MOTION FEATURE AUGMENTED RECURRENT NEURAL NETWORK FOR SKELETON-BASED DYNAMIC HAND GESTURE RECOGNITION

Xinghao Chen　*Hengkai Guo*　*Guijin Wang*[⋆]　*Li Zhang*

Department of Electronic Engineering, Tsinghua University, Beijing, China

ABSTRACT

Dynamic hand gesture recognition has attracted increasing interests because of its importance for human computer interaction. In this paper, we propose a new motion feature augmented recurrent neural network for skeleton-based dynamic hand gesture recognition. Finger motion features are extracted to describe finger movements and global motion features are utilized to represent the global movement of hand skeleton. These motion features are then fed into a bidirectional recurrent neural network (RNN) along with the skeleton sequence, which can augment the motion features for RNN and improve the classification performance. Experiments demonstrate that our proposed method is effective and outperforms start-of-the-art methods.

Index Terms— Skeleton, Dynamic Hand Gesture Recognition, Recurrent Neural Network, Feature Augmentation

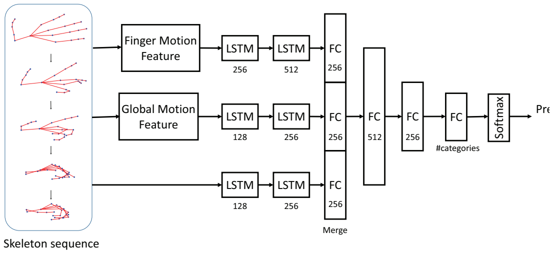


Fig. 1. The framework of our proposed method. Finger motion features and global motion features are extracted from the input dynamic hand gesture skeleton sequence. These motion features, along with the skeleton sequence, are fed into a recurrent neural network (RNN) to get the predicted class of input gesture.

1. INTRODUCTION

Due to its flexibility and expressiveness, hand gesture can provide an efficient and natural way for human computer interaction (HCI). Hand gesture recognition has been researched for decades and has great potentials for applications in sign language recognition, remote control and virtual reality etc [1, 2, 3, 4, 5, 6, 7]. Dynamic hand gesture recognition aims to understand what a hand sequence conveys. It remains a challenging task due to high intra-class variance because the way of performing a gesture differs from person to person.

Previous works on dynamic hand gesture recognition usually took RGB images and depth images [8, 9] as input [5]. Some of them used multi-modal input including IR images [6] or audio stream [7]. Recent progresses on hand pose estimation [10, 11, 12, 13] have greatly promoted the research on dynamic hand gesture recognition from 3D hand skeleton sequences. Smedt et al. [14] proposed a skeleton-based approach for dynamic hand gesture recognition and demonstrated its superiority over depth-based approaches. In their approach, a temporal pyramid representation was utilized to model temporal information. Shape of connected joints, histogram of hand directions and wrist rotations were used to characterize hand shape and hand movement. However, the amplitude of gesture is not considered in their approach and the temporal pyramid representation may lose some motion information.

The most important clues for dynamic hand gesture are articulated movements of fingers and the global movements of the hand. In prior works, some sort of joint angle features [4, 15] were utilized to describe the hand shape. However, these features are not sufficient enough to characterize the full pose of a hand.

In this paper, we propose a motion feature augmented RNN for skeleton-based dynamic hand gesture recognition. We extract the angles of bones from the hand skeleton, which is efficient and concise representation of the finger articulated movements. To describe the global movements of the hand, we extract the global rotation and global translation of the hand. A distance adaptive discretization scheme is given to better model the amplitude of the gestures. The finger motion features and global features are fed into a bidirectional RNN along with the skeleton sequence to predict the class of input gesture. Experiments on the publicly-available skeleton-based DHG-14/28 dataset [14] demonstrate the effectiveness of our proposed method.

This work was partially supported by National Science Foundation of China (No. 61271390) and State High-Tech R&D Program of China (No. 2015AA016304). We gratefully acknowledge NVIDIA for GPU donation.

