Response to Reviewer 1

We thank the reviewer for his time and comments and positive appreciation of the paper. Below, we provide our answers to his comments and to the unclear points that were raised, and what we have done to clarify the paper and take comments into account. Note that in our answer, all references (Equations, Figures) refer to the new version unless stated otherwise.

Comment: *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 arguably a little behind the maximum performance, and this may be due to the network initialisation and multimodal neural network learning.

Note that the revised version has undergone an important rewriting which hopefully should further improve the clarity of the method description, as well as the DBN and 3DCNN motivation (Section III-C). It also includes extra experimental analysis and an extra intermediate fusion implementation and evaluation to further extend the broadness of the paper.

Response to Reviewer 2

We thank the reviewer for his time and comments. Note that the article has been reworked significantly to better introduce and enhance the modeling elements and their motivation (see Section III-C). In addition, further experiments with an intermediate fusion scheme learning a joint multimodal representation have been conducted. Below, we provide our specific answers to his comments and to the unclear points that were raised, and what we have done to clarify the paper and take comments into account. Note that in our answer, all references (Equations, Figures) refer to the new version unless stated otherwise.

Comment: *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 agree that in this journal version of the paper, some self-contained information had been omitted from the conference paper, and that there were unclear points. We have moved the related work section after the introduction section. Moreover, we have included the discussions of literature that utilise temporal models, e.g."Wang *et al.* [37] introduced a more elaborated discriminative hidden-state approach for the recognition of human gestures. However, relying on only one layer of hidden states, their model alone might not be powerful enough to learn a higher level representation of the data and take advantage of very large corpus. In this paper, we adopt a different approach by focusing on deep feature learning within a temporal model." We also include more literatures discussing the benefits of deep learning using RGB-D data for object detection or classification tasks such as:"Socher *et al.* [39] 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." Regarding your concern about the novelty of each individual technique, we agree that the hybrid combination of HMM and ANN can be traced back to earlier works in continuous speech recognition [33] [32]. However, it is important to point out that the problem we addressed includes not only learning emission probability but also learning features from raw multi-stream of inputs which do not share the same characteristics. To our best knowledge, this problem has not been explored before in the context of gesture recognition. This can be further argued by comparing with other state-of-the-art works which rely on hand-crafted features to some degree [57] [43].

We have significantly reworked the paper and improved its structure to both improve clarity and account for the reviewer's comment.

- Section III now describe the method overview, both in terms of HMM modeling with more intuition about the temporal modeling(Section III-A and III-B), and in terms of Neural network modeling (Section III-C), including with new figures;
- more motivation and intuition behind the use of learned higher level representation and the advantages offered by DBN models for emission probability modeling over the use of GMMs, including the crucial importance of the Gaussian-Bernouilli Restricted Bolzman Machines for pretraining and initializing the deep belief network.
- the experimental part includes a proper description of the dataset, experimental protocol (including performance measures), and more result analysis.
- finally, an intermediate fusion scheme that learns a shared representation has also been implemented and evaluated. Its description is provided in Section V-C.

We hope and believe that this revision is now much more self-contained and will better suit the quality standard of a journal paper.

Comment: *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. In this revision, we implement the intermediate fusion scheme in Section IV-D2 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 results for the intermediate fusion system are reported in Tab. I. The fusion network is initialised by the pre-trained model and stacked with one hidden layer with 2024 hidden unites. We fine-tune the network and stop the training when the validation error rate stops decreasing ($\sim$15 epochs). However, we can see from Tab. I that the intermediate fusion system does not outperform the late fusion system. The result is counter-intuitive because we expect that the intermediate fusion multimodal feature learning would extract semantically meaningful shared representations, outperforming individual modalities, and the intermediate fusion schemes efficacy against the traditional method of late fusion [65]. One possible explanation could be that one individual module has dominant effect on the learning process so as to skew the network towards learning that specific module. The mean activations of the neurons for each module in Fig. 6 indicate the aforementioned conjecture: the large difference between the mean activations of the skeleton module neurons which are predominantly larger than those of the RGB-D ConvNet's (0.57 *vs.* 0.056) can be an indicator of such a bias during the multimodal fine-tuning phase and support this conjecture, even if these mean activations are not directly comparable due to the neuron heterogeneity (the skeleton DBN has logistic units whereas the 3DCNN ConvNet has relu units). Note that such heterogeneity was not present when fusing modalities in [22],where better registration and less spatial registration variability in lip images allowed to also resort to the same stacked RBMs for the visual modality (rather than *3DCNN*) and the audio one. More investigation on how to handle heterogeneous networks should be conducted.

