Large-scale Multimodal Gesture Recognition Using Heterogeneous Networks

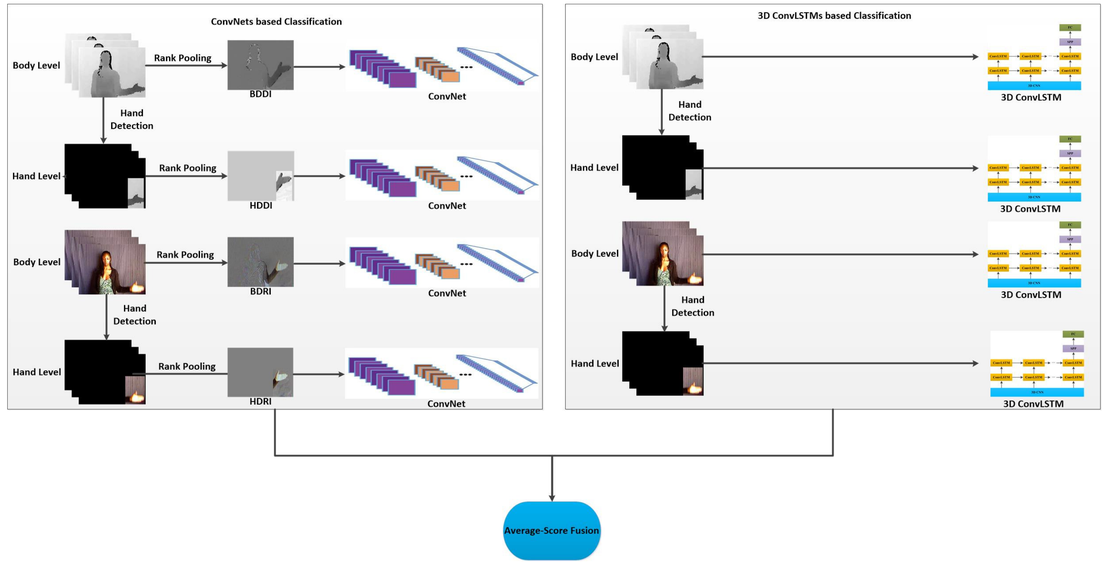
Huogen Wang*[1], Pichao Wang*[2], Zhanjie Song[3], Wanqing Li[4]

ICCV17 International Conference on Computer Vision 2017

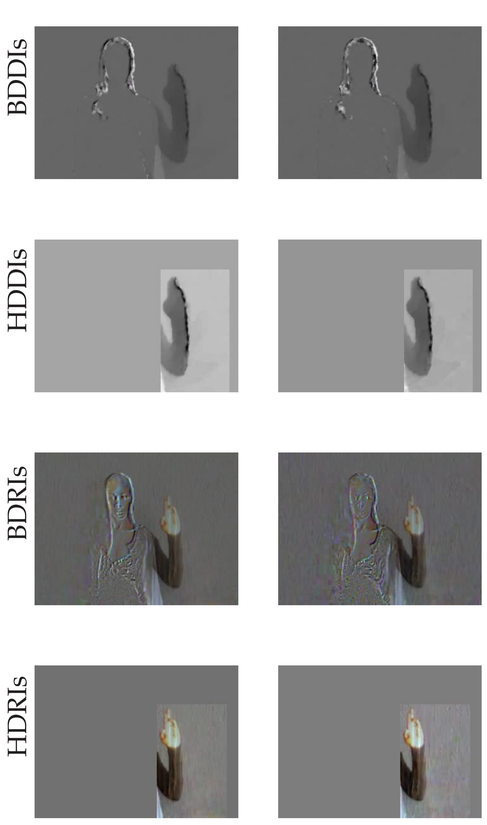
Introduction

This paper presents the method designed for the 2017 ChaLearn LAP Large-scale Gesture Recognition Challenge. The method converts a video sequence into multiple body level and hand level dynamic images as the inputs to ConvNets respectively through rank pooling and adopts ConvLSTM Networks to learn long-term spatiotemporal features from short-term spatiotemporal features extracted using a 3DCNN at body and hand level. Such a heterogeneous network system learns effectively different levels of spatiotemporal features that are complementary to each other to improve the recognition accuracy. The method has been evaluated on the 2017 isolated and continuous ChaLearn LAP Large-scale Gesture Recognition Challenge datasets and the results are ranked among the top performances.

