Large-scale Isolated Gesture Recognition Using Convolutional Neural Networks

Pichao Wang[1], Wanqing Li[1], Song Liu[1], Zhimin Gao[1], Chang Tang[2] and Philip Ogunbona[1]

[1]Advanced Multimedia Research Lab, University of Wollongong, Australia
[2]School of Information Science and Engineering, Wuhan University of Science and Technology, Wuhan, China

pw212@uowmail.edu.au, {wanqing, songl}@uow.edu.au, zg126@uowmail.edu.au
tangchang@wust.edu.cn, philipo@uow.edu.au

***Abstract*—This paper proposes three simple, compact yet effective representations of depth sequences, referred to respectively as Dynamic Depth Images (DDI), Dynamic Depth Normal Images (DDNI) and Dynamic Depth Motion Normal Images (DDMNI). These dynamic images are constructed from a sequence of depth maps using bidirectional rank pooling to effectively capture the spatial-temporal information. Such image-based representations enable us to fine-tune the existing ConvNets models trained on image data for classification of depth sequences, without introducing large parameters to learn. Upon the proposed representations, a convolutional Neural networks (ConvNets) based method is developed for gesture recognition and evaluated on the Large-scale Isolated Gesture Recognition at the ChaLearn Looking at People (LAP) challenge 2016. The method achieved 55.57% classification accuracy and ranked 2^{nd} place in this challenge but was very close to the best performance even though we only used depth data.**

***Index Terms*—gesture recognition; depth map sequences; Convolutional Neural Networks**

I. Introduction

Gestures are naturally performed by humans, produced as part of deliberate actions, signs or signals, or subconsciously revealing intentions or attitude [1]. While they may involve the motion of all parts of the body, the studies of gestures usually focus on arms and hands which are essential in gesture communication. Recognition of gestures has recently attracted increasing attention due to its indubitable importance in many applications such as Human Computer Interaction (HCI), Human Robot Interaction (HRI) and assistive technologies for the handicapped and the elderly.

Gestures are one type of actions and many action recognition methods can be applied to gesture recognition. Recognition of human actions from depth/skeleton data is one of the most active research topics in multimedia signal processing in recent years due to the advantages of depth information over conventional RGB video, e.g. being insensitive to illumination changes. Since the first work of such a type [2] reported in 2010, many methods [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13] have been proposed based on specifical hand-crafted feature descriptors extracted from depth/skeleton. With the recent development of deep learning, a few methods have been developed based on Convolutional Neural Networks (ConvNets) [14], [15], [16], [17], [18] and Recurrent Neural Networks (RNNs) [19], [20], [21], [22]. However, it remains unclear how video could be effectively represented and fed to deep neural networks for classification. For example, one can conventionally consider a video as a sequence of still images with some form of temporal smoothness, or as a subspace of images or image features, or as the output of a neural network encoder. Which one among these and other possibilities would result in the best representation in the context of gesture recognition is not well understood.

Inspired by the recent work in [14], [15], [16], [23], this paper proposes for gesture recognition three simple, compact and effective representations of depth sequences which effectively decribe a short depth sequence with images. Such representations make it possible to use a standard ConvNet architecture to learn suitable "dynamic" features from the sequences by utilizing the ConvNet models trained from image data. Consequently, it avoids training millions of parameters from scratch and is especially valuable in the cases that lack sufficient annotated training video data. For instance, the large-scale isolated gesture recognition challenge [24] has on average only 144 video clips per class compared to 1200 images per class in ImageNet.

The proposed three representations are Dynamic Depth Image (DDI), Dynamic Depth Normal Image (DDNI) and Dynamic Depth Motion Normal Image (DDMNI). They are all constructed from a sequence of depth maps based on bidirectional rank pooling to encode the spatial (i.e. posture) and temporal (i.e. motion) information at different levels and are complementary to each other. Experimental results have shown that the three representations can improve the recognition accuracy substantially.

The rest of this paper is organized as follows. Section II briefly reviews the related works on gesture/action recognition based on depth and deep learning. Details of the proposed method are described in Section III. Experimental results are presented in Section IV. Section V concludes the paper.