[⋆]Corresponding author: wangguijin@tsinghua.edu.cn

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

2. PROPOSED FRAMEWORK

The framework of our proposed algorithm is shown in Figure 1. A hand skeleton sequence is taken as input and the class of gesture is predicted by RNN. Firstly the global motion features and finger motion features are extracted from the input skeleton sequence. The hand skeleton can be directly and effectively represented by a kinematic hand model whose parameters are the angles of bones, the global translation and global rotation[12, 13]. Therefore, these hand parameters can serve as efficient and discriminating features for dynamic hand gesture recognition. In our approach, theses features with offset and dynamic pose modelling are utilized as the motion features to represent dynamic hand gestures. The details of motion feature extraction will be presented in Section 3.

We exploit the recurrent neural network (RNN) to model temporal information for its success in temporal sequences recognition tasks[16, 6]. Though RNN can somehow learn features from the input sequences, some information may be absent or weakened, which will hinder the classification performance. Some previous works [17, 18] combined features extracted from deep neural network with hand-crafted features to enhance the discriminability of the features. Inspired by these works, in this paper we utilized a RNN which is augmented by motion features to classify dynamic hand gestures from skeleton sequence. The finger motion features and global features are fed into a RNN along with the skeleton sequence to predict the class of input gesture, which will be discussed in Section 4.

3. MOTION FEATURE EXTRACTION

In this section we will describe how to extract finger motion features $\mathcal{H}(\mathcal{S})$ and global motion features $\mathcal{G}(\mathcal{S})$ from the input hand skeleton sequence $\mathcal{S} = \{s^t\}_{t=1}^T$, where $s^t = \{x_i^t, y_i^t, z_i^t\}_{i=1}^J$ denotes a hand skeleton for frame t, T is the number of frames of this sequence and J is the number of joints for hand skeleton.

3.1. Global Motion Feature

The global motion features (global rotation and global translation) are important for dynamic hand gesture. Typically, the global status of the hand can be represented by the wrist joint, palm joint and metacarpophalangeal (MCP) joints, which is denoted by p^t. We use Kabsch algorithm [19] to infer the global rotation $\mathcal{G}_r$ and global translation $\mathcal{G}_l$, as shown in Equation (1):

$$[\mathcal{G}_l, \mathcal{G}_r] = Kabsch(p^t, p_0) \tag{1}$$

where $\mathcal{G}_r = (r_x, r_y, r_z)$ represents the rotations along three axis and $\mathcal{G}_l = (\rho, \theta, \phi)$ is the spherical coordinates of global translation. p_0 is a fake palm that centers at $(0, 0, 0)$ and faces the camera.

The amplitudes of hand gestures differ from person to person for the same gesture. Therefore previous work [14] ignored the amplitude part ρ of global translation. However, sometimes the amplitude is critical for gestures. For example, gesture $Grab$ and gesture $Pinch$ are quite similar except for the amplitude of the gesture. To this end, we propose a distance adaptive discretization(DAD) method to extract global translation amplitude feature, inspired by Distance Adaptive Scheme [20, 4] which is used for feature selection. The DAD method discretizes ρ into M bins using the threshold $\{\eta_i\}_{i=1}^M$. A gaussian distribution kernel $g(x)$ is used to generate the thresholds.

$$\int_0^{\eta_i} g(x)dx = \frac{i}{M}\int_0^{\sigma} g(x)dx \tag{2}$$

where σ is the standard deviation of the gaussian function. In our experiments, $\sigma = 1.5 r_{palm}$ where r_{palm} is the radius of the palm. The global feature can be written as Equation (3):

$$\Phi^t = [\rho_{bin}, \theta, \phi, r_x, r_y, r_z] \tag{3}$$

where ρ_{bin} is the discrete representation of ρ using the thresholds determined by Equation (2).

Similarly to previous works [21], we use offset pose Φ_{op}^t and dynamic pose Φ_{dp}^t to model the finger motion features. The offset pose represents the offset from current pose to the pose of first frame of the gesture sequence. The dynamic pose represents the difference of global features between current frame and several previous frames. There features can enhance the representability of the global motion of the hand and thus can model the temporal information of dynamic hand gesture.

$$\Phi_{op}^t = \Phi^t - \Phi^1 \tag{4}$$

$$\Phi_{dp}^t = \{\Phi^t - \Phi^{t-s} | s = 1, 5, 10\} \tag{5}$$

All above features are concatenated to form the global motion features $\mathcal{G}^t(\mathcal{S}) = [\Phi^t, \Phi_{op}^t, \Phi_{dp}^t]$ for frame t.

