[46] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing atari with deep reinforcement learning," *arXiv preprint arXiv:1312.5602*, 2013.
[47] K. Jarrett, K. Kavukcuoglu, M. Ranzato, and Y. LeCun, "What is the best multi-stage architecture for object recognition?" in *Computer Vision, 2009 IEEE 12th International Conference on*. IEEE, 2009, pp. 2146–2153.
[48] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," *arXiv preprint arXiv:1207.0580*, 2012.
[49] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, "On the importance of initialization and momentum in deep learning," in *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, 2013, pp. 1139–1147.
[50] L. Pigou, A. V. D. Oord, S. Dieleman, M. V. Herreweghe, and J. Dambre, "Beyond temporal pooling: Recurrence and temporal convolutions for gesture recognition in video," *CoRR*, vol. abs/1506.01911, 2015. [Online]. Available: http://arxiv.org/abs/1506.01911
[51] D. Wu and L. Shao, "Multimodal dynamic networks for gesture recognition," in *Proceedings of the ACM International Conference on Multimedia*. ACM, 2014, pp. 945–948.
[52] J. Y. Chang, "Nonparametric gesture labeling from multi-modal data," in *European Conference on Computer Vision and Pattern Recognition Workshops*, 2014.
[53] F. Bastien, P. Lamblin, R. Pascanu, J. Bergstra, I. J. Goodfellow, A. Bergeron, N. Bouchard, and Y. Bengio, "Theano: new features and speed improvements," in *Deep Learning and Unsupervised Feature Learning NIPS 2012 Workshop*, 2012.
[54] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

APPENDIX B
REVIEW

B.1 Editor Comments

Associate Editor Comments to the Author: Reviews of the paper have been received. In general comments are positive. Still one reviewer suggests some interesting extra analyses of the proposed method. Please carefully address all reviewers comments for a second review round of the paper. Please provide your revision by July 31th.

Response: Thank you for associate editor's general positive comment of the paper. We also would like to thank reviewers' careful and insightful suggestions for improving the paper. Following are some common points that were mentioned during reviewers' comment and we listed the most notable changes as below:

- Previous version for multimodal fusion, we used the late fusion scheme $s = a * s1 + (1 - a) * s2$ where a is chosen by cross-validation. We implement an early fusion scheme in Section 4.4.2 that a new-top level perceptron layer is created to combine two models' output as in Fig. 10. The new multimodal neural network's parameters are initialised by the previously trained individual module, taking advantage of different module's intrinsic properties and making the network converge much faster. The early fusion system using pre-trained weights. The results are reported in Section 4.4.2 and Table 1.
- The Related Works section has been moved after the Introduction section. We also follow reviewers' suggestions and include discussion of related works concerning with works of: 1) exploiting temporal models in the context of gesture recognition; 2) literatures for RGBD data using deep learning "Wang *et al.* [31] introduced a discriminative hidden-state approach for the recognition of human gestures. However, their discriminative training is limited for pre-segmented gesture sequence and one layer of hidden states might not learn power enough high level representation for larger corpus." "In the field of deep learning from RGBD data, Socher *et al.* [32] proposed a single convolutional neural net layer for each modality as inputs to multiple, fixed-tree RNNs in order to compose higher order features for 3D object classification. The single convolutional neural net layer provides useful translational invariance of low level features such as edges and allows parts of an object to be deformable to some extent. The pooled filter responses are then give to a recursive neural network to learn compositional features and part interactions. Gupta *et al.* [33] proposed a geocentric embedding for depth images that encodes height above ground and angle with gravity for each pixel in addition to the horizontal disparity. The depth image was represented by three channels: horizontal disparity, height above ground and angle with gravity. This augmented representation allows the CNN to learn strong features than by using disparity (or depth) alone. "
- Experimental analysis. We have included some time analysis in Section 5.4 and visualisation of response maps after learnt filters in Fig 8.
- Explanation of intuition behind higher level presentation of the skeleton features. We include section 3.2 to explain the intuition behind higher level representation for skeleton joins features which appeared in previous cvpr paper but didn't include in previous submission. We think this part is one of major contributions of the paper and inclusion of the section makes the journal submission more self-contained.