Comment: *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 you for your detailed comments. We have amended the results table accordingly. The less than maximum performance could be 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.

Comment: *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: In the revision, we include the response maps after filtering for both body and hand parts as in Fig. 5. Because our filter size is 5×5 (smaller filter sizes tend to generalize better, [66] used 3×3 convolution filters), their interpretation is indeed harder to visualise, although one can notice that depth filters capture smooth depth transitions, and the combined image and depth filter (see architecture description in Section IV-C2 and Fig. 5) can represent two types of information: segmentation, and edge. We have modified the text to provide better insight as follows:

" *The convolutional filter weights of the first layer are depicted in Fig. 5. The unique characteristics from the kernels are clearly visible: as hand input images (RGB and depth) have larger homogenous areas than the body inputs, the resulting filters are smoother than their body counterpart. In addition, while being smoother overall than the grayscale filters, depth filters exhibit stronger edges, as also reported in [39]. Finally, by looking at the joint depth-image response maps, we can notice that some filters better capture segmentation like information, while other are more edge oriented.* "

Response to Reviewer 3

We thank the reviewer for his time and insightful comments. Below, we provide our specific answers to his comments and to the unclear points that were raised, and what we have done to clarify the paper and take comments into account. Note that in our answer, all references (Equations, Figures) refer to the new version unless stated otherwise.

Comment: *The paper proposes the fusion of the output from a Gaussian-Bernoulli Deep Belief network operating on skeletal features and the output of a Convolutional Neural Network operating on RGBD data to perform gesture segmentation and recognition. The paper advances the field of gesture recognition by using both data sources and deep learning architectures within a Hidden Markov Model chain. The results are improved compared to using either architecture independently.*

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. The paper might not have enough new material to warrant a PAMI publication wrt to the previous conference versions. I also find that it is not that well written for a journal paper (see below). 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 had been omitted from the conference paper, and that there were unclear points. We have significantly reworked the paper and improved its structure to both improve clarity and account for the reviewer's comment:

- a related works section has been added;
- Section III now describe the method overview, both in terms of HMM modeling (Section III-A and III-B), and in terms of Neural network modeling (Section III-C), including with new figures;
- more intuition about the temporal modeling (Section III-A);
- more motivation and intuition behind the use of learned higher level representation and the advantages offered by DBN models for emission probability modeling over the use of GMMs, including the crucial importance of the Gaussian-Bernouilli Restricted Bolzman Machines for pretraining and initializing the deep belief network.
- the experimental part includes a proper description of the dataset, experimental protocol (including performance measures), and more result analysis.
- finally, an intermediate fusion scheme that learns a shared representation has also been implemented and evaluated. Its description is provided in Section V-C. However, to the contrary of what we had expected, it did not perform better than the late fusion scheme.

We hope and believe that this revision is now much more self-contained and will better suit the quality standard of a journal paper.

Comment: *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 / RGB-D) we have an unobserved variable H_t: describe at a high level the information that the unobserved variables capture, mention examples*

Response: We have considerably reworked Section III to provide an overview of the method along with intuition and motivations. Please check the new version. More specific parts of this section addressing your more specific comments are provided below.

Regarding the variables:

" *A continuous-observation HMM is adopted for modelling higher level temporal relationships. At each time step t, we have one observed random variable X_t composed of the skeleton input X_t^s and RGB-D input images X_t^r as shown in the graphical representation in Fig. 1. The hidden state variable H_t takes on values in a finite set $\mathcal{H}$ composed of $N_{\mathcal{H}}$ states related to the different gestures. The intuition motivating the HMM model is that a gesture is composed of a sequence of poses where the relative duration of each pose may vary. This variance is captured by allowing flexible forward transitions within a Markov chain. In practice, H_t can be interpreted as being in a particular phase of a gesture* **a**. "

Or related to more concrete example (Section III-B abut Markov state diagram):

" *For each given gesture $a \in \mathcal{A}$, a set of states $\mathcal{H}_a$ is introduced to defined a Markov model of that gesture. For example, for action sequence "tennis serving", the action sequence can implictly be dissected into $h_{a_1}, h_{a_2}, h_{a_3}$ as: 1) raising one arm 2) raising the racket 3) hitting the ball.* "

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

Response: We have now written a proper related work section after the introduction section.