3.2. Finger Motion Feature

For many dynamic hand gesture, the movement of fingers are critical because the global movement may be non-significant, especially for fine-grained gestures. We use 20 DoFs (degree of freedoms) to model the finger movement. For the MCP joints, there are 2 DoFs for each joint. For proximal interphalangeal (PIP) and distal interphalangeal (DIP) joints, 1 DoF is used to describe the angle of bone. These parameters can retain rich information for the shape of the hand skeleton. We use $\mathcal{IK}(\cdot)$ to denote the inverse kinematics function that derive hand parameters from the original hand skeleton s^t.

$$\Theta^t = \mathcal{IK}(s^t) \tag{6}$$

Similarly, we use dynamic pose Θ_{dp}^t and offset pose Θ_{op}^t to model the finger motion feature.

$$\Theta_{op}^t = \Theta^t - \Theta^1 \tag{7}$$

Table 1. Recognition rates (%) of self-comparison experiments on DHG-14 dataset.

Method	fine			coarse			both		
	best	worst	avg±std	best	worst	avg±std	best	worst	avg±std
Skeleton	86.0	42.0	61.2 ± 12.37	**97.78**	**74.44**	86.44 ± 7.94	93.57	**67.86**	77.43 ± 6.82
Motion Features	84.0	46.0	71.5 ± 11.44	96.67	64.44	81.94 ± 8.17	90.0	58.57	78.21 ± 7.49
Ours	**90.0**	**56.0**	$\mathbf{76.9 \pm 9.19}$	**97.78**	72.22	$\mathbf{89.0 \pm 7.55}$	**94.29**	**67.86**	$\mathbf{84.68 \pm 6.67}$

$$\Theta^t_{dp} = \{\Theta^t - \Theta^{t-s} | s = 1, 5, 10\} \qquad (8)$$

These features are concatenated to form the finger motion features $\mathcal{F}^t(\mathcal{S}) = [\Theta^t, \Theta^t_{op}, \Theta^t_{dp}]$ for frame t.

4. DYNAMIC HAND GESTURE RECOGNITION

RNN has shown great successes in human action recognition and hand gesture recognition. Although RNN can learn features for the input data, the representability of the features may be absent in some aspects. To this end, we augment features for RNN by combining the hand-crafted global and finger motion features and the original skeleton. The framework of our proposed method for skeleton-based dynamic hand gesture recognition is shown in Figure 1. The finger motion features and global motion features are extracted from the input skeleton sequence. These motion features and the input skeleton sequence are fed into the RNN. Each branch contains two long short term memory (LSTM) layers and one fully connected (FC) layer. Outputs from three branches are concatenated together, followed by three FC layers and a softmax layer for class prediction. All layers are followed by a dropout layer and FC layers are followed by a ReLU function.

5. EXPERIMENTS

5.1. Dataset

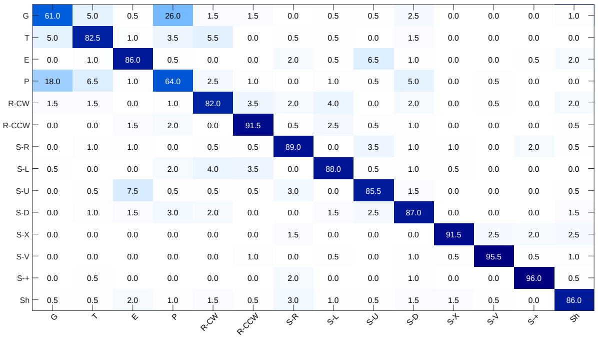
DHG-14/28 [14] is a public dynamic hand gesture dataset that provides hand gesture sequences with depth images and skeletons. Since our proposed method bases on hand skeleton, we only use the skeleton information of the dataset to conduct our experiments. DHG-14/28 is a challenging dataset since it contains hand gesture from 20 subjects and has 14 gestures with two different finger configurations.