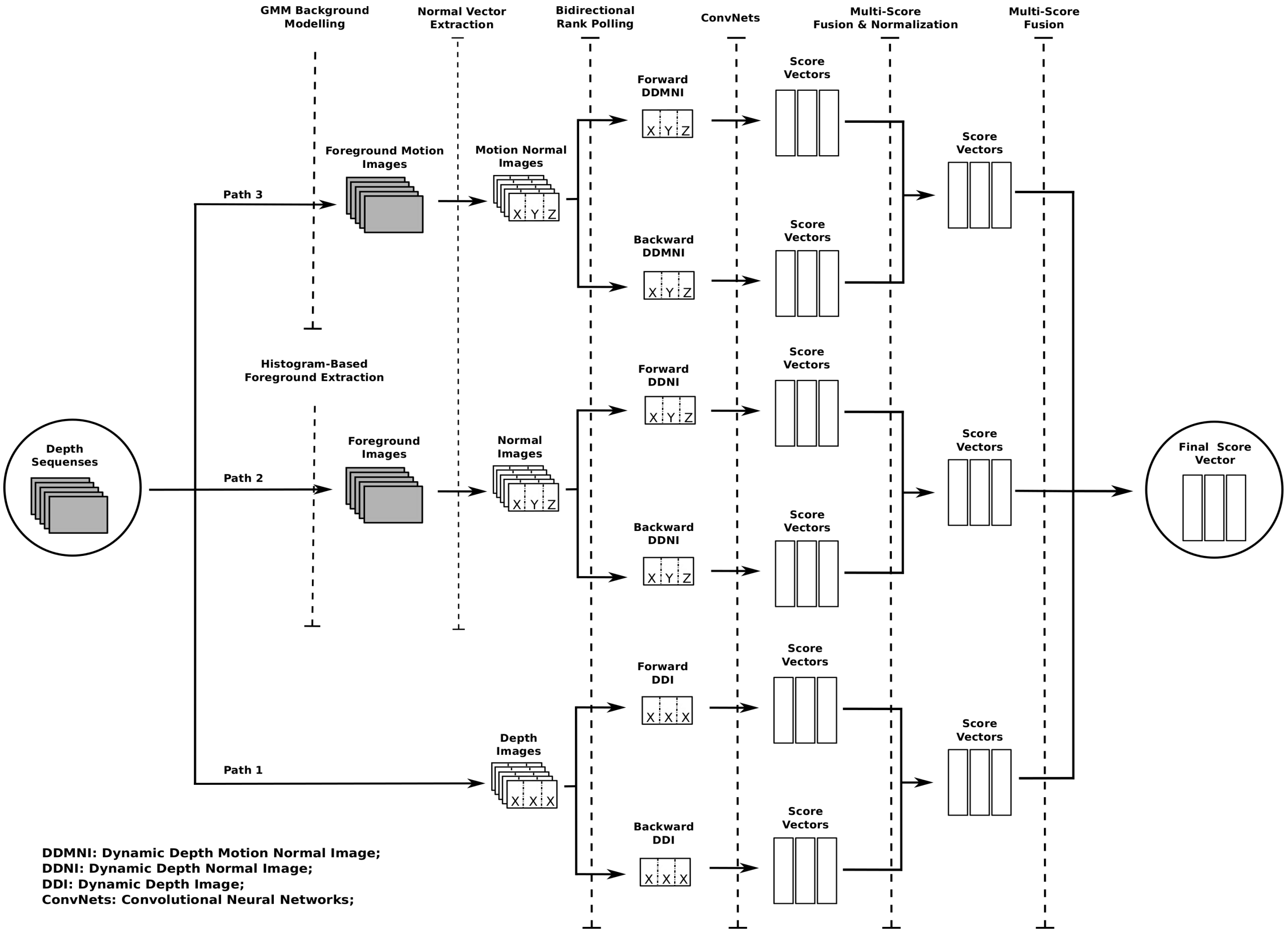

Fig. 1: The framework for proposed method.

