August 9th, 2015

To: Editor TPAMI

TPAMISI-2015-01-0002 submission -
Deep Dynamic Neural Networks for Multimodal Gesture Segmentation and Recognition.

Dear Guest Editor,

This letter is in response to the review of our submitted manuscript referenced above on multimodal gesture recognition using dynamic deep neural networks. We would first like to thank the reviewers and guest editor for their time and valuable comments. We have taken into careful consideration each one of these comments, and have prepared a detailed response in a separate document adjoint to this letter. We have made this answer as self-contained as possible to facilitate the review process. Furthermore, addressing these comments led to many improvements of the manuscript. Before summarizing the main changes in the paper, we would like to recall the main characteristics and contributions of the paper:

- A Gaussian-Bernoulli Deep Belief Network is proposed to extracts high-level skeletal joint features and the learned representation is used to estimate the emission probability needed to infer gesture sequences;
- A learning framework is proposed to extract temporal features jointly from multiple channel inputs of RGB images and depth images. Because the features are learned from raw 2D images stacked along the 1D temporal domain, we refer our approach as 3D Convolutional Neural Network;
- Intermediate fusion and late fusion are investigated as different strategies to model emission probability within the temporal modeling. Both strategies show that multiple-channel fusions outperform each individual module.

Major modifications: We now would like to summarize the main changes done to the manuscript:

- *Intermediate fusion:* In the previous version for multimodal fusion, we used the late fusion scheme $s = a * s1 + (1 - a) * s2$ where a is chosen by cross-validation. In this revision, we implement an intermediate fusion scheme in Section 4.4.2 that a new-top level perceptron layer is created to combine two models' output as in Fig. 6. The new multimodal neural network's parameters are initialised by the previously trained individual module, taking advantage of different modules intrinsic properties and making the network converge much faster. The intermediate fusion system uses pre-trained weights. The results are reported in Section 4.4.2 and Table 1.

- *Related Works section:* The Related Works section has been moved after the Introduction section. We also follow reviewers' suggestions and include discussions of related works concerning works of: 1) exploiting temporal models in the context of gesture recognition, notably a discriminative hidden-state approach for the recognition of human gestures introduced by Wang *et al.* [1] ; 2) literature for RGBD data using deep learning, which includes the use of recurrent neural networks by Socher *et al.* [2] and applying convolutional neural networks on top of geocentric embedding for depth images by Gupta *et al.* [3]

- *Experimental analysis:* We have included some time analysis in Section 5.4 and visualisation of response maps after learnt filters in Fig. 8. We also gave qualitative remarks on these filter banks. Regarding the quantitative results, we have added more analysis on failure patterns and lessons learnt from the experiments.

- *Explanation of intuition behind higher level presentation of the skeleton features:* We include Section 3.3 to explain the intuition behind higher level representation for skeleton joint features which appeared in our previous CVPR paper but was not included in the previous submission. We think this part is one of major contributions of the paper and inclusion of this section makes the journal paper more self-contained.

Minor modifications: We have addressed all reviewer's comments, most with direct modifications in the paper. We would like to highlight a few relevant points:

- *3D Convolutional Neural Networks*: We clarified accordingly to Rev4 and the readers that the 3^{rd} dimension of the input is indeed the time axis. However, RGB and Depth data are jointly processed by the neural networks, which justifies the multiple-channel naming convention.

- *Parameters and formula interpretation* (cf Rev3): While discussing model formulation, we have added more description and intuitive explanation of unobserved variable H_t. We also corrected and clarified the number of frames assigned to each hidden state.

We hope that these new experiments, clarifications, and paper modifications will satisfy the reviewers as well as address your own comments.

We thank you again for your time and consideration of our manuscript.

Sincerely,

Di Wu, Lionel Pigou, Pieter-Jan Kindermans, Nam Le, Ling Shao, Joni Dambre, and Jean-Marc Odobez

References

[1] S. B. Wang, A. Quattoni, L.-P. Morency, D. Demirdjian, and T. Darrell, "Hidden conditional random fields for gesture recognition," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, vol. 2. IEEE, 2006, pp. 1521–1527.

[2] R. Socher, B. Huval, B. Bath, C. D. Manning, and A. Y. Ng, "Convolutional-recursive deep learning for 3d object classification," in *Advances in Neural Information Processing Systems*, 2012.

[3] S. Gupta, R. Girshick, P. Arbeláez, and J. Malik, "Learning rich features from rgb-d images for object detection and segmentation," in *ECCV*. Springer, 2014.