5.2. Implementation

The proposed RNN framework is implemented in Keras [22]. We use Adam [23] algorithm with mini-batch of 32 to train the network. The parameters of Adam are set to default setting suggested in [23], with learning $lr = 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 1e^{-08}$.The network is trained for 100 epochs. In our experiments, M of Equation (2) is set to $M = 5$. Every skeleton sequence is subtracted by the palm position of the first frame and scaled the amplitude to 1 before fed into third branch in Figure 1.

5.3. Self-comparison

To verify the contributions of our proposed method, we conduct two self-comparison baseline experiments on DHG-14 dataset, which has 14 gesture classes. The first baseline (Motion Features) only takes motion features as input and remove the third branch of the framework shown in Figure 1. The second baseline (Skeleton) only use the skeleton sequences as input. We follow same experimental setup as [14], using a leave-one subject-out cross-validation (LOOCV) strategy for all following experiments. The proposed network is trained on data from 19 subjects and tested on the rest one. Therefore, these experiments are repeated 20 times, with different subject being used for testing. Previous work [14] only reported the average classification accuracy, which is not sufficient to evaluate the performance of the algorithm. In this paper, we report the worst, best and average results of 20 different splitting protocol as well as the standard derivation.



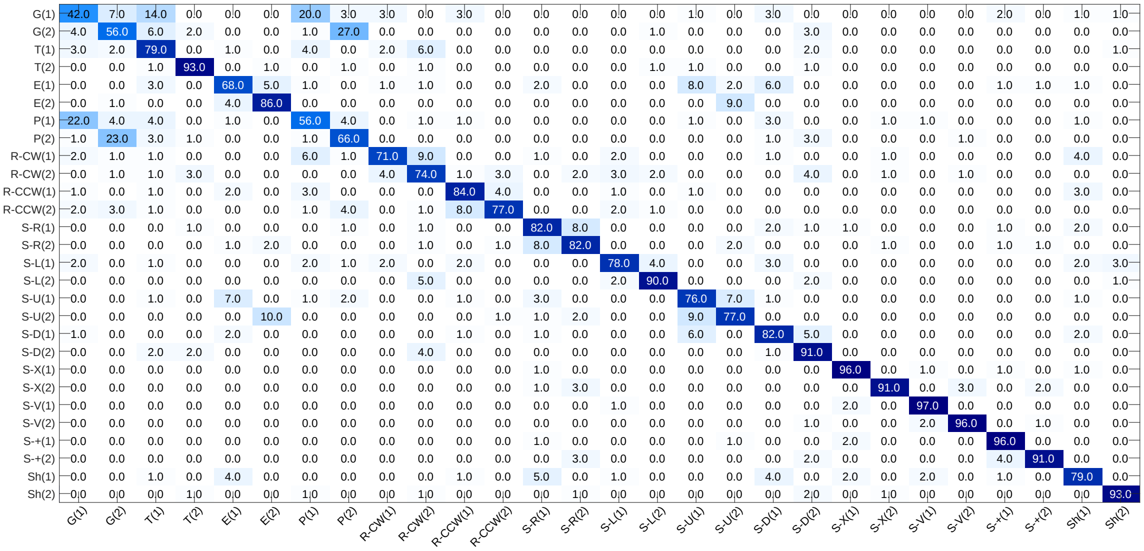
	G	T	E	P	R-CW	R-CCW	S-R	S-L	S-U	S-D	S-X	S-V	S-+	Sh
G	61.0	5.0	0.5	26.0	1.5	1.5	0.0	0.5	0.5	2.5	0.0	0.0	0.0	1.0
T	5.0	82.5	1.0	3.5	5.5	0.0	0.5	0.5	0.0	1.5	0.0	0.0	0.0	0.0
E	0.0	1.0	86.0	0.5	0.0	0.0	2.0	0.5	6.5	1.0	0.0	0.0	0.5	2.0
P	18.0	6.5	1.0	64.0	2.5	1.0	0.0	1.0	0.5	5.0	0.0	0.5	0.0	0.0
R-CW	1.5	1.5	0.0	1.0	82.0	3.5	2.0	4.0	0.0	2.0	0.0	0.5	0.0	2.0
R-CCW	0.0	0.0	1.5	2.0	0.0	91.5	0.5	2.5	0.5	1.0	0.0	0.0	0.0	0.5
S-R	0.0	1.0	1.0	0.0	0.5	0.5	89.0	0.0	3.5	1.0	1.0	0.0	2.0	0.5
S-L	0.5	0.0	0.0	2.0	4.0	3.5	0.0	88.0	0.5	1.0	0.5	0.0	0.0	0.0
S-U	0.0	0.5	7.5	0.5	0.5	0.5	3.0	0.0	85.5	1.5	0.0	0.0	0.0	0.5
S-D	0.0	1.0	1.5	3.0	2.0	0.0	0.0	1.5	2.5	87.0	0.0	0.0	0.0	1.5
S-X	0.0	0.0	0.0	0.0	0.0	0.0	1.5	0.0	0.0	0.0	91.5	2.5	2.0	2.5
S-V	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.5	0.0	1.0	0.5	95.5	0.5	1.0
S-+	0.0	0.5	0.0	0.0	0.0	0.0	2.0	0.0	0.0	1.0	0.0	0.0	96.0	0.5
Sh	0.5	0.5	2.0	1.0	1.5	0.5	3.0	1.0	0.5	1.5	1.5	0.5	0.0	86.0