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

Response: The Algorithm description has been removed from the paper. We have privileged a more structured and more textual description of the method, and thus needed to remove the Algorithms for space reason. Nevertheless, we believe that given the method overview in Section III, and the more detailed elements in Section IV, the method steps should be fairly easily understandable.

Comment: *"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 main intuition behind this number was that a gesture is often composed of 5 phases: 1) an onset; 2) arm motions to reach 3) a more static pose (often characterized by a distinct hand posture); and a4) motion back to 5) stop in the rest pose. However, we agree that this number might not be optimal, and that different gestures could have different number of states. Also, from a more heuristic point of view, note that we had performed experiments with 10 states per class, and that it performed similarly.

To account for the reviewer's comment we have updated the text as follows:

"*Note that intuitively, 5 states represents a good granularity as most gestures in the Clalearn are composed of 5 phases: an onset, followed by arm motions to reach a more static pose (often characterized by a distinct hand posture), and the motion back to the rest place. In the future, optimal section of this number*[6] *and of different number of states per gesture could be investigated.* "

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

Response: Thank you for the careful observation. This is actually a written error. The corrected tet reads now:

[6]Experiments with 10 states led to similar performance.

"*To do so, a force alignment is used which means that if the i^{th} sequence is a gesture* **a***, then the first $\lfloor \frac{T_i}{5} \rfloor$ frames are assigned to state $h^1_{\mathbf{a}}$ (the first state of gesture* **a***), the following $\lfloor \frac{T_i}{5} \rfloor$ frames are assigned to $h^2_{\mathbf{a}}$, and so forth.* "

Comment: *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 tend to generalize better, [66] used 3×3 convolution filters), their interpretation is indeed harder to visualise, although one can notice that depth filters capture smooth depth transitions, and the combined image and depth filter (see architecture description in Section IV-C2 and Fig. 5) can represent two types of information: segmentation, and edge. We have modified the text to provide better insight as follows:

" *The convolutional filter weights of the first layer are depicted in Fig. 5. The unique characteristics from the kernels are clearly visible: as hand input images (RGB and depth) have larger homogenous areas than the body inputs, the resulting filters are smoother than their body counterpart. In addition, while being smoother overall than the grayscale filters, depth filters exhibit stronger edges, as also reported in [39]. Finally, by looking at the joint depth-image response maps, we can notice that some filters better capture segmentation like information, while other are more edge oriented.* "

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

Response: Thank you for the suggested citations. We find those work interesting, e.g. for Gupta *et al.* [40], as it shows that CNN do not necessarily need to be trained from the raw images, and some handcrafted features may better help the network to learn more meaningful, higher level representations. We have added these references in the paper in the following way:

" *However, the advent of Kinect-like sensors has put more emphasis on RGB-D data for gesture recognition, but not only. For instance, the benefits of deep Learning using RGB-D data have been explored for object detection or classification tasks. Socher* et al. *[39] 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. To address object detection, Gupta* et al. *[40] 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.* "

Comment: *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 works. We came across both papers. "Multimodal Deep Learning" essentially is the prototype for an intermediate fusion model. Regarding Wang *et al.* [37]), the similarity with our proposed method is that both methods use a/several hidden layer for learning higher level representations. However, authors in Wang *et al.* [37]) observed that one hidden layer is limited for learning a larger class corpus. In our case, we believe that higher level representation learning with more layers, which is an essential part of our paper, is very important for gesture classification. Recent advancement in multi-layer feature learning and pre-training for DBN renders our proposed method more meaningful. We have included the references as follows:

For Ngiam et al, as it does not relate to gesture recognition, we have cited this method in the introduction.

