Large-scale Multimodal Gesture Segmentation and Recognition based on Convolutional Neural Networks

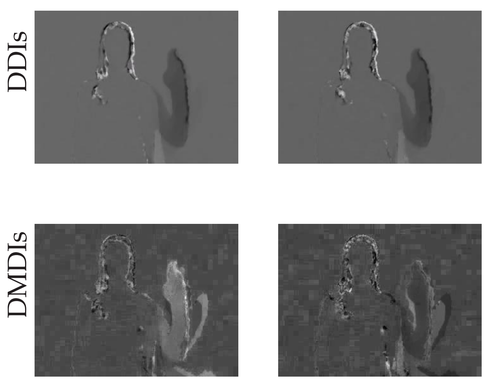
Huogen Wang*[1], Pichao Wang*[2], Zhanjie Song[3], Wanqing Li[4]

ICCV17 International Conference on Computer Vision 2017

Introduction

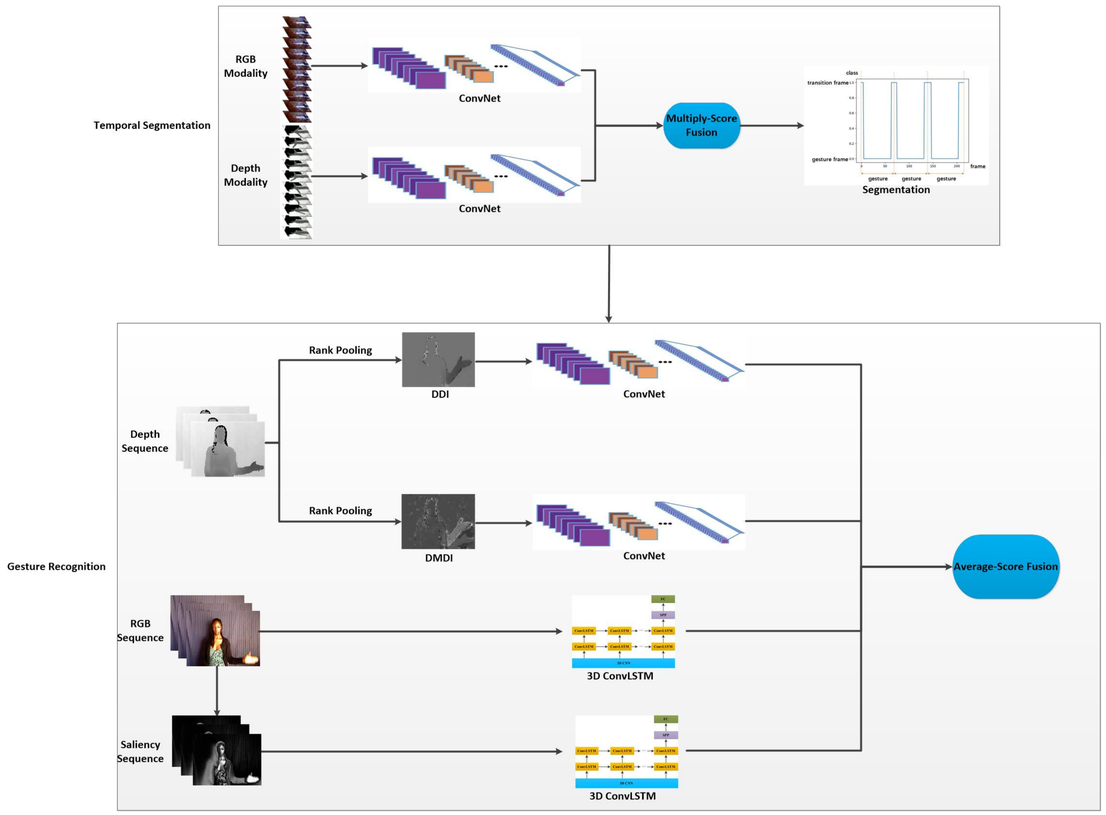
This paper presents an effective method for continuous gesture recognition. The method consists of two modules: segmentation and recognition. In the segmentation module, a continuous gesture sequence is segmented into isolated gesture sequences by classifying the frames into gesture frames and transitional frames using two stream CNNs. In the recognition module, our method exploits the spatiotemporal information embedded in RGB and depth sequences. For the depth modality, our method converts a sequence into Dynamic Images and Motion Dynamic Images through rank pooling and input them to Convolutional Neural Networks respectively. For the RGB modality, our method adopts Convolutional LSTM Networks to learn long-term spatiotemporal features from short-term spatiotemporal features obtained by a 3D convolutional neural network. Our method has been evaluated on ChaLearn LAP Large-scale Continuous Gesture Dataset and achieved the state-of-the-art performance.