II. Related Work

A. Depth Based Action Recognition

With Microsoft Kinect Sensors researchers have developed methods for depth map-based action recognition. Li et al. [2] sampled points from a depth map to obtain a bag of 3D points to encode spatial information and employ an expandable graphical model to encode temporal information [25]. Yang et al. [4] stacked differences between projected depth maps as a depth motion map (DMM) and then used HOG to extract relevant features from the DMM. This method transforms the problem of action recognition from spatio-temporal space to spatial space. In [5], a feature called Histogram of Oriented 4D Normals (HON4D) was proposed; surface normal is extended to 4D space and quantized by regular polychorons. Following this method, Yang and Tian [7] cluster hypersurface normals and form the polynormal which can be used to jointly capture the local motion and geometry information. Super Normal Vector (SNV) is generated by aggregating the low-level polynormals. In [10], a fast binary range-sample feature was proposed based on a test statistic by carefully designing the sampling scheme to exclude most pixels that fall into the background and to incorporate spatio-temporal cues.

B. Deep Leaning Based Recognition

Exiting deep learning approach can be generally divided into four categories based on how the video is represented and fed to a deep neural network. The first category views a video either as a set of still images [26] or as a short and smooth transition between similar frames [27], and each color channel of the images is fed to one channel of a ConvNet. Although obviously suboptimal, considering the video as a bag of static frames performs reasonably well. The second category is to represent a video as a volume and extends ConvNets to a third, temporal dimension [28], [29] replacing 2D filters with 3D ones. So far, this approach has produced little benefits, probably due to the lack of annotated training data. The third category is to treat a video as a sequence of images and feed the sequence to a RNN [30], [19], [20], [21], [22]. A RNN is typically considered as memory cells, which are sensitive to both short as well as long term patterns. It parses the video frames sequentially and encode the frame-level information in their memory. However, using RNNs did not give an improvement over temporal pooling of convolutional features [26] or over hand-crafted features. The last category is to represent a video in one or multiple compact images and adopt available trained ConvNet architectures for fine-tuning [14], [15], [16], [23]. This category has achieved state-of-the-art results of action recognition on many RGB and depth/skeleton datasets. The proposed method in this paper falls into the last category.

III. Proposed Method

The proposed method consists of three stages: construction of the three sets of dynamic images, ConvNets training and score fusion for classification, as illustrated in Fig. 1. Details are presented in the rest of this section.

A. Construction of Dynamic Images

The three sets of dynamic images, Dynamic Depth Images (DDIs), Dynamic Depth Normal Images (DDNIs) and Dynamic Depth Motion Normal Images (DDMNIs) are constructed from a sequence of depth maps through rank pooling [23]. They aim to capture both posture and motion information for gesture recognition.

1) Rank Pooling: Let $I_1, ..., I_T$ denote the frames in a sequence of depth maps, and $\varphi(I_t) \in \mathbb{R}^d$ be a representation or feature vector extracted from each individual frame I_t. Let $V_t = \frac{1}{t}\sum_{\tau=1}^{t} \varphi(I_t)$ be time average of these features up to time t. The ranking function associates to each time t a score $S(t|\mathbf{d}) =< \mathbf{d}, V_t >$, where $\mathbf{d} \in \mathbb{R}^d$ is a vector of parameters. The function parameters $\mathbf{d}$ are learned so that the scores reflect the rank of the frames in the video. In general, later times are associated with larger scores, $i.e.$ $q > t \Rightarrow S(q|\mathbf{d}) > S(t|\mathbf{d})$. Learning $\mathbf{d}$ is formulated as a convex optimization problem using RankSVM [31]:

$$\begin{aligned} \mathbf{d}^* &= \rho(I_1, ..., I_T; \varphi) = \arg\min_{\mathbf{d}} E(\mathbf{d}), \\ E(\mathbf{d}) &= \frac{\lambda}{2} \parallel \mathbf{d} \parallel^2 + \\ &\frac{2}{T(T-1)} \times \sum_{q>t} max\{0, 1 - S(q|\mathbf{d}) + S(t|\mathbf{d})\}. \end{aligned} \tag{1}$$

The first term in this objective function is the usual quadratic regular term used in SVMs. The second term is a hinge-loss soft-counting how many pairs $q > t$ are incorrectly ranked by the scoring function. Note in particular that a pair is considered correctly ranked only if scores are separated by at least a unit margin, $i.e.$ $S(q|\mathbf{d}) > S(t|\mathbf{d}) + 1$.

