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 helped us to improve the analysis of the proposed method and the overall presentation of the paper. We also appreciate the overall positive feedback to the revised paper and the improvements with respect to the first version manuscript. We have taken into careful consideration each of the comments made by the reviewers, and have prepared a detailed response in a separate document. We have made this document as self-contained as possible to facilitate the review process. We would like to summarize the main contributions of the paper as follows. We developed an integrated framework that can simultaneously perform segmentation and recognition of a continuous stream of gestures. To achieve this, we have integrated the following in an HMM framework.

- A Gaussian-Bernoulli Deep Belief Network with pre-training is proposed to extract high-level skeletal joint features and the learned representation is used to estimate the emission probability needed to infer gesture sequences;

- A learning method is proposed to extract temporal features jointly from RGB images and depth images. Because the features are learned from 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 combine the heterogenous input (skeleton data and RGB-D data) to model the emission probability of the HMM compenent. Both strategies show that multimodal fusion outperforms the use of a single feature modality.

- The difference between the mean activations of the different modalities used for intermediate fusion is caused by the different activation functions (non-linearitities). This effect is further analyzed and can be seen as a minor contribution in itself. The main goal of this analysis is actually to stimulate further investigations on how to effectively fuse multi-modal data and networks with various activation functions.

Modifications: In what follows, we give an overview of the main changes made to the manuscript.

- *Extensive proof reading and grammatical corrections:* We have corrected the typos and the grammatical mistakes pointed out by the reviewers. We have paid extra attention to thoroughly proofread the revised manuscript.

- *Related works section (cf Reviewer 3):* We followed the reviewers their advice and included discussions of additional related work. For the ChaLearn2013 competition we discussed the extensive usage of HMMs ([1, 2] and RNNs ([3], [2, 4]). We also explain the key differences between the aforementioned papers and the approach we proposed. A key distinction lies in the fact that we use HMM for modelling hidden states of gestures over the joint feature space whilst their HMM models are purely for audio input [1, 2]. Furthermore, our proposed system uses DBN with pre-training to learn the skeleton features instead of hand crafted features [3]. Moreover, we explore the late and intermediate fusion strategies instead of the basic weighted likelihood that is adopted by [1]. Even though the intermediate fusion scheme does not outperform late fusion, this observation is an interesting contribution in itself.

- *Updates to the figures:* We have updated the figures to make them more clear.

- *Force alignment interpretation (cf Reviewer 1):* : 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 the clarifications and the paper modifications properly address the reviewers' comments as well 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.