwardly used to recognize hand gestures [14, 15]. Notably, in standard HMMs, a Gaussian distribution is typically considered to be the mixture component. Standard HMMs frequently use maximum-likelihood estimation (MLE) to evaluate the parameters in the Gaussian distribution, usually resulting in an overtrained model. To solve the overfitting problem, Saon *et al.* [16] proposed a Bayesian sensing hidden Markov model (BS-HMM), which incorporates Bayesian compressive sensing, and applied it to speech recognition. Our previous work [17] examined the effectiveness of the BS-HMM for HGR. It is noteworthy that the HMM and the BS-HMM are implemented using data in the form of a frame.

Unlike our previous work [17], this paper uses depth image [18] rather than conventional color image. Besides, the concept of the sequential distributions is developed and incorporated into the BS-HMM. Each distribution is composed of k consecutive frames, and regarded as an observation. The likelihood of each observation is then formulated based on Bhattacharyya divergence. In the proposed system, the BS-HMM with embedded Bhattacharyya divergence (BDBS-HMM) is utilized to model each class of hand gestures. The parameters of the proposed BDBS-HMM are trained using an expectation-maximization (EM) procedure. The main contributions of this work are two-fold. First, the proposed HGR system learns the hidden state of the depth image-based features. The proposed system provides stronger model regularization than the HMM-based system. Second, Bhattacharyya divergence is incorporated into the BS-HMM. The proposed BDBS-HMM provides an intuitive way to handle the depth image-based features in the form of a distribution.

*These authors contributed equally to this work.

978-1-5090-2175-8/17/$31.00 ©2017 IEEE

2. PRELIMINARY

This section briefly reviews the BS-HMM [16]. The BS-HMM exhibits stronger model regularization than the standard HMM and was originally used for speech recognition. In our previous work [17], the BS-HMM was utilized for HGR. The 3D coordinates that were extracted from each frame are regarded as the features of an RGB image. Given sequential features $X = \{\mathbf{x}_t\}_{t=1}^{T}$, $\mathbf{x}_t \in \mathbb{R}^D$, the BS-HMM [16] assumes that each observation $\mathbf{x}_t$ can be represented as a linear combination of a basis $\mathbf{\Phi}_i$ for state i. The likelihood of the observation $\mathbf{x}_t$ at state i is given as

$$\begin{aligned} & p(\mathbf{x}_t|\mathbf{w}_t, \lambda_i) \\ & \propto |\mathbf{R}_i|^{\frac{1}{2}} \exp\left[-\frac{1}{2}\left((\mathbf{x}_t - \mathbf{\Phi}_i\mathbf{w}_t)^{\mathrm{T}}\mathbf{R}_i\left(\mathbf{x}_t - \mathbf{\Phi}_i\mathbf{w}_t\right)\right)\right] \end{aligned} \quad (1)$$

where $\mathbf{R}_i$ is a state-dependent precision matrix; $\mathbf{w}_t$ is the sensing weight that is drawn from a prior $\mathcal{N}(\mathbf{0}, \mathbf{A}_i^{-1})$, and $\lambda_i = \{\mathbf{A}_i, \mathbf{\Phi}_i, \mathbf{R}_i\}$ are the state parameters. The marginal likelihood of sequential features X is obtained by integrating out the sensing weight $\mathbf{w}_t$, which is written as

$$p(X|\lambda) = \sum_{S=\{s_t\}}\left[\pi_{s_1}p(\mathbf{x}_1|\lambda_{s_1})\prod_{t=2}^{T}a_{s_{t-1}s_t}p(\mathbf{x}_t|\lambda_{s_t})\right] \quad (2)$$

where $p(\mathbf{x}_t|\lambda_{s_t}) = \int p(\mathbf{x}_t|\mathbf{w}_t, \lambda_{s_t})p(\mathbf{w}_t)d\mathbf{w}_t$, π_i is the initial probability associated with state i, a_{ij} is the probability of transition from state i to state j. $\lambda = \{\pi_i, a_{ij}, \mathbf{A}_i, \mathbf{\Phi}_i, \mathbf{R}_i\}$ are the parameters. Learning in the BS-HMM consists of optimizing Eq. (2) with respect to the parameters λ. The EM algorithm is used to solve the optimization problem.