Optimizing the above equation defines a function $\rho(I_1, ..., I_T; \varphi)$ that maps a sequence of T depth video frames to a single vector d^*. Since this vector contains enough information to rank all the frames in the video, it aggregates information from all of them and can be used as a video descriptor. This process is called rank pooling.

2) Construction of DDI: Given a sequence of depth maps, the ranking pooling method [23] described above is employed to generate a dynamic depth image (DDI). The DDI is fed to the three channel of a ConvNet. Different from [23] the rank pooling is applied in a bidiretional way to convert one depth map sequence into two DDIs. As shown in Fig. 2, DDIs effectively capture the posture information, similar to key poses.

3) Construction of DDNI: In order to simultaneously exploit the posture and motion information in depth sequences, it is proposed to extract normals from depth maps and construct the so called DDNIs (dynamic depth normal images). For each depth map, the surface normal (n_x, n_y, n_z) at each location is calculated. Thus, three channels (N_x, N_y, N_z), referred to as a Depth Normal Image (DNI), are generated from the calculated normals, where (N_x, N_y, N_z) represents normal images for the three components (n_x, n_y, n_z) respectively. The sequence of DNIs goes through bidirectional rank pooling to generate two DDNIs, one being from forward ranking pooling and the other from backward rank pooling.

To minimise the interference of the background, it is assumed that the background in the histogram of depth maps occupies the last peak representing far distances. Specifically, pixels whose depth values are greater than a threshold defined by the last peak of the depth histogram minus a fixed tolerance (0.1 was set in our experiments) are considered as background and removed from the calculation of DDNIs by re-setting their depth values to zero. Through this simple process, most of the background can be removed and has much contribution to the DDNIs. Samples of DDNIs can be seen in Fig. 2.

4) Construction of DDMNI: The purpose of construction of a DDMNI is to further exploit the motion in depth maps. Gaussian Mixture Models (GMM) is applied to depth sequences to detect moving foreground. The same process as the construction of a DDNI (but without using histogram-based foreground extraction) is employed to the moving foreground. This process generates two DDMNIs, which specifically capture the motion information as illustrated in Fig. 2.

B. Network Training

After the construction of DDIs, DDNIs and DDMNIs, there are six dynamic images, as illustrated in Fig. 2, for each depth map sequence. Six ConvNets were trained on the six channels individually. Different layer configurations were used for the validation and testing sets provided by the Challenge. For validation, the layer configuration of six ConvNets follows the one in [32]. For testing, VGG-16 [33] was adopted for fine-tuning. The implementation is derived from the publicly available Caffe toolbox [34] based on three NVIDIA Tesla K40 GPU cards for both validation and testing.

The training procedure for validation is similar to the one in [32]. The network weights were learned using the mini-batch stochastic gradient descent with the momentum being set to 0.9 and weight decay being set to 0.0005. All hidden weight layers use the rectification (RELU) activation function. At each iteration, a mini-batch of 256 samples is constructed