B.2 Reviewer Comments

B.2.1 Reviewer1

Recommendation: Accept With No Changes

Comments: *The article is easy to read and well structured. The methodology is not strictly novel but its application in the gesture domain with the multimodal fusion makes the article worth reading. Although results are arguably a little behind the maximum performance ones the overall impression of the article is favorable and I believe the community may benefit to check the ideas included in this paper.*

The article proposes a framework for dynamic data augmenting a HMM with deep learning techniques and apply this to gesture segmentation and recognition. Gestures segmentation and recognition is a difficult problem. The

article tackles this difficulty by means of pure data driven approaches similar to the ones used for speech recognition. The particularities of the computer vision domain are handled accordingly.

Response: Thank you for your review and positive outlook of the paper. We are also aware that our results are arguable a little behind the maximum performance, this may due to the network initialisation and multimodal neural network learning. We have included extra experimental analysis and early fusion implementation to further extend the broadness of the paper.

B.2.2 Reviewer2

Recommendation: Revise and resubmit as new

Comments: *In general, the manuscript is well written and is easy to follow. In the given case, it would be preferable to have the "Related work" section right after the introduction, as otherwise paper's contributions are not completely clear. Furthermore, there is certainly a vast literature on exploiting HMMs in the context of gesture recognition (as well as other temporal models, such as recurrent neural networks), which should be briefly summarized, the differences with the proposed solution should be highlighted.*

Response: Thank you for your comments and the recognition of easy readability of the paper. We have rearranged the related work section after the introduction section. Moreover, we have included the discussions of literatures that utilise temporal models, e.g.,"Wang *et al.* [31] introduced a discriminative hidden-state approach for the recognition of human gestures. However, their discriminative training is limited for pre-segmented gesture sequence and one layer of hidden states might not learn power enough high level representation for larger corpus", " [32] proposed a single convolutional neural net layer for each modality as inputs to multiple, fixed-tree RNNs in order to compose higher order features for 3D object classification."

Comments: *Authors claim to learn a model in the joint multi-modal space is a slight overstatement, as neural networks processing different modalities are trained completely independently with following averaging of produced scores.*

Response: Thank you for your comments. Upon receiving the review, we implement the early fusion scheme in Section 4.4.2 that "we adopt another layer of perceptron for cross modality learning taking the input from each individual net's penultimate layer. The parameters of two neural networks (for skeleton and depth) are loaded from the previously trained individual module...The result for early fusion system are reported in Tab. 1. The fusion network is initialise by the pre-trained model and stack with one hidden layer with 2024 hidden unites. We fine-tune the network and the stop the training when validation error rate stop decreasing ($\sim$15 epochs)...However, we can see from Tab. 1 that the early fusion system didn't outperform the late fusion system. The result is counter-intuitive because we expect the early fusion multimodal feature learning will extract semantically meaningful shared representations, outperforming individual modalities, and the early fusion schemes efficacy against the traditional method of late fusion [51]. One possible explanation could be that one individual module has dominant effect on the learning process so as to screw the network towards learning that specific module. The mean activations of the neurons for each modules in Fig. 10 indicate the aforementioned conjecture."

Comments: *State of the art in the experimental section should be mentioned more consistently. For fair comparison, first three lines in Table 3 should be: [39] Deep learning (step 4): skeleton 0.7891, video 0.7990, fusion 0.8449 [39] Deep learning (multiscale): skeleton 0.8080, video 0.8096, fusion 0.8488 [40] 3 sets of skeletal features and HoG: skeleton 0.791, fusion 0.8220 Therefore, it shows that both learning-based and feature extraction-based approaches outperform the proposed method on each modality, as well as on a combination of them. Furthermore, it would be interesting to see how the HMM contributes in the performance in comparison with simple voting based on frame-based predictions.*

Response: Thank for your careful comment. We have amended the result table accordingly. The less than maximum performance could due to the less than ideal settings and initialisations of the neural network. Nonetheless, we would like to argue that one major contribution of the paper is using the learning method for feature extraction and the utilisation of HMM for simultaneous gesture segmentation and recognition. We also present some brief analysis of why the fusion network didn't achieve expected performance gain and hope the experimental analysis could cast some light on the future research directions of the related problems.