3. PROPOSED SYSTEM

3.1. Preprocessing and Feature Extraction

The proposed system uses the depth sensor, Microsoft Kinect, to obtain a sequence of depth images that contain geometric information [18]. Then, skeletal information is used to track the hand among frames. The background/foreground separation is easily achieved using a depth threshold [19]. Afterwards, a full depth image is cropped as determined by the position of the hand and resized to 50 × 50 pixels. The next step is to normalize the segmented depth images that were obtained in the preceding step. For a segmented depth image $\mathbf{I} \in \mathbb{R}^{N\times M}$, each element I_{nm} is normalized as $\widetilde{I}_{nm} = 255 \times (I_{nm} - I_{\min})/(I_{\max} - I_{\min})$, where $I_{\min} = \min(\mathbf{I})$ and $I_{\max} = \max(\mathbf{I})$. To refine the image texture, the image contrast is increased by histogram equalization. Also, the median filter [20] is applied to reduce the noise in each image. In this work, a 5×5 median filter is used.

The effectiveness of two features– HOG [7] and HON4D [10]– to the proposed BDBS-HMM is investigated. A 900-dimensional HOG feature is extracted from each frame. PCA is then used to reduce the dimensionality of the HOG feature, yielding sequential features $X = \{\mathbf{x}_t\}_{t=1}^{T}$ with the dimension of 20. To obtain observations for the BDBS-HMM, the HOG feature vectors of k consecutive frames are calculated to form a distribution, yielding a sequential distribution $G = \{g_l\}_{l=1}^{L}$, $g_l = (\mu_l, \mathbf{\Sigma}_l)$ with $L = T/k$. The original HON4D feature is applied to a sequence of depth images to generate a video-level feature vector [10]. In this work, every four consecutive frames from a sequence of depth images are used to generate an HON4D feature. PCA is again used to reduce its dimensionality. The frame-level HON4D sequential features are used to train the HMM and the BS-HMM. As in the previous process, the sequential distributions are formed for BDBS-HMM training.

3.2. Bhattacharya Divergence Based Bayesian Sensing Hidden Markov Model (BDBS-HMM)

Notably, the BS-HMM models a sequence of data points. To model a sequence of distributions, this work proposes the BDBS-HMM, which incorporates Bhattacharyya divergence into the BS-HMM. The Bhattacharyya divergence D_B [21, 22], which measures the difference among probability distributions, is defined as

$$D_B(p, q) = -\ln\left(\sum_y \sqrt{p(y)q(y)}\right) \quad (3)$$

where $p(y)$ and $q(y)$ are arbitrary distributions over the same field. Given a sequential distribution $G = \{g_l\}_{l=1}^{L}$, $g_l = (\mu_l, \mathbf{\Sigma}_l)$ with $\mu_l \in \mathbb{R}^D, \mathbf{\Sigma}_l \in \mathbb{R}^{D\times D}$, the likelihood of the observation g_l in state i is written as

$$\begin{aligned} & p(\mu_l, \mathbf{\Sigma}_l|\mathbf{w}_l, \lambda_i) \\ & \propto |\mathbf{R}_i|^{\frac{1}{2}} \exp\Bigg[-\frac{1}{2}(\mu_l - \mathbf{\Phi}_i\mathbf{w}_l)^{\mathrm{T}}\mathbf{R}_i\left(\mu_l - \mathbf{\Phi}_i\mathbf{w}_l\right) \\ & \quad + \ln\left(\frac{\det\mathbf{R}_i}{\sqrt{\det\mathbf{\Sigma}_l\det(2\mathbf{R}_i - \mathbf{\Sigma}_l)}}\right)\Bigg] \end{aligned} \quad (4)$$

The BDBS-HMM has one more term than the BS-HMM (refer to Eq. (1)). Each observed distribution g_l is generated by the BDBS-HMM parameters $\lambda = \{\pi_i, a_{ij}, \mathbf{A}_i, \mathbf{\Phi}_i, \mathbf{R}_i\}$.