" *Multimodal deep learning technique were also investigated [22] to learn cross-modality representation, for instance in the context of audio-visual speech recognition.* "

However, as their method is very similar to the intermediate fusion scheme we have now implemented, we have added the following in Section IV-D2:

" *Note that this is very similar to the approach proposed in [22] for audio-visual speech recognition. An important difference is that in [22], the same stacked RBMs/*DBN *architecture was used to represent both modalities before fusion, whereas in our case, a stacked RBMs/*DBN *and a* 3DCNN *are used. Also, [22] proposed the use of a multimodal autoencoder to handle predictions when potentially only one modality migth be present, a point that we do not address.* "

For Wang et al:

" *Wang* et al. *[37] introduced a more elaborated discriminative hidden-state approach for the recognition of human gestures. However, relying on only one layer of hidden states, their model alone might not be powerful enough to learn a higher level representation of the data and take advantage of very large corpus. In this paper, we adopt a different approach by focusing on deep feature learning within a temporal model.* "

Response to Reviewer 4

We thank the reviewer for his time and insightful comments. Below, we provide our specific answers to his comments and to the unclear points that were raised, and what we have done to clarify the paper and take comments into account. Note that in our answer, all references (Equations, Figures) refer to the new version unless stated otherwise.

Comment: *The paper purports 3 contributions. (1) the authors use deep learning to estimate emission probabilities for a HMM predicting gesture. (2) 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 sub-sequences. I think, Fig. 6 could be clearer. (3) Emission probability models are trained for both the skeletal data and depth data. They are then averaged (weighted) and used in an HMM.*

Overall, I am convinced this paper solves the problem of gesture recognition with a novel combination of techniques. However, I am not convinced (1) any of the technical techniques themselves are particularly novel nor (2) that the chosen combination is the right one. Finally, (3) the results aren't particularly impressive (only matching state-of-the-art). Moreover, I have technical/philosophical objections which I'll elaborate on in the comments.

Learning to model HMM emission and transition parameters is an old idea (going back decades, to at least the well known Baum-Welch algorithm) and 3D convolutional networks for video were explored by [11].

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: 1-13. [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:10.1007/978-3-319-10584-0 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 25, (i), 665-673.

Response: Thanks for the comments. Note that we have significantly reworked the paper and improved its structure to both improve clarity and make the paper more self-contained, while taking into account reviewers' comments. Regarding your specific issues above, we can state the following and have made the following changes:

- Regarding your concern about the novelty of each individual technique, we agree that the hybrid combination of HMM and ANN can be traced back to earlier works in continuous speech recognition [33] [32]. However, it is important to point out that the problem we addressed includes not only learning emission probability but also learning features from raw multi-stream of inputs which do not share the same characteristics. To our best knowledge, this problem has not been explored before in the context of gesture recognition. This can be further argued by comparing with other state-of-the-art works which rely on hand-crafted features to some degree [57] [43].

  In response to your question if we have chosen the right combination, it is true that there are several techniques appropriate for each model, some of which are presented in Related Works. Nonetheless, given the limited scope of the paper we focus on the combination which has the most potential. For each module, we can justify our choice of techniques as following: First, at the feature learning stage, we have to deal with 3 input streams: skeleton input, RGB images, and depth images. RGB images and depth images share correlated spatial information. We believe it is the right choice to learn features jointly by combining these two streams as a single 2-channel input. By expanding the CNN into temporal domain, the learnt features extracted by 3DCNN not only describe high level visual but also dynamic information about movements. On another hand, skeleton input appears to be sparser but contains more robust information. DBN architecture initialsed with Gaussian-Bernoulli RBM can exploit high level correlation among the set of upper body joints and the learnt features

also can be compatible with features learnt from 3DCNN. Secondly, at the temporal modeling stage, HMM is a clearly a very good candidate given its pros in simultaneous segmentation and inference. Therefore, we followed the previous works in speech recognition and embedded our dynamic deep neural networks in HMM. With the points noted out, we believe our paper has sufficient novelty and contributions, and very much deserves publication in a TPAMI devoted to the gesture recognition task.