Rank Pooling

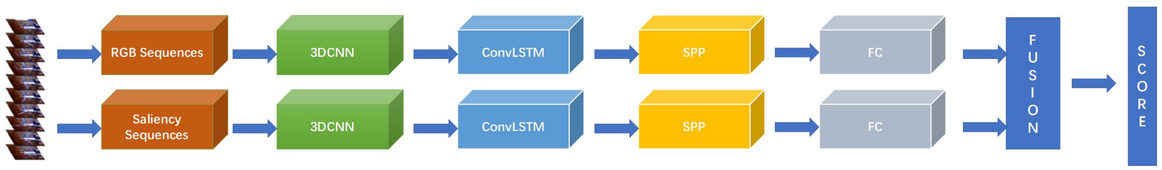


Samples of generated DDIs and DMDIs for gesture Mudra1/Ardhapataka, the left images are dynamic images for forward, the right images are dynamic images for backward. From up to bottom: DDIs and DMDIs. DDIs are constructed from depth sequence. Unlike DDIs, DMDIs are constructed from the absolute differences between consecutive frames through an entire depth sequence.

Proposed Method



3D ConvLSTM

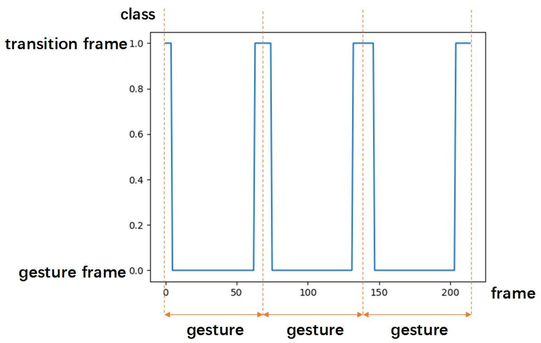


Challenge Result

Rank by test set	Team	Mean Jaccard Index $\overline{J_S}$ (valid set)	Mean Jaccard Index $\overline{J_S}$ (test set)
1	ICT_NHCI	0.5163	**0.6103**
2	AMRL	**0.5957**	0.5950
3	PaFiFA	0.3646	0.3744
4	Deepgesture	0.3190	0.3164
–	Proposed Method	0.5214	0.5307

The result of ChaLearn LAP Large-scale Continuous Gesture Recognition Challenge. Our mean Jaccard Index is 0.5307 in test set. It can be seen that our method is among the top performance.

Temporal Segmentation



An example of the temporal segmentation. The sequence is segemented into three isolated gesture sequences, the middle point of continuous transitional frames is defined as the boundary of two gestures.

Saliency Sequence

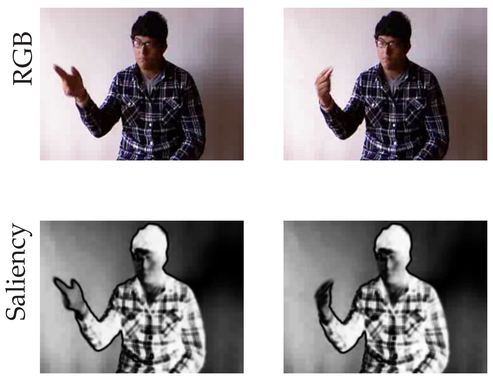


Illustration of a RGB sequence (top) and its saliency sequence (bottom).

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