The MLE for λ is determined using the EM algorithm. In the BDBS-HMM, the state sequence $S = \{s_l\}_{l=1}^{L}$ is the latent variables. In the E-step, the expected value of the complete-data likelihood under the posterior distribution of the latent variable is computed as

$$\mathbb{E}\left\{\log p\left(G, S|\lambda\right)|G, \lambda^{\mathrm{old}}\right\} = \textstyle\sum_S p(S|G, \lambda^{\mathrm{old}})\log p(G, S|\lambda) \quad (5)$$

where λ^{old} is the current parameter value and the second term can be mainly focused on the following computation [16].

$$
\begin{aligned}
&p(\mu_l, \mathbf{\Sigma}_l|\lambda_i) = \int p(\mu_l, \mathbf{\Sigma}_l|\mathbf{w}_l, \lambda_i)p(\mathbf{w}_l|\mathbf{A}_i)d\mathbf{w}_l \\
&\propto |\mathbf{R}_i|^{\frac{1}{2}}\,|\mathbf{A}_i|^{\frac{1}{2}}\,|\mathbf{\Lambda}_i|^{\frac{1}{2}} \\
&\times \exp\left[-\frac{1}{2}\left(\mu_l^{\mathrm{T}}\mathbf{R}_i\mu_l - \mathbf{m}_{li}^{\mathrm{T}}\mathbf{\Lambda}_i^{-1}\mathbf{m}_{li} + \ln\left(\frac{\det \mathbf{R}_i}{\sqrt{\det \mathbf{\Sigma}_l \det(2\mathbf{R}_i - \mathbf{\Sigma}_l)}}\right)\right)\right]
\end{aligned} \tag{6}
$$

where $\mathbf{\Lambda}_i^{-1} = \mathbf{\Phi}_i^{\mathrm{T}}\mathbf{R}_i\mathbf{\Phi}_i + \mathbf{A}_i$ and $\mathbf{m}_{li} = \mathbf{\Lambda}_i\mathbf{\Phi}_i^{\mathrm{T}}\mathbf{R}_i\mu_l$. In the M-step, the updated parameter values λ^{new} are obtained by maximizing Eq. (5). Owing to limitations of space, only the closed-form solution for parameter $\mathbf{R}_i$ is shown here.

$$
\begin{aligned}
&(\mathbf{R}_i^{\mathrm{new}})^{-1} = \mathbf{\Phi}_i\mathbf{\Lambda}_i\mathbf{\Phi}_i^{\mathrm{T}} - \mathbf{R}_i^{-1} \\
&+\frac{\sum_l \gamma_l(i)\left(\left(\mu_l - \mathbf{\Phi}_i\mathbf{w}_l\right)\left(\mu_l - \mathbf{\Phi}_i\mathbf{w}_l\right)^{\mathrm{T}} + \frac{1}{2}\mathbf{\Sigma}_l^{-1} + (2\mathbf{R}_i - \mathbf{\Sigma}_\mathbf{l})^{-1}\right)}{\sum_l \gamma_l(i)}
\end{aligned} \tag{7}
$$

where $\gamma_l(i) = p(s_l = i|G)$. The closed-form solutions for parameters $\mathbf{\Phi}_i$ and $\mathbf{A}_i$ are similar to those obtained using the BS-HMM [16], but with the observed data point replaced by the mean of the observed distribution. The procedure for initializing $\mathbf{\Phi}_i$ follows the BS-HMM [16]. The sensing weight $\mathbf{w}_l$ can be estimated by maximizing a posterior (MAP) [16], which is given as $\mathbf{w}_l^{\mathrm{MAP}} = \mathbf{m}_{li}$.

4. EXPERIMENTAL RESULTS

4.1. Database and Evaluation Criteria

The effectiveness of the proposed method is experimentally evaluated using the hand gesture recognition task. The recognition performance is evaluated using F-measures, including precision, recall, and F1-score. The recognition rate is also considered. The experiments are conducted on two databases-both of raw depth video information that was captured by a Microsoft Kinect device. The first is the MSRGesture3D database [23]. MSRGesture3D comprises 12 classes of gesture. A total of 336 videos, including 12 dynamic American Sign Language (ASL) gestures, made by ten persons, were recorded. Each video consists of from 30 to 60 frames. Fig. 1 (a) displays an example of a depth image from the MSRGesture3D database. The second database includes self-recorded data that were recorded under real conditions. Our database consists of 150 videos with ten classes of gesture, including Up, Down, Left, Right, Rotate, No, Stop, Come, Zoom, and Ok. Each video is a sequence of 60 frames. Fig. 1 (b) displays an example of a depth image from our database. Most of the gesture videos in MSRGesture3D are well-segmented, and so show only the wrist and the full palm. However, the videos in our database are not well-segmented, as shown in Fig. 1 (b), so hand gesture localization is used to preprocess the data in our database. All of the experiments use half of the files for training and the other half for testing.