Comments: *Visualization of the filter banks (section 3.3.4) in its current state is unnecessary as it does not provide any interesting insights on the interpretation of the learned features. Instead, the poorly formed filters rather indicate undertraining, or lack of training data given the model complexity, or suboptimality of training procedure.*

Response: We include the response maps after filtering for both body and hand part. We observe some interesting properties from the visualisation of the filter banks as in [32] that "Depth images have sharper edges and generally are smoother than the grayscale filters, though the distinctions are less obvious compared with the body versus hand filters."

B.2.3 Reviewer3

Recommendation: Author Should Prepare A Minor Revision

Comments:

In general, I would be more excited if shared representations were learned from the skeleton and the RGB data, as done in multimodal deep learning. This is left for future work. On the positive side, the CNN and DBN are technically sound and the results from their fusion are interesting.

One would expect that the journal version of the paper would be more self-contained and easier to follow than the conference versions, but here I observe the opposite trend. For example, the older conference version [21] explains the intuition behind the higher level representation of the skeleton features, but the journal version does not. The conference paper explains how the coordinate frames are built for the features, while this paper skips this part. The conference paper explains the datasets and visualizes the Viterbi paths better.

Response: Thank you for your careful and positive review. We agree that in this journal version of the paper, some self-contained information has been omitted from the conference paper. Specifically, 1) we include the Problem formation in Sec 3.1 that explains the intuition behind the higher level representation and the advantages offered by feed forward neural over GMMs. 2) we include another section for learning the higher level representation for skeleton joints features in Section 3.2 from a pre-training point of view. The pre-training step is of crucial importance in learning the right initialisation of the deep belief networks using the sigmoid activation function.

Comments: *Section 2 does not help much the reader understand the formulation. For example: "At each time step, we have one observed random variable X_t: explain what these variables represent early (raw skeletan input / RGBD) we have an unobserved variable H_t: describe at a high level the information that the unobserved variables capture, mention examples*

Response: We have include the interpretation part as follows: "A continuous-observation HMM with discrete hidden states is adopted for modelling higher level temporal relationships. At each time step t, we have one observed random variable X_t which can be the skeleton input or depth/RGB input. The unobserved variable H_t taking on values in a finite set $\mathcal{H} = (\bigcup_{a\in\mathcal{A}} \mathcal{H}_a)$, where $\mathcal{H}_a$ is a set of states associated with an individual gesture $\boldsymbol{a}$ by force-alignment. The unobserved variable H_t can be interpreted as a segment of an action $\boldsymbol{a}$. For example, for action sequence "tennis serving", the action sequence can be dissected into $\mathcal{H}_{a_1}, \mathcal{H}_{a_2}, \mathcal{H}_{a_3}$ as: 1) raising one arm 2) raising the racket 3) hitting the ball."

Comments: *The related work section is out of place after the technical sections and before the experiments.*

Response: We have rearranged the related work section after the introduction section.

Comments: *There is no point writing a loop for m=1:2 in Algorithm 1 and 2.*

Response: We have rewritten the Algorithm 1 and 2 and merge the algorithms into a more succinct format.

Comments: *"the number of states ... is chosen as 5": any intuition here?*

Response: Thank you for the comment and this is a very good observation. The number of hidden states is chosen uniformly as 5 in the paper might not be the optimal way of setting the number of hidden states for each gesture. We also experimented segmenting gestures into 10 states and obtained similar result. We reduce the hidden states from 10 to 5 in order to reduce the number of predicting classes and avoiding overfitting. The interpretation of chosen the number of hidden states for Markove Model is similar to choosing the number of hidden states for neural networks: it's more heuristically based. Ideally, we could set the number of hidden states according to the average length of the gesture sequence. But due to time constraint, we didn't train such neural networks.

Comments: *"10 frames are assigned to hidden state ...": why 10?*

Response: Thank you for the careful observation. This is actually a written error due to different number of hidden states used for our experiments. We rewrote the section as: "***Hidden states*** **($\mathcal{H}_a$):** Force alignment is used to extract the hidden states, *i.e.* if a gesture token is 100 frames, the first $20 = \frac{100}{5(N_{\mathcal{H}_a})}$ frames are assigned to hidden state ***1***, the following 20 frames are assigned to hidden state ***2***, and so forth."