- Regarding the 3DCNN. It was not our intention to confuse the reader. We have made the text more explicit in the introduction:

  " *A 3D Convolutional Neural Network is proposed to extract features from 2D multiple channel inputs like depth and RGB images stacked along the 1D temporal domain;* "
  and similarly, when presenting the 3DCNN in section Section IV.C.2

  " *The 3D convolution itself is achieved by convolving a 3D kernel to the cuboid formed by stacking multiple contiguous frames together.* "
  We have also updated the 3DCNN figure (now figure 4 in the paper), specifying the input, intermediate layers and their corresponding modalities more explicitly, which we hope is not ambiguous.
- multimodal fusion: we have now implemented and evaluated a more unified neural network to this end. See the answer to your next comment.
- Related work. Thank you for suggesting very relevant works using RGB-D with deep learning. We have included them in the related litterature as follows:

  " *However, the advent of Kinect-like sensors has put more emphasis on RGB-D data for gesture recognition, but not only. For instance, the benefits of deep learning using RGB-D data have been explored for object detection or classification tasks. Dosovistskiy* et al. *[38] presented a generic feature learning for training a convolutional network using only unlabeled data. In contrast to supervised network training, the resulting feature representation is not class specific and are advantageous on geometric matching problems, outperforming the SIFT descriptor. Socher* et al. *[39] 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. To address object detection, Gupta* et al. *[40] 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. This augmented representation allows CNN to learn stronger features than when using disparity (or depth) alone.* "

Comment: *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 actually believe that this should improve the system. We thus implemented and evaluated such an intermediate fusion scheme in this revision. However, this approach did not improve the results over the late fusion scheme, providing similar results.

To account for this new model, the text was updated as follows in the model description and the result analysis. In Section IV-D2, describing the Intermediate fusion:

" *As an alternative to the late fusion scheme, we can take advantage of the high-level representation implicitly learned by each module (and represented by the V^s and V^r nodes of the penultimate layer of the respective networks, before the softmax) to fuse the modality in an intermediate fashion by concatenating these two layers in one layer of 2024 hidden unites and learning a cross-modality emission probability predictive network. Note that this is very similar in spirit to the approach proposed in [22] for audio-visual speech recognition. An important difference is that in [22], the same stacked RBMs/*DBN *architecture was used to represent both modalities before fusion, whereas in our case, a stacked RBMs/*DBN *and a* 3DCNN *are used. Also, [22] proposed the use of a multimodal autoencoder to handle predictions when potentially only one modality migth be present, a point that we do not address.*

The resulting architecture is trained by first initializing the weights of the deeper layers from the previously trained module, and then jointly fine tuning the whole network (including learning the last layer parameters) and stop the training when the validation error rate stops decreasing (∼15 epochs). We argue that using the "pre-trained" parameters is important due to the heterogeneity of the inputs of the system, and that the joint training should adjust parameters to handle this heterogeneity and produce the final estimates. "

A specific section in the analysis is devoted to the result analysis, which reads:

"

Late vs. Intermediate fusion. *The results in Tab. I and II show that the intermediate fusion system improved individual modalities, but without outperforming the late fusion system. The result is counter-intuitive, as we would have expected the cross-modality learning in the intermediate fusion scheme to result in better emission probability predictions, as compared to the simple score fusion in the late system. One possible explanation is that the independance assumption of the late scheme better preserves both the complementarity and redundancy of the different modalities, properties which are important for fusion. Another related explanation is that in the intermediate fusion learning process, one modality may dominate and skew the network towards learning that specific module and lowering the importance of the other one. The large difference between the mean activations of the skeleton module neurons which are predominantly larger than those of the RGB-D ConvNet's (0.57* vs. *0.056) can be an indicator of such a bias during the multimodal fine-tuning phase and support this conjecture, even if these mean activations are not directly comparable due to the neuron heterogeneity (the skeleton DBN has logistic units whereas the 3DCNN ConvNet has relu units). Note that such heterogeneity was not present when fusing modalities in [22], where better registration and less spatial registration variability in lip images allowed to also resort to the same stacked RBMs for the visual modality (rather than* 3DCNN*) and the audio one. More investigation on how to handle heterogeneous networks should be conducted.* "

Comment: *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 experiments. We have included more analysis in the Experiment and Analysis, see th whole section V.C. Below we provide the text related to major changes:

(1) discussion of confusion matrices:

" *The confusion matrices (in log-form) in Fig. 9 better illustrate the complementarity of the behaviors of the two modalities.*