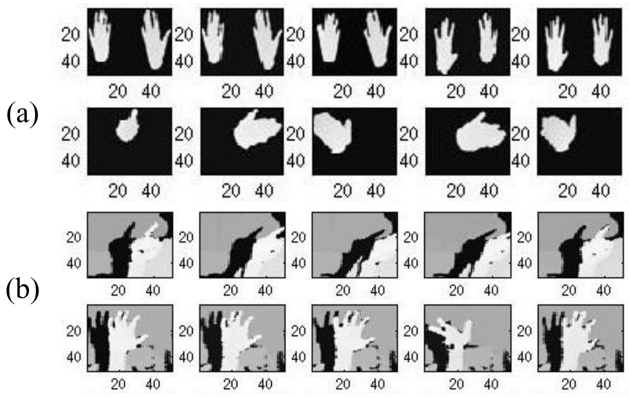

Fig. 1. Example depth images from (a) MSRGesture3D database. (b) Our database.

4.2. Baseline Methods and Experimental Settings

To confirm the efficiency of the proposed method, a standard HMM-based method (HMM) is chosen as the baseline. Since the HMM generally causes model overfitting, BS-HMM [16] is selected as another baseline method. The proposed method is also compared to four state-of-the-art methods from 1) Oreifej *et al.* [10], 2) Wang *et al.* [11], 3) Yang *et al.* [9], and 4) Kläser *et al.* [8]. Three experiments are performed. The first experiment studies the effect of the number of mixture components in the HMM, the BS-HMM and the BDBS-HMM. The second experiment examines the similarity of basis vectors in the BS-HMM and the BDBS-HMM during the training process. The automatic relevance determination (ARD) parameter is also investigated. The third experiment compares the proposed BDBS-HMM with the four state-of-the-art methods. In the first two experiments, two feature sets are studied. The first one is HOG [7] and the second is HON4D [10]. For a fair comparison, the HMM, the BS-HMM and the BDBS-HMM use the same feature set. The basis vectors of the BS-HMM and the BDBS-HMM are initialized by considering the 64 components of the standard HMM. Each experiment was conducted 20 times with random partitioning and the average results are reported. The detailed settings are as described in section 3.1.

4.3. Results and Discussions

First, the number of hidden states of the HMM, the BS-HMM, and the BDBS-HMM was set to two. In each state, the number of mixture components was set to two, four, eight and 16. Table 1 and Table 2 show the experimental results obtained using the MSRGesture3D dataset and our dataset, respectively. Experimental results indicate that the proposed BDBS-HMM outperforms the baseline methods in most cases. Notably, the BS-HMM and the BDBS-HMM use the 64-component HMM for initialization.

Next, the BS-HMM is compared with the proposed

Table 1. Recognition rate (%) obtained using various number of mixture components on MSRGesture3D database

Num. mix / state	2	4	8	16	64
HOG+HMM	84.46	87.32	**88.30**	87.32	82.08
HOG+BS-HMM	89.43	89.04	89.55	**91.07**	-
HOG+BDBS-HMM	88.78	91.34	92.47	**93.27**	-
HON4D+HMM	80.60	86.55	89.49	90.86	**91.09**
HON4D+BS-HMM	94.97	95.68	**95.74**	95.33	-
HON4D+BDBS-HMM	92.67	94.97	95.38	**95.89**	

Table 2. Recognition rate (%) obtained using various number of mixture components on our database