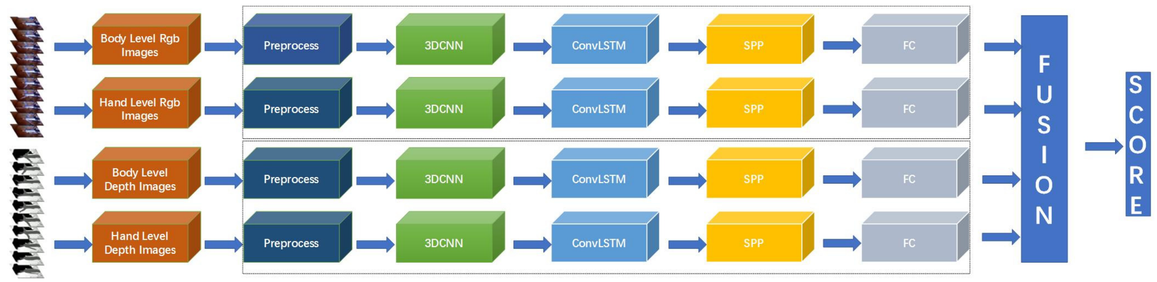
Proposed Method



Rank Pooling



3D ConvLSTM



Challenge Result

The result of ChaLearn LAP Large-scale Isolated Gesture Recognition Challenge.

Rank by test set	Team	Recognition Rate r (valid set)	Recognition Rate r (test set)
1	ASU	**64.40%**	**67.71%**
2	SYSU_ISEE	59.70%	67.02%
3	Lostoy	62.02%	65.97%
4	**AMRL(ours)**	60.81%	65.59%
5	XDETVP	58.00%	60.47%
-	baseline [7]	49.17%	67.26%

The result of ChaLearn LAP Large-scale Continuous Gesture Recognition Challenge.

Rank by testing dataset	Team	Mean Jaccard Index $\overline{Js}$ (valid set)	Mean Jaccard Index $\overline{Js}$ (test set)
1	ICT_NHCI	0.5163	**0.6103**
2	**AMRL(ours)**	**0.5957**	0.5950
3	PaFiFA	0.3646	0.3744
4	Deepgesture	0.3190	0.3164

Evaluation on IsoGD

Methods	Accuracy
Body Level (ConvNet)	49.14%
Hand Level (ConvNet)	50.36%
Hand Level + Body Level (ConvNet)	53.65%
Body Level (3D ConvLSTM)	51.31%
Hand Level (3D ConvLSTM)	48.32%
Hand Level + Body Level (3D ConvLSTM)	53.09%
Body Level (ConvNet+3D ConvLSTM)	57.85%
Hand Level (ConvNet+3D ConvLSTM)	54.67%
All-Score Fusion	**60.81%**

The results of different schemes on validation set are listed.Conclusions: (i) body level and hand level are complementary, as their fusion improves on both; (ii) Score fusion of ConvNet and 3D ConvLSTM greatly improves the final result.

Evaluation on IsoGD

Methods	Accuracy
MFSK [35]	18.65%
MFSK+DeepID [35]	18.23%
Scene Flow [40]	36.27%
Wang et al. [41]	39.23%
Pyramidal C3D [46]	45.02%
Duan et al. [7]	49.17%
Li et al. [24]	49.2%
C3D+ConvLSTM [47]	51.02%
Proposed Method	**60.81%**

Comparison of proposed method with other method on the validation set of ChaLearn LAP IsoGD.

References

[1] B. Fernando, E. Gavves, J. Oramas, A. Ghodrati, and T. Tuytelaars. Rank pooling for action recognition IPAMI,2017

[2] G. Zhu, L. Zhang, P. Shen, and J. Song. Multimodal gesture recognition using 3d convolution and convolutional lstm. IEEE Access,2017

[3] J. Wan, Y. Zhao, S. Zhou, I. Guyon, S. Escalera, and S. Z. Li. Chalearn looking at people rgb-d isolated and continuous datasets for gesture recognition. in CVPR,2016

[4] W. Jun, S. Escalera, A. Gholamreza, H. J. Escalante, X. Bar ÌĄo, I. Guyon, M. Madadi, A. Juri, G. Jelena, L. Chi, and X. Yiliang. Results and analysis of chalearn lap multi-modal isolated and continuous gesture recognition, and real versus fake expressed emotions challenges.