Fig. 3. The confusion matrix of the proposed approach for DHG-14.

The recognition rates of these two baselines are shown in Table 1. In most cases, Our proposed method outperforms two baselines in terms of worst, best, average accuracy and the stand derivation, which verify the effectiveness of the proposed framework.

5.4. Comparison with State-of-the-arts

We compare our work with state-of-the art method [14] on DHG-14/28 dataset. The recognition rates of different methods on DHG-14 and DHG-28 dataset are shown in Table 2. It shows that our proposed method outperforms state-of-the-art



	G(1)	G(2)	T(1)	T(2)	E(1)	E(2)	P(1)	P(2)	R-CW(1)	R-CW(2)	R-CCW(1)	R-CCW(2)	S-R(1)	S-R(2)	S-L(1)	S-L(2)	S-U(1)	S-U(2)	S-D(1)	S-D(2)	S-X(1)	S-X(2)	S-V(1)	S-V(2)	S-+(1)	S-+(2)	Sh(1)	Sh(2)
G(1)	42.0	7.0	14.0	0.0	0.0	0.0	20.0	3.0	3.0	0.0	3.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	3.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	1.0	1.0
G(2)	4.0	56.0	6.0	2.0	0.0	0.0	1.0	27.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	3.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
T(1)	3.0	2.0	79.0	0.0	1.0	0.0	4.0	0.0	2.0	6.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
T(2)	0.0	0.0	1.0	93.0	0.0	1.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
E(1)	0.0	0.0	3.0	0.0	68.0	5.0	1.0	0.0	1.0	1.0	0.0	0.0	2.0	0.0	0.0	0.0	8.0	2.0	6.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0	0.0
E(2)	0.0	1.0	0.0	0.0	4.0	86.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	9.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
P(1)	22.0	4.0	4.0	0.0	1.0	0.0	56.0	4.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	3.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	1.0	0.0
P(2)	1.0	23.0	3.0	1.0	0.0	0.0	1.0	66.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	3.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0
R-CW(1)	2.0	1.0	1.0	0.0	0.0	0.0	6.0	1.0	71.0	9.0	0.0	0.0	1.0	0.0	2.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	4.0	0.0
R-CW(2)	0.0	1.0	1.0	3.0	0.0	0.0	0.0	0.0	4.0	74.0	1.0	3.0	0.0	2.0	3.0	2.0	0.0	0.0	0.0	4.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0
R-CCW(1)	1.0	0.0	1.0	0.0	2.0	0.0	3.0	0.0	0.0	0.0	84.0	4.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.0	0.0
R-CCW(2)	2.0	3.0	1.0	0.0	0.0	0.0	1.0	4.0	0.0	1.0	8.0	77.0	0.0	0.0	2.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
S-R(1)	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0	82.0	8.0	0.0	0.0	0.0	0.0	2.0	1.0	1.0	0.0	0.0	0.0	1.0	0.0	2.0	0.0
S-R(2)	0.0	0.0	0.0	0.0	1.0	2.0	0.0	0.0	0.0	1.0	0.0	1.0	8.0	82.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0	1.0	0.0	0.0
S-L(1)	2.0	0.0	1.0	0.0	0.0	0.0	2.0	1.0	2.0	0.0	2.0	0.0	0.0	0.0	78.0	4.0	0.0	0.0	3.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	3.0
S-L(2)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	5.0	0.0	0.0	0.0	0.0	2.0	90.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
S-U(1)	0.0	0.0	1.0	0.0	7.0	0.0	1.0	2.0	0.0	0.0	1.0	0.0	3.0	0.0	0.0	0.0	76.0	7.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
S-U(2)	0.0	0.0	0.0	0.0	0.0	10.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	2.0	0.0	0.0	9.0	77.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
S-D(1)	1.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	6.0	0.0	82.0	5.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0
S-D(2)	0.0	0.0	2.0	2.0	0.0	0.0	0.0	0.0	0.0	4.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	91.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
S-X(1)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	96.0	0.0	1.0	0.0	1.0	0.0	1.0	0.0
S-X(2)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	3.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	91.0	0.0	3.0	0.0	2.0	0.0	0.0
S-V(1)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	97.0	0.0	0.0	0.0	0.0	0.0
S-V(2)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	2.0	96.0	0.0	1.0	0.0	0.0
S-+(1)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	2.0	0.0	0.0	0.0	96.0	0.0	0.0	0.0
S-+(2)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0	0.0	4.0	91.0	0.0	0.0
Sh(1)	0.0	0.0	1.0	0.0	4.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	5.0	0.0	1.0	0.0	0.0	0.0	4.0	0.0	2.0	0.0	2.0	0.0	1.0	0.0	79.0	0.0
Sh(2)	0.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	93.0