The higher underdetection of RGB-D is clearly visible (whiter matrix, except last 'undetected' column). We can also notice that some gestures are more easily recognized than others, or catch the difficult instances of other gestures. This is the case of the "Basta" gesture, whose arms motion resembles the start and end of the arm motion of many other gesture (see Fig. 7). Whatever the modality, its model thus tends to recognize few instance of all other gesture classes, whenever their likelihood are low when being evaluated using the HMM states associated with their true label due to too much variability. Similarly, the hand movement and pose of the "Buenissimo" gesture is present in several other gesture classes, whose instances are then often confused with "Buenissimo" when relying on the skeleton information alone. However, as these gestures differ primarily in their hand pose, such confusion is much more reduced using the RGB-D domain, or when fusing the skeleton and RGB-D modules. The complementary properties of the two modalities is also illustrated from the Viterbi path decoding plot in Fig. 8. In general, the benefit of this complementarity between arm pose/gesture and hand pose can be observed from the whiter confusion matrix than in the skeleton case (less confusion due to hand pose information from RGB-D) and much less under-detection than in the RGB-D case (better upper-body pose discrimination thanks to skeleton input).

However, the modalities by themselves have more difficulties to correct the recognition errors which are due to variations coming from the performer, like differentiating people that gesticulate more (see Fig. 11). "

(2) discussion of late and intermediate feature fusion (see answer to your previous comment).

(3) analysis of the temporal modelling benefits:

"

HMM benefit. *As the emission probabilities are learned in a discriminative manner, one could wonder whether the HMM brings benefit beyond smoothing. To investigate this issue, we removed the temporal structure as follows: for a given gesture* $\mathbf{a}$*, we computed its score at time* t*,* $Score(\mathbf{a}, t)$*, by summing the emission probabilities* $p(X_t|H_t = h)$ *for all nodes associated to that gesture,* i.e. $h \in \mathcal{H}_\mathbf{a}$*. This score is then smoothed in the temporal domain (using a window of 5 frames) to obtain* $\widehat{Score}(\mathbf{a}, t)$*. Finally, following [57], the gesture recognition proceeds in two steps: first finding gesture segments by thresholding the score of the ergodic state; then, for each resulting gesture segment, the recognized gesture is defined as the one whose average score within the segment is the highest. Fig. 10 illustrates this process along with the DDNN and ground-truth. In general, we could observe that better decisions on the presence of gestures and on the boundaries where a gesture starts and ends are achieved with the proposed* DDNN *thanks to the use of the state-diagram defined in Fig. 2, as compared to the above method, where deciding on a gesture detection threshold is rather unstable and quite sequence dependent. Indeed, the overall performance of the above scheme without the HMM temporal sequencing is reduced to* $JI = 0.66$*, while the* Recognized*,* Confused *and* Missed *corresponding to Table II for the test set are* 76.6 *,* 5.3 *and* 18.1*. However, note that the above method relying on only the gesture probability learned using neural networks on 5 frame inputs still outperforms the Jaccard index of* 0.413 *obtained by [58] when using a 5 frames template matching system where all the features are handcrafted.* "

(4) better understanding of the challenges of the dataset caused by different performers' body movement, as illustrated in Fig. 11:

" *Most performers tend to keep their upper-body static while performing the gesture, leading to good recognition performance (Jaccard index of person on the top is 0.95 for the late fusion system). Some persons are more involved and move more vehemently (person at the bootom, Jaccard index of 0.61), which can affect the recognition algorithm itslef (bottom left samples) or even the skeleton tracking (bottom right; note that normally cropped images are centered vertically on the head position).*

Examples of overall upper body movements influence on the system performance. Left (score: 0.94) performer almost kept a static upper body whilst performing Italian sign language. Right (score: 0.34) performer moved vehemently when performing

the gestures.11 "

Comment: *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 developed such a joint training and included more qualitative interpretation of the results. 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 be left to the future work, and have included the following in the conclusion:

" *In addition, while the proposed HMM provided a good basis for the temporal modeling of gestures, other more discriminant temporal approaches such as Conditional Random Field or further and better variants [37] could be directly exploited at their advantage in conjuction with our deep neural network learning approach. Ultimately, in a logical way, these two research directions converge into the investigation of a single and unified deep learning framework fusing heterogeneous modalities by using recent Recurrent Neural Networks such as Long Short Term Memory [64] for modelling the temporal component of the problem.* "

Comment: *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 sub-sequences. I think, Fig. 6 could be clearer.*

Response: See earlier our answer to this issue.