Comments: *it is hard to interpret the learned features on Figure 8. There is no intuition what the depth filters capture.*

Response: Because our filter size is 5×5 (smaller filter sizes tends to generalize better, [54] used 3×3 convolution filters), the interpretation will be hard to interpret. However, we observe the similar effect as in paper [32] for depth image filters and we include the following analysis: "Visualisation of the 5×5 filters in the first layer for the different input channels. Interestingly, we observe the same effect as [32] that the resulting filters from depth images have sharper edges which arise due to the strong discontinuities at object boundaries. While the depth channel is often quite noisy most of the features are still smooth."

Comments: *Citations that could be added in the context of deep learning from RGBD data: "Convolutional-Recursive Deep Learning for 3D Object Classification", Socher et al., NIPS 2012*

Response: Thank you for the related works suggestions. And we include in the related works section as follows:"In the field of deep learning from RGBD data, Socher *et al.* [32] proposed a single convolutional neural net layer for each modality as inputs to multiple, fixed-tree RNNs in order to compose higher order features for 3D object classification. The single convolutional neural net layer provides useful translational invariance of low level features such as edges and allows parts of an object to be deformable to some extent. The pooled filter responses are then give to a recursive neural network to learn compositional features and part interactions."

We also observe the same CNN filters as the paper that "one interesting result is that depth channel edges are much sharper. This is due to the large discontinuities between object boundaries and background. While the depth channel is often quite noisy most of the features are still smooth".

Comments: *"Learning Rich Features from RGB-D Images for Object Detection and Segmentation", Gupta et al., ECCV 2014*

Response: We find the works in Gupta *et al.* [33] interesting in a sense that CNN does not necessarily need to train from the raw images, some crafted features may better help network to learn more meaningful, higher level representation. And we have include the related works as follows. "Gupta *et al.* proposed a geocentric embedding for depth images that encodes height above ground and angle with gravity for each pixel in addition to the horizontal disparity. The depth image was represented by three channels: horizontal disparity, height above ground and angle with gravity. This augmented representation allows the CNN to learn strong features than by using disparity (or depth) alone."

Comments: *Another related work is the "Multimodal Deep Learning" by Ngiam et al., ICML 11. I would also like to see some discussion wrt "Hidden Conditional Random Fields for Gesture Recognition", Wang et al, CVPR 2006*

Response: Thank you for suggesting very relevant literatures. We came across both papers. "Multimodal Deep Learning" essentially is the prototype for early fusion model. For Wang *et al.* [31]) observed that one hidden layer is limited for learning larger class corpus. Feature learning for skeleton module is an essential part of this paper and we believe higher level representation is more beneficial for gesture classification. Moreover, the partition function of CRF makes the discriminative training more difficult to train. The similarity in the aforementioned paper with our proposed method is that both methods used hidden layer for learning higher level representation. Recent advancement in feature learning and pre-training for DBN renders our proposed more meaningful. We have included the following in the related works section: "Wang *et al.* [31] introduced a discriminative hidden-state approach for the recognition of human gestures. However, their discriminative training is limited for pre-segmented gesture sequence and one layer of hidden states might not learn power enough high level representation for larger corpus."

B.2.4 Reviewer4

Recommendation: Author Should Prepare A Major Revision For A Second Review

Comments: *Late fusion: my greatest technical concern is that two deep models are trained and then combined with a weighted average: s = a * s1 + (1-a)*s2 where a is chosen by cross-validation. Instead, the authors could combine the two models by creating a new top-level perceptron layer which takes the two models as input. Then this whole structure could be trained jointly with back-propagation. I'd expect results to be (1)at least as good and (2) more philosophically unified.*

Response: We agree with your insightful observation and continue experimenting with the early fusion scheme in this newer submission. One of our previous paper [51] utilized the early fusion scheme for audio and skeleton module for action recognition. We followed that strategy and perform the early fusion scheme using penultimate layer as in Section4.4.2. "However, we can see from Tab. 1 that the early fusion system didn't outperform the late fusion system. The result is counter-intuitive because we expect the early fusion multimodal feature learning will extract semantically meaningful shared representations, outperforming individual modalities, and the early fusion schemes efficacy against the traditional method of late fusion [51]. One possible explanation could be that one individual module has dominant effect on the learning process so as to screw the network towards learning that specific module. The mean activations of the neurons for each modules in Fig. 10 indicate the aforementioned conjecture."