Fig. 2. The confusion matrix of the proposed approach for DHG-28.

Table 2. Comparison of recognition rates (%) on DHG-14/28 dataset.

Method	DHG-14			DHG-28
	fine	coarse	both	both
Smedt et al. [14]	73.60	88.33	83.07	80.0
Ours	**76.9**	**89.0**	**84.68**	**80.32**

work [14] on DHG-14 dataset in terms of coarse, fine and all gestures. To better illustrate the performance of our proposed algorithm, the confusion matrix of 14 classes is shown in Figure 3. It can be observed that the confusion between gesture *Grab* and *Pinch* is severe, due to the high similarity of these two gestures. However, our algorithm does improve the performance of these two gestures compared with those of [14]. It can be observed that our method promotes the classification accuracy of fine-grained gestures a lot. The improvement of recognition rate of coarse-grained gestures is comparatively little because it's already quite good.

As shown in Table 2, our method is also better than [14] when considering the more complicated 28-gestures classification task, which demonstrates the effectiveness of our proposed algorithm. The confusion matrix of 28 classes is shown in Figure 2. A metric called Loss of Accuracy when Removing the Finger Differentiation (LARFD) was proposed in [14] to evaluate to what degree we can blame the loss of accuracy from 14-gestures to 28-gestures classification on the intra-gesture confusion. The smaller LARFD metric is, the less loss of accuracy is due to intra-gesture confusion. The LARFD metric of different methods is listed in Table 3. We can see that our proposed algorithm outperforms [14].

Table 3. Comparison of LAFED metric.

Method	Smedt et al. [14]	Ours
LAFED	0.0114	**0.0075**

6. CONCLUSION

This paper proposes an algorithm to augment the motion features for recurrent neural network to recognize skeleton-based dynamic hand gestures. The finger motion features are extracted from the input skeleton sequence to describe the articulated movements of fingers and the global motion features are extracted to represent the global translation and rotation of the hand. The motion features, along with the skeleton sequence, are fed into a RNN to predict the class of input gesture. Experiments on the public DHG-14/28 dataset demonstrate that our proposed method outperforms state-of-the-art methods. Future work may focus on a hierarchical coarse to fine framework to achieve better classification performance.

7. REFERENCES

[1] Sushmita Mitra and Tinku Acharya, "Gesture recognition: A survey," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 37, no. 3, pp. 311–324, 2007.

[2] Bobo Zeng, Guijin Wang, and Xinggang Lin, "A hand gesture based interactive presentation system utilizing

heterogeneous cameras," *Tsinghua Science and Technology*, vol. 17, no. 3, pp. 329–336, 2012.