Num. mix / state	2	4	8	16	64
HOG+HMM	90.25	87.44	90.25	**90.31**	90.19
HOG+BS-HMM	89.75	93.94	**94.69**	94.44	-
HOG+BDBS-HMM	95.50	**96.69**	96.37	95.50	-
HON4D+HMM	90.94	91.63	91.63	93.94	**94.62**
HON4D+BS-HMM	93.06	93.69	**94.37**	94.31	-
HON4D+BDBS-HMM	94.00	**94.50**	93.94	93.62	-

BDBS-HMM in more detail. Theoretically, the set of basis vectors in a mixture component need to be more independent during learning. The first five iterations in the training process were studied. The left half of Fig. 2 presents the average cosine similarity between pairs of basis vectors in each iteration. The right half of Fig. 2 presents the mean and standard deviation of the ARD parameters in each iteration. The value of ARD represents the precision of the sensing weights. The comparison between the BS-HMM and the BDBS-HMM verifies that their convergence speeds are similar. Table 3 compares the BS-HMM and the BDBS-HMM in terms of the F-measures with the MSRGesture3D database. Each score is an average across all classes. Experimental results reveal that the BDBS-HMM outperforms the BS-HMM.

Finally, the proposed BDBS-HMM is experimentally compared to four state-of-the-art methods. Table 4 shows the results obtained with the MSRGesture3D database. For the BDBS-HMM, the number of mixture components is fixed at 16. Oreifej *et al.* [10] uses the HON4D as a feature descriptor. An SVM with a polynomial kernel is used for classification. The comparison between the proposed method and that of the Oreifej *et al.* [10] shows the superiority of temporal model.

Table 3. Comparison between BS-HMM and BDBS-HMM in terms of precision, recall and F1-score.

Metrics	Precision	Recall	F1-score
HON4D+BS-HMM	0.946	0.946	0.945
HON4D+BDBS-HMM	**0.959**	**0.958**	**0.957**

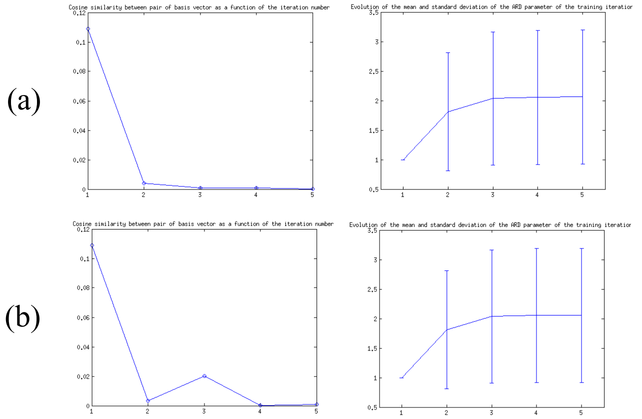

Fig. 2. Cosine similarity and ARD parameter as a function of the iteration number. (a) Results obtained using BS-HMM; (b) Results obtained using BDBS-HMM.

5. CONCLUSIONS AND FUTURE WORK

This work proposed an HGR system that is based on depth information. The main novelty of this system is its ability to deal with probabilistic features. To handle features in the form of sequential distributions, the Bhattacharyya divergence is incorporated into the framework of the BS-HMM. The parameters in the proposed BDBS-HMM are estimated based on MLE. An additional benefit is that model regularization is considered. A recursive solution for the parameters is derived using the EM procedure. The performance of the BDBS-HMM is compared with those of the standard HMM and the BS-HMM. The developed approach is also compared with the state-of-the-art methods. Experimental results demonstrate the superiority of the BDBS-HMM when applied the MSRGesture3D database. Future work may involve the integration of deep generative models, which learn a sequence of probabilistic features.

Table 4. Comparison between proposed method and state-of-the-art methods in terms of recognition rate (%).

Metrics	(%)
Wang *et al.* [11]	88.50
Yang *et al.* [9]	89.20
Kläser *et al.* [8]	85.23
Oreifej *et al.* [10]	92.45
HON4D+BDBS-HMM	**95.89**

6. REFERENCES

[1] Q. Pu, S. Gupta, S. Gollakota, and S. Patel, "Whole-home gesture recognition using wireless signals," in *Proc. MobiCom*, 2013, pp. 27–38.