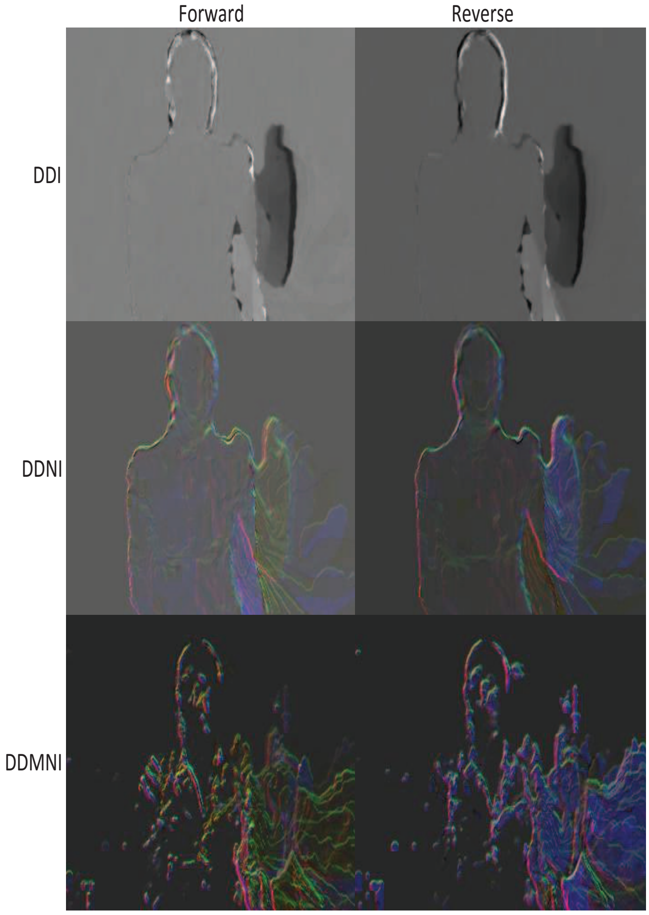

Fig. 2: Samples of generated forward and backward DDIs, DDNIs and DDMNIs for gesture Mudra1/Ardhapataka.

by sampling 256 shuffled training samples. All the images are resized to 256 × 256. The learning rate was set to 10^{-3} for fine-tuning with pre-trained models on ILSVRC-2012, and then it is decreased according to a fixed schedule, which is kept the same for all training sets. For each ConvNet the training undergoes 20K iterations and the learning rate decreases every 5K iterations. For all experiments, the dropout regularisation ratio was set to 0.5 in order to reduce complex co-adaptations of neurons in the nets.

For testing, the training procedure is similar to the one in [33]. The network weights were learned using the mini-batch stochastic gradient descent with the momentum being set to 0.9 and weight decay being set to 0.0005. All hidden weight layers use the rectification (RELU) activation function. At each iteration, a mini-batch of 32 samples was constructed by sampling 256 shuffled training samples. All the images are resized to 224 × 224. The learning rate was set to 10^{-3} for fine-tuning with pre-trained models on ILSVRC-2012, and then it is decreased according to a fixed schedule, which is kept the same for all training sets. For each ConvNet the training undergoes 50K iterations and the learning rate decreases every 20K iterations. For all experiments, the dropout regularisation ratio was set to 0.9 in order to reduce complex co-adaptations of neurons in the nets.

C. *Score Fusion for Classification*

Given a testing depth video sequence (sample), three pairs of dynamic images (DDIs, DDNIs, DDMNIs) are generated and fed into six different trained ConvNets. For each image pair, multiply-score fusion was used. The score vectors outputted by the two pair ConvNets are multiplied in an element-wise way and then the resultant score vectors are normalized using L_1 norm. The three normalized score vectors are then multiplied in an element-wise fashion and the max score in the resultant vector is assigned as the probability of the test sequence being the recognized class. The index of this max score corresponds to the recognized class label.

IV. Experiments

In this section, the Large-scale Isolated Gesture Recognition Dataset at the ChaLearn LAP challenge 2016 (ChaLearn LAP IsoGD Dataset) [35] and the evaluation protocol are described. The experimental results of the proposed method on this dataset are presented.

A. *Dataset*

The ChaLearn LAP IsoGD Dataset is derived from the ChaLearn Gesture Dataset (CGD) [36]. It includes 47933 RGB-D depth sequences, each RGB-D video representing one gesture instance. There are 249 gestures performed by 21 different individuals. The detailed information of this dataset are shown in Table I. In this paper, only depth maps were used to evaluate the performance of the proposed method. Some samples of depth sequences are shown in Fig. 3.

B. *Evaluation Protocal*

The dataset is divided into training, validation and test sets. All three sets consist of samples of different subjects so ensure that the gestures of one subject in validation and test sets will not appear in the training set.

For the isolated gesture recognition challenge, recognition rate r is used as the evaluation criteria. The recognition rate is calculated as:

$$r = \frac{1}{n}\delta(p_l(i), t_l(i)) \qquad (2)$$

where n is the number of samples; p_l is the predicted label; t_l is the ground truth; $\delta(j_1, j_2) = 1$, if $j_1 = j_2$, otherwise $\delta(j_1, j_2) = 0$.

C. *Experimental Results*

The results obtained by the proposed method on the validation and test sets are listed and compared with the baseline methods [37] (MFSK and MFSK+DeepID) in Table II. The codes and models can be downloaded at the author's homepage:https://sites.google.com/site/pichaossites/.