[3] Xinghao Chen, Chenbo Shi, and Bo Liu, "Static hand gesture recognition based on finger root-center-angle and length weighted mahalanobis distance," in *SPIE Photonics Europe*. International Society for Optics and Photonics, 2016, pp. 98970U–98970U.

[4] Cao Dong, Ming C Leu, and Zhaozheng Yin, "American sign language alphabet recognition using microsoft kinect," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2015, pp. 44–52.

[5] Eshed Ohn-Bar and Mohan Manubhai Trivedi, "Hand gesture recognition in real time for automotive interfaces: A multimodal vision-based approach and evaluations," *IEEE Transactions on Intelligent Transportation Systems*, vol. 15, no. 6, pp. 2368–2377, 2014.

[6] Pavlo Molchanov, Xiaodong Yang, Shalini Gupta, Kihwan Kim, Stephen Tyree, and Jan Kautz, "Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4207–4215.

[7] N. Neverova, C. Wolf, G. Taylor, and F. Nebout, "Moddrop: Adaptive multi-modal gesture recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 8, pp. 1692–1706, Aug 2016.

[8] Guijin Wang, Xuanwu Yin, Xiaokang Pei, and Chenbo Shi, "Depth estimation for speckle projection system using progressive reliable points growing matching," *Applied optics*, vol. 52, no. 3, pp. 516–524, 2013.

[9] Chenbo Shi, Guijin Wang, Xuanwu Yin, Xiaokang Pei, Bei He, and Xinggang Lin, "High-accuracy stereo matching based on adaptive ground control points," *IEEE Transactions on Image Processing*, vol. 24, no. 4, pp. 1412–1423, 2015.

[10] James S Supancic, Grégory Rogez, Yi Yang, Jamie Shotton, and Deva Ramanan, "Depth-based hand pose estimation: data, methods, and challenges," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1868–1876.

[11] Markus Oberweger, Paul Wohlhart, and Vincent Lepetit, "Training a feedback loop for hand pose estimation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3316–3324.

[12] Danhang Tang, Jonathan Taylor, Pushmeet Kohli, Cem Keskin, Tae-Kyun Kim, and Jamie Shotton, "Opening the black box: Hierarchical sampling optimization for estimating human hand pose," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3325–3333.

[13] Qi Ye, Shanxin Yuan, and Tae-Kyun Kim, "Spatial attention deep net with partial pso for hierarchical hybrid hand pose estimation," in *The European Conference on Computer Vision (ECCV)*, 2016.

[14] Quentin De Smedt, Hazem Wannous, and Jean-Philippe Vandeborre, "Skeleton-based dynamic hand gesture recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2016, pp. 1–9.

[15] Wei Lu, Zheng Tong, and Jinghui Chu, "Dynamic hand gesture recognition with leap motion controller," *IEEE Signal Processing Letters*, vol. 23, no. 9, pp. 1188–1192, 2016.

[16] Yong Du, Wei Wang, and Liang Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1110–1118.

[17] Pichao Wang, Zhaoyang Li, Yonghong Hou, and Wanqing Li, "Combining convnets with hand-crafted features for action recognition based on an hmm-svm classifier," *arXiv preprint arXiv:1602.00749*, 2016.

[18] Sajith Kecheril Sadanandan, Petter Ranefall, and Carolina Wählby, "Feature augmented deep neural networks for segmentation of cells," in *European Conference on Computer Vision Workshops*. Springer, 2016, pp. 231–243.

[19] Wolfgang Kabsch, "A solution for the best rotation to relate two sets of vectors," *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography*, vol. 32, no. 5, pp. 922–923, 1976.

[20] Hui Liang, Junsong Yuan, and Daniel Thalmann, "Parsing the hand in depth images," *IEEE Transactions on Multimedia*, vol. 16, no. 5, pp. 1241–1253, 2014.

[21] Hongzhao Chen, Guijin Wang, Jing-Hao Xue, and Li He, "A novel hierarchical framework for human action recognition," *Pattern Recognition*, vol. 55, pp. 148–159, 2016.

[22] François Chollet, "Keras," `https://github.com/fchollet/keras`, 2015.

[23] Diederik Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.