Comments: *The analysis is a bit brief. More experiments and ablative analysis could be added. Specifically, can we interpret the failure patterns of the proposed model(s) and prior work? It would be interesting to see statements like [40] fails more often on gestures of X kind because HOG erases Y useful information or [39] does worse for Z because it handles time at an earlier stage of the pipeline. Then, also giving some qualitative examples of these failures.*

Response: We agree that there is a lack of experiments analysis, especially the failure patterns and lessons learnt from the prior experiment. We have include more analysis in the Experiment and Analysis section as follows: "Examples of overall upper body movement's influence on system performance. Left (score: 0.94) performer almost kept static upper body whilst performing Italian sign language. Right (score: 0.34) performer moved vehemently when performing the gestures.13"

Comments: *These extra experiments (considering joint training of a combined emission probability model) and qualitative interpretation could significantly affect the paper. Overall, the research is solid but needs significantly more work before publication. RCNN: Last, it is entirely possible to train a recurrent neural network to perform Viterbi decoding. This may be difficult (requiring more training data) but would make the entire paper fit into a the deep learning framework. I cannot hold this against the authors, but some discussion might help.*

Response: We have included more qualitative interpretation of the result. We agree that a recurrent neural network could potentially replace the Viterbi decoding part to make the system as a more unified end-to-end system. This, however, may left to the future works.

Comments: *They use a 3d convolutional network. While the introduction makes it sound like this is for multiple-channels (e.g. RGB + Depth), sec. 3.3.2 makes it clear the 3rd dimension is time as the model processes 4 frame subsequences. I think, Fig. 6 could be clearer.*

Response: Thank you for your detailed observation. Yes, the 3rd dimension of the input network is indeed the time axis. However, RGB and Depth data are treat as the two channels during the input phase. We detailed the description of the Figure as follows. "The 3rd dimension of the input is time with 4 frames stacked together. The depth and RGB data are stacked (concatenated) together at Input. Hand and body part output are concatenated at H7."

Comments: *Using RGB-D with deep learning is a common idea, explored by many concurrent works e.g. [A,B,C]. [A] Dosovitskiy, A., Springenberg, J. T., Riedmiller, M., & Brox, T. (2014). Discriminative Unsupervised Feature Learning with Convolutional Neural Networks. arXiv Preprint arXiv: 113.*
[B] Gupta, S., Girshick, R., Arbelez, P., & Malik, J. (2014). Learning Rich Features from RGB-D Images for Object Detection and Segmentation. arXiv Preprint arXiv:1407.5736, 116. doi:10.1007978-3-319-10584-0-23
[C] Socher, R., Huval, B., Bhat, B., Manning, C. D., & Ng, A. Y. (2012). Convolutional-Recursive Deep Learning for 3D Object Classification. Advances in Neural Information Processing Systems

Response: Thank you for suggesting related works. We find the works in Gupta *et al.* [33] interesting in a sense that CNN does not necessarily need to train from the raw images, some crafted features

may better help network to learn more meaningful, higher level representation. We have included the materials using deep learning for RBG-D data with discussions in the related works section: "Gupta *et al.* proposed a geocentric embedding for depth images that encodes height above ground and angle with gravity for each pixel in addition to the horizontal disparity. The depth image was represented by three channels: horizontal disparity, height above ground and angle with gravity. This augmented representation allows the CNN to learn strong features than by using disparity (or depth) alone." "In the field of deep learning from RGBD data, Socher *et al.* proposed a single convolutional neural net layer for each modality as inputs to multiple, fixed-tree RNNs in order to compose higher order features for 3D object classification. The single convolutional neural net layer provides useful translational invariance of low level features such as edges and allows parts of an object to be deformable to some extent. The pooled filter responses are then give to a recursive neural network to learn compositional features and part interactions."