The results showed that the proposed method significantly outperformed the baseline methods, even though only single modality, i.e. depth data, was used while the baseline method used both RGB and depth videos.

TABLE I: Information of the ChaLearn LAP IsoGD Dataset.

Sets	# of labels	# of gestures	# of RGB videos	# of depth videos	# of subjects	label provided
Training	249	35878	35878	35878	17	Yes
Validation	249	5784	5784	5784	2	No
Testing	249	6271	6271	6271	2	No
All	249	47933	47933	47933	21	-

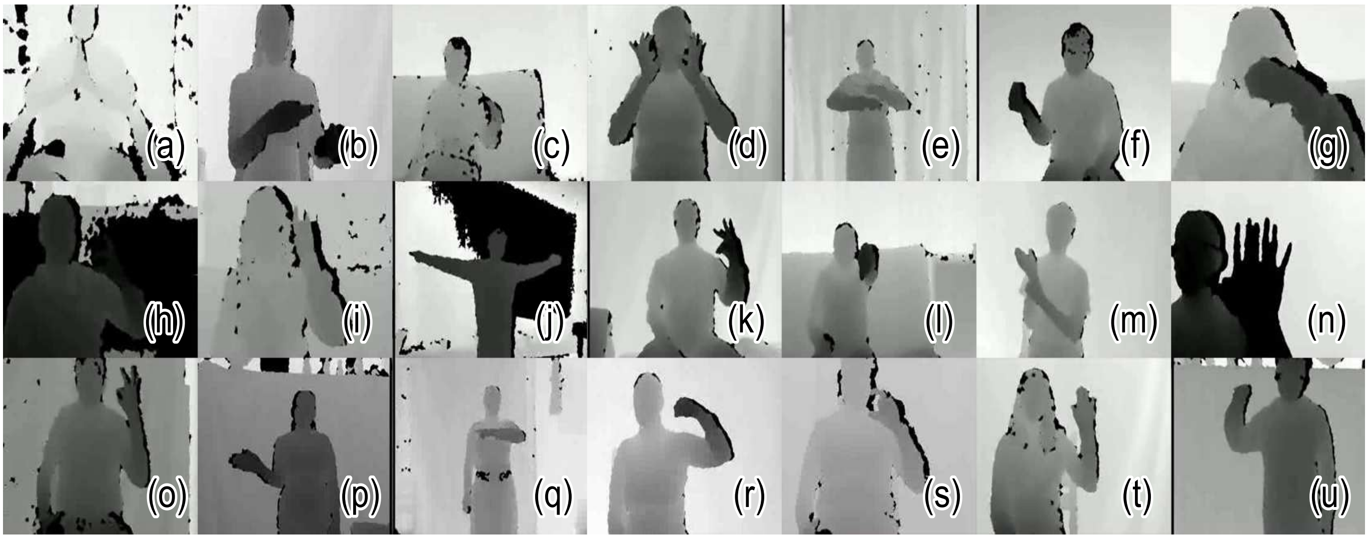

Fig. 3: The samples of 21 out of 249 gestures. From top left to bottom right, they are:
(a) ItalianGestures/Madonna; (b) GestunoTopography/92_harbour_port; (c) TaxiSouthAfrica/TaxiHandSigns2;
(d) GestunoSmallAnimals/129_cat_chat;(e) RefereeWrestlingSignals2/Reversal; (f) DivingSignals3/NotUnderstood;
(g) SurgeonSignals/ArmyNavyRetractor; (h) GangHandSignals1/EastSide; (i) SwatHandSignals1/DogNeeded;
(j) HelicopterSignals/MoveLeft; (k) GangHandSignals2/Killas; (l) TaxiSouthAfrica/TaxiHandSigns6;
(m) DivingSignals4/HowMuchAir; (n) ChineseNumbers/wu,TaxiSouthAfrica/TaxiHandSigns7;
(o) Mudra2/Vitarka,DivingSignals4/OK,GangHandSignals2/OK; (p) DivingSignals1/Around;
(q) CanadaAviationGroundCirculation1/DirigezVousVers; (r) MusicNotes/do; (s) GangHandSignals1/Crip;
(t) SwatHandSignals1/Stop; (u) RefereeWrestlingSignals2/Stalling,SwatHandSignals1/Breacher.