[2] C. I. Penaloza, Y. Mae, F. F. Cuellar, M. Kojima, and T. Arai, "Brain machine interface system automation considering user preferences and error perception feedback," *IEEE Trans. Autom. Sci. Eng.*, vol. 11, no. 4, pp. 1275–1281, Oct. 2014.

[3] G. Pavlakos, S. Theodorakis, V. Pitsikalis, A. Katsamanis, and P. Maragos, "Kinect-based multimodal gesture recognition using a two-pass fusion scheme," in *Proc. ICIP*, 2014, pp. 1495–1499.

[4] J. Suarez and R. R. Murphy, "Hand gesture recognition with depth images: A review," in *Proc. RO-MAN*, 2012, pp. 411–417.

[5] M. B. Kaâniche and F. Brémond, "Recognizing gestures by learning local motion signatures of HOG descriptors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2247–2258, Nov. 2012.

[6] M. B. Kaâniche and F. Brémond, "Tracking HOG descriptors for gesture recognition," in *Proc. AVSS*, 2009, pp. 140–145.

[7] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. CVPR*, 2005, vol. 1, pp. 886–893.

[8] A. Kläser, M. Marszalek, and C. Schmid, "A spatio-temporal descriptor based on 3D-gradients," in *Proc. BMVC*, 2008, pp. 1–10.

[9] X. Yang, C. Zhang, and Y. Tian, "Recognizing actions using depth motion maps-based histograms of oriented gradients," in *Proc. ACM MM*, 2012, pp. 1057–1060.

[10] O. Oreifej and Z. Liu, "HON4D: Histogram of oriented 4D normals for activity recognition from depth sequences," in *Proc. CVPR*, 2013, pp. 716–723.

[11] J. Wang, Z. Liu, J. Chorowski, Z. Chen, and Y. Wu, "Robust 3D action recognition with random occupancy patterns," in *Proc. ECCV*, 2012, pp. 872–885.

[12] C. Wolf, G. W. Taylor, and F. Nebout, "Multi-scale deep learning for gesture detection and localization," in *Proc. ECCV*, 2014, pp. 474–490.

[13] Z. Li and R. Jarvis, "Real time hand gesture recognition using a range camera," in *Proc. ACRA*, 2009, pp. 529–534.

[14] S. Kim, G. Park, S. Yim, S. Choi, and S. Choi, "Gesture-recognizing hand-held interface with vibrotactile feedback for 3D interaction," *IEEE Trans. Consum. Electron.*, vol. 55, no. 3, pp. 1169–1177, Aug. 2009.

[15] Y. Wang, T. Yu, L. Shi, and Z. Li, "Using human body gestures as inputs for gaming via depth analysis," in *Proc. ICME*, 2008, pp. 993–996.

[16] G. Saon and J. T. Chien, "Bayesian sensing hidden Markov models," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 1, pp. 43–54, Jan. 2012.

[17] A. Hernawan, Y. S. Lee, A. Santoso, C. Y. Wang, and J. C. Wang, "Bayesian sensing hidden Markov model for hand gesture recognition," in *Proc. ASE BigData & SocialInformatics*, 2015, pp. 56:1–56:5.

[18] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake., "Real-time human pose recognition in parts from single depth images.," in *Proc. CVPR*, 2011, pp. 1297–1304.

[19] X. Liu and K. Fujimura, "Hand gesture recognition using depth data," in *Proc. FGR*, 2004, pp. 529–534.

[20] T. Huang, G. Yang, and G. Tang, "A fast two-dimensional median filtering algorithm," *IEEE Trans. Acoustics, Speech, Signal, Process.*, vol. 27, no. 1, pp. 13–18, Feb. 1979.

[21] J. Hershey, P. Olsen, and S. Rennie, "Variational Kullback-Leibler divergence for hidden Markov models," in *Proc. ASRU*, 2007.

[22] J. R. Hershey and P. A. Olsen, "Variational Bhattacharyya divergence for hidden Markov models," in *Proc. ICASSP*, 2008, pp. 4557–4560.

[23] A. Kurakin, Z. Zhang, and Z. Liu, "A real time system for dynamic hand gesture recognition with a depth sensor," in *Proc. EUSIPCO*, 2012, pp. 1975–1979.