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
{chen-xh13@mails, wangguijin@}tsinghua.edu.cn

ICIP 2017
IEEE International Conference on Image Processing
September 17-20, 2017, Beijing, China

INTRODUCTION

- 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.

FRAMEWORKS

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.

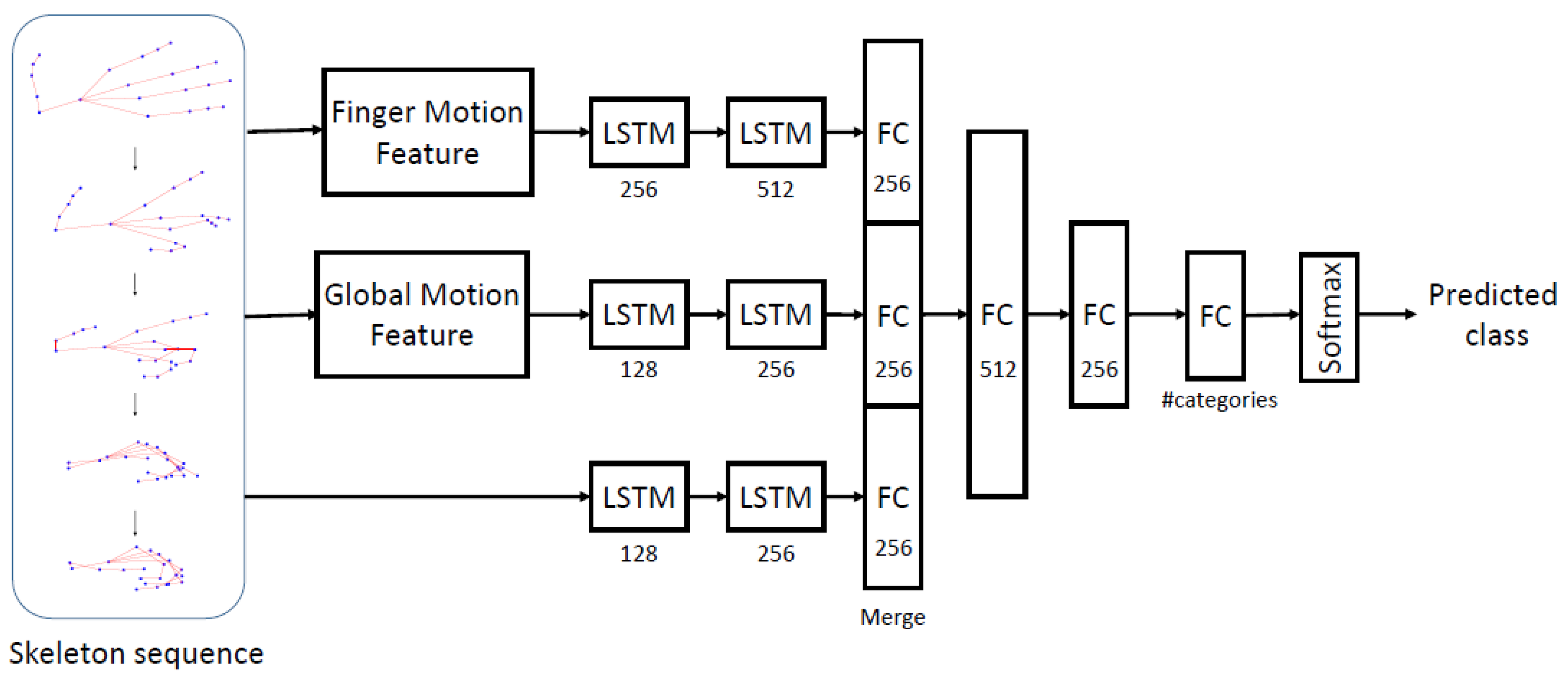

Figure 1: Proposed framework

METHODS

- Finger Motion Feature

Motivation: Joint coordinates are highly correlated, Hand model parameters are more compact and effective representations

- Finger Motion Feature

How to represent a hand skeleton: angles between conjunct bones $\Theta^t = \mathcal{IK}(s^t)$

Impose temporal information into the features

$$\Theta_{op}^t = \Theta^t - \Theta^1 \qquad \Theta_{dp}^t = \{\Theta^t - \Theta^{t-s} | s = 1, 5, 10\}$$

Concatenate all above features to get finger motion features $\mathcal{F}^t(\mathcal{S}) = [\Theta^t, \Theta_{op}^t, \Theta_{dp}^t]$

- Global Motion Feature

Motivation: to model the global rotation and direction of hand skeleton trajectory

First get the rotation and translation

$$[\mathcal{G}_l, \mathcal{G}_r] = Kabsch(p^t, p_0) \quad \mathcal{G}_r = (r_x, r_y, r_z) \quad \mathcal{G}_l = (\rho, \theta, \phi)$$