TABLE II: Comparative accuracy of proposed method and baseline methods on the ChaLearn LAP IsoGD Dataset.

Method	Set	Recognition rate r
MFSK	Validation	18.65%
MFSK+DeepID	Validation	18.23%
Proposed Method	Validation	**39.23%**
MFSK	Testing	24.19%
MFSK+DeepID	Testing	23.67%
Proposed	Testing	**55.57%**

TABLE III: Comparsion the performance of our submission with those of other teams. Our team secures the second place in the ICPR ChaLearn LAP challenge 2016.

Rank	Team	Recognition rate r
1	FLiXT [38]	56.8968%
2	**AMRL (ours)**	55.5733%
3	XDETVP-TRIMPS [39]	50.9329%
4	ICT_NHCI	46.8027%
5	XJTUfx	43.9164%
6	TARDIS	40.1531%
7	NTUST	20.3317%

The challenge results are summarized in Table III. We can see that our method is among the top performers and our recognition rate is very close to the best performance of this challenge (55.5733% vs. 56.8968%), even though we only used depth data for proposed method while the winner [38] adopted both depth and RGB modalities.

V. Conclusions

This paper presented three simple, compact yet effective representations of depth sequences for gesture recognition using convolutional Neural networks. They are all based on bidirectional rank pooling method converting the depth sequences into images. Such representations enables the use of existing ConvNets models directly on video data with fine-tuning without introducing large parameters to learn. The three representations represent the posture and motion in different levels and they are complementary to each other and improve the recognition accuracy largely. Experimental results on ChaLearn LAP IsoGD Dataset verified the effectiveness of the proposed method.

Acknowledgment

The authors would like to thank NVIDIA Corporation for the donation of a Tesla K40 GPU card used in this challenge.

References

[1] S. Escalera, V. Athitsos, and I. Guyon, "Challenges in multimodal gesture recognition," *Journal of Machine Learning Research*, vol. 17, no. 72, pp. 1–54, 2016.

[2] W. Li, Z. Zhang, and Z. Liu, "Action recognition based on a bag of 3D points," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2010, pp. 9–14.

[3] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 1290–1297.

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

[5] O. Oreifej and Z. Liu, "HON4D: Histogram of oriented 4D normals for activity recognition from depth sequences," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 716–723.

[6] M. A. Gowayyed, M. Torki, M. E. Hussein, and M. El-Saban, "Histogram of oriented displacements (HOD): Describing trajectories of human joints for action recognition," in *Proc. International Joint Conference on Artificial Intelligence (IJCAI)*, 2013, pp. 1351–1357.

[7] X. Yang and Y. Tian, "Super normal vector for activity recognition using depth sequences," in *Proc. IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 804–811.

[8] H. Rahmani, A. Mahmood, D. Q. Huynh, and A. Mian, "HOPC: Histogram of oriented principal components of 3D pointclouds for action recognition," in *Proc. European Conference on Computer Vision (ECCV)*, 2014, pp. 742–757.

[9] P. Wang, W. Li, P. Ogunbona, Z. Gao, and H. Zhang, "Mining mid-level features for action recognition based on effective skeleton representation," in *Proc. International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, 2014, pp. 1–8.

[10] C. Lu, J. Jia, and C.-K. Tang, "Range-sample depth feature for action recognition," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 772–779.

[11] H. J. Escalante, I. Guyon, V. Athitsos, P. Jangyodsuk, and J. Wan, "Principal motion components for one-shot gesture recognition," *Pattern Analysis and Applications*, pp. 1–16, 2015.

[12] R. Vemulapalli and R. Chellappa, "Rolling rotations for recognizing human actions from 3d skeletal data," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 1–9.

[13] J. Zhang, W. Li, P. O. Ogunbona, P. Wang, and C. Tang, "RGB-D-based action recognition datasets: A survey," *Pattern Recognition*, vol. 60, pp. 86–105, 2016.

