December 8th, 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 second round review of our submitted manuscript referenced above on “Deep Dynamic Neural Networks for Multimodal Gesture Segmentation and Recognition”. We would first like to thank again the reviewers and guest editor for their diligence and valuable feedback which help us to improve the analysis of the proposed method and overall presentation of the paper. We also appreciate the overall positive outlooks of the revised paper and the improvements with respect to the first version manuscript. 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 document as self-contained as possible to facilitate the review process. Before summarizing the main changes in the manuscript, we would like to recall the main contributions of the paper as follows. Within an HMM framework allowing for the simultaneous gesture recognition and segmentation, we propose:

- A Gaussian-Bernoulli Deep Belief Network with pre-training 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.

- The difference of mean activations in intermediate fusion due to different activation functions is analyzed which is a contribution itself so as to spur further investigation to effectively fuse multi-model, various activations.

Modifications: We now would like to summarize the modifications of manuscript as the following points:

- *Related works section (cf Rev3):* The Related Works section has been enriched following reviewers’ comments by including our analysis of the suggested related works in the context of ChaLearn 2013 competition. This modification emphasizes on the exploitations of HMMs ([1, 2] and RNNs ([3], [2, 4]). We also explain the key differences between the aforementioned papers and the proposed approach is that we use HMM for modelling hidden stats of gesture over the joint feature space whilst their HMM models are purely for audio input [1, 2]. Our proposed system uses DBN with pre-training to learn the skeleton features instead of the hand crafted features [3]. Moreover, we explore the late and intermediate fusion scheme instead of the weighted likelihood that is adopted by [1]. Albeit the intermediate fusion scheme does not outperform late fusion, the discrepancy is a contribution in itself.

- *Extensive proof reading and grammatical correction:* We have corrected the typos and grammatical mistakes according to the reviewers' comments and we have thoroughly proof read the revised manuscript for this revision.
- *Updates of figures:* We have updated and clarified figures in a clearer manner.
- *Forced alignment interpretation (cf Rev1):* : While discussing the model formulation, we have added more description and potential improvement of force alignment scheme in Section 4.1 from the works of speech recognition community [5].

We hope that these new 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] K. Nandakumar, K. W. Wan, S. M. A. Chan, W. Z. T. Ng, J. G. Wang, and W. Y. Yau, "A multi-modal gesture recognition system using audio, video, and skeletal joint data," in *Proceedings of the 15th ACM on International conference on multimodal interaction.* ACM, 2013.

[2] J. Wu, J. Cheng, C. Zhao, and H. Lu, "Fusing multi-modal features for gesture recognition," in *ACM International Conference on Multimodal Interaction*, 2013.

[3] N. Neverova, C. Wolf, G. Paci, G. Sommavilla, G. W. Taylor, and F. Nebout, "A multi-scale approach to gesture detection and recognition," in *Computer Vision Workshops (ICCVW), 2013 IEEE International Conference on.* IEEE, 2013.

[4] 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.

[5] D. Yu and L. Deng, *Automatic Speech Recognition.* Springer, 2012.