Discretize the amplitude

$$\int_0^{\eta_i} g(x)dx = \frac{i}{M}\int_0^{\sigma} g(x)dx \qquad \Phi^t = [\rho_{bin}, \theta, \phi, r_x, r_y, r_z]$$

Motion features

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

$$\Phi_{dp}^t = \{\Phi^t - \Phi^{t-s} | s = 1, 5, 10\} \qquad \mathcal{G}^t(\mathcal{S}) = [\Phi^t, \Phi_{op}^t, \Phi_{dp}^t]$$

EXPERIMENTS

- Datasets
  DHG-14/28 [14]: 14/28 gestures, 20 participants 2800 sequences.
- Self Comparison
  Table 1
- Comparison with State-of-the-arts
  Table 2 & Figure 2

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**

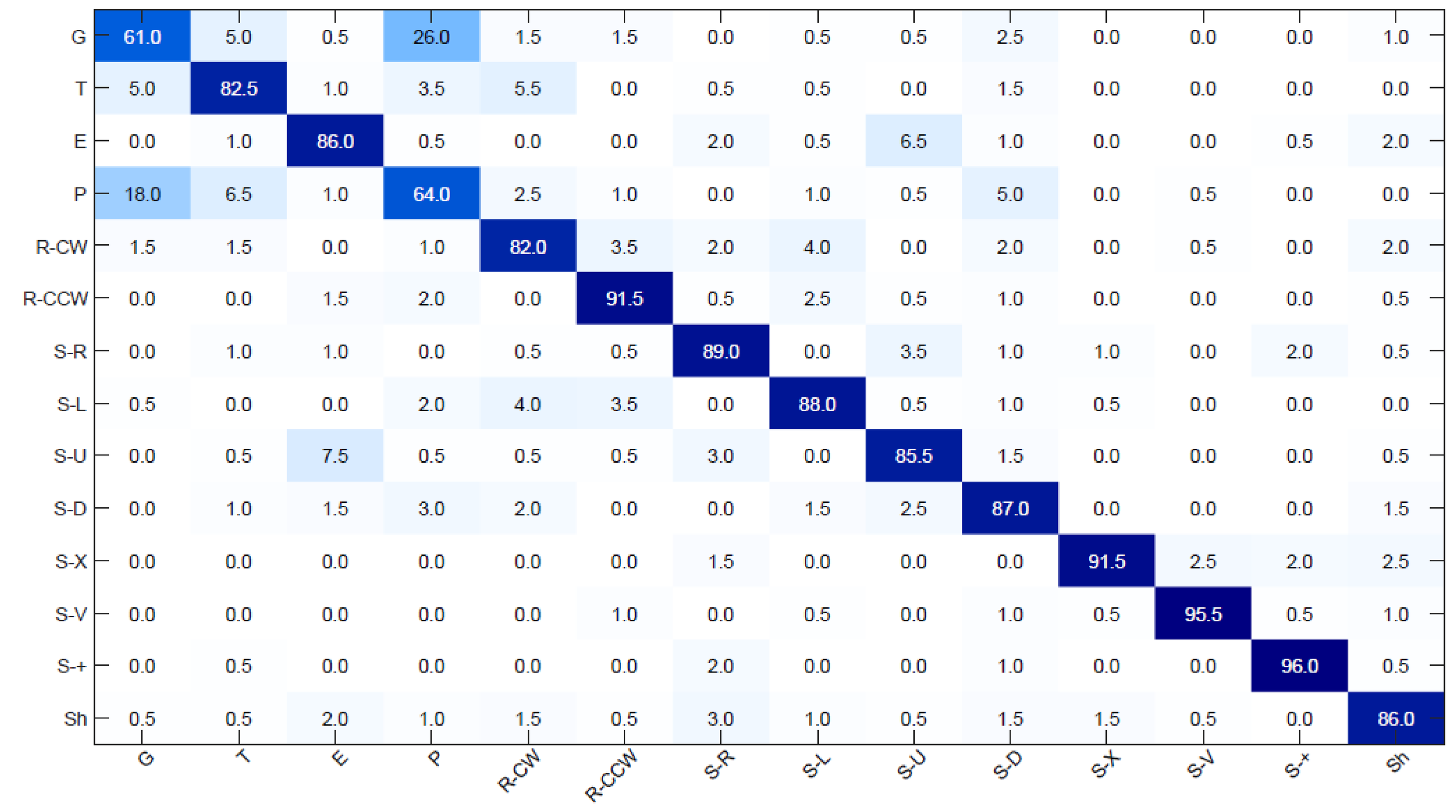

Figure 2: Confusion matrix

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**	$\mathbf{84.68 \pm 6.67}$

REFERENCES

[14] Smedt et al., CVPRW(2016): Skeleton-Based Dynamic Hand Gesture Recognition