[14] P. Wang, W. Li, Z. Gao, C. Tang, J. Zhang, and P. O. Ogunbona, "Convnets-based action recognition from depth maps through virtual cameras and pseudocoloring," in *Proc. ACM international conference on Multimedia (ACM MM)*, 2015, pp. 1119–1122.

[15] P. Wang, W. Li, Z. Gao, J. Zhang, C. Tang, and P. Ogunbona, "Action recognition from depth maps using deep convolutional neural networks," *Human-Machine Systems, IEEE Transactions on*, vol. 46, no. 4, pp. 498–509, 2016.

[16] P. Wang, Z. Li, Y. Hou, and W. Li, "Action recognition based on joint trajectory maps using convolutional neural networks," in *Proc. ACM international conference on Multimedia (ACM MM)*, 2016, pp. 1–5.

[17] P. Wang, W. Li, S. Liu, Y. Zhang, Z. Gao, and P. Ogunbona, "Large-scale continuous gesture recognition using convolutional neural networks," in *Proceedings of ICPRW*, 2016.

[18] Y. Hou, Z. Li, P. Wang, and W. Li, "Skeleton optical spectra based action recognition using convolutional neural networks," in *Circuits and Systems for Video Technology, IEEE Transactions on*, 2016, pp. 1–5.

[19] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1110–1118.

[20] V. Veeriah, N. Zhuang, and G.-J. Qi, "Differential recurrent neural networks for action recognition," in *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 4041–4049.

[21] W. Zhu, C. Lan, J. Xing, W. Zeng, Y. Li, L. Shen, and X. Xie, "Co-occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks," in *The 30th AAAI Conference on Artificial Intelligence (AAAI)*, 2016.

[22] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "NTU RGB+ D: A large scale dataset for 3D human activity analysis," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[23] H. Bilen, B. Fernando, E. Gavves, A. Vedaldi, and S. Gould, "Dynamic image networks for action recognition," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[24] J. Wan, S. Z. Li, Y. Zhao, S. Zhou, I. Guyon, and S. Escalera, "Chalearn looking at people rgb-d isolated and continuous datasets for gesture recognition," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2016, pp. 1–9.

[25] W. Li, Z. Zhang, and Z. Liu, "Expandable data-driven graphical modeling of human actions based on salient postures," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 18, no. 11, pp. 1499–1510, 2008.

[26] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond short snippets: Deep networks for video classification," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 4694–4702.

[27] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. Annual Conference on Neural Information Processing Systems (NIPS)*, 2014, pp. 568–576.

[28] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 1, pp. 221–231, 2013.

[29] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 4489–4497.

[30] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 2625–2634.

[31] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and computing*, vol. 14, no. 3, pp. 199–222, 2004.

[32] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Annual Conference on Neural Information Processing Systems (NIPS)*, 2012, pp. 1106–1114.

[33] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[34] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. B. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding." in *Proc. ACM international conference on Multimedia (ACM MM)*, 2014, pp. 675–678.

[35] H. J. Escalante, V. Ponce-Lpez, J. Wan, M. A. Riegler, B. Chen, A. Claps, S. Escalera, I. Guyon, X. Bar, P. Halvorsen, H. Mller, and M. Larson, "Chalearn joint contest on multimedia challenges beyond visual analysis: An overview," in *Proceedings of ICPRW*, 2016.

[36] I. Guyon, V. Athitsos, P. Jangyodsuk, and H. J. Escalante, "The chalearn gesture dataset (CGD 2011)," *Machine Vision and Applications*, vol. 25, no. 8, pp. 1929–1951, 2014.

[37] J. Wan, G. Guo, and S. Z. Li, "Explore efficient local features from rgb-d data for one-shot learning gesture recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 8, pp. 1626–1639, Aug 2016.

[38] Y. Li, Q. Miao, K. Tian, Y. Fan, X. Xu, R. Li, and J. Song., "Large-scale gesture recognition with a fusion of rgb-d data based on the c3d model," in *Proceedings of ICPRW*, 2016.

[39] G. Zhu, L. Zhang, L. Mei, J. Shao, J. Song, and P. Shen, "Large-scale isolated gesture recognition using pyramidal 3d convolutional networks," in *Proceedings of ICPRW*, 2016.