Multimodal Dynamic Networks for Gesture Recognition

Di Wu, Ling Shao
Department of Electronic and Electrical Engineering
The University of Sheffield, Sheffield S1 3JD, UK.
{elp10dw,ling.shao}@sheffield.ac.uk

ABSTRACT

Multimodal input is a real-world situation in gesture recognition applications such as sign language recognition. In this paper, we propose a novel bi-modal (audio and skeleton joints) dynamic network for gesture recognition. First, state-of-the-art dynamic Deep Belief Networks are deployed to extract high level audio and skeletal joints representations. Then, instead of traditional late fusion, we adopt another layer of perceptron for cross modality learning taking the input from each individual net's penultimate layer. Finally, to account for temporal dynamics, the learned shared representations are used for estimating the emission probability to infer action sequences. In particular, we demonstrate that multimodal feature learning will extract semantically meaningful shared representations, outperforming individual modalities, and the early fusion scheme's efficacy against the traditional method of late fusion.

Categories and Subject Descriptors

H.1.2 [**User/Machine Systems**]: Human information processing; I.2.14 [**Artificial Intelligence**]: [video analysis, motion]

General Terms

Algorithm, Experimentation

Keywords

Gesture Recognition, Human-Computer Interaction, Multimodal Fusion, Deep Belief Networks

1. INTRODUCTION

Gesture recognition has been a popular research field in recent years due to its promising application prospects in human-computer interaction. In the early days of gesture recognition research, most approaches were controller-based, in which users had to wear or hold certain hardware for motion data capturing. In vision-based approaches, usersŠ motion data are captured by cameras and numerous computer vision methods have been successfully adopted into this area for further data analysis and understanding. Over the last few years, with the immense popularity of the Kinect, there has been renewed interest in developing methods for human gesture and action recognition from both 3D skeletal data and audio data captured synchronously by the device.

Deep learning is an emerging field of machine learning focusing on learning representations of data and has recently found success in a variety of domains, from computer vision to speech recognition, natural language processing, web search ranking, and even online advertising. The ability of deep learning methods to capture the semantics of data is, however, limited by both the complexity of the models and the intrinsic richness of the input to the system. In particular, current methods only consider a single modality leading to an impoverished model of the world. Sensory data are inherently multimodal instead: images are often associated with text; videos contain both visual and audio signals; text is often related to social content from public media; etc. It is expected that the cross-modality structure may yield a big leap forward in machine understanding of the world.

Multimodal learning involves relating information from multiple sources. For example, images and depth scans are correlated at first-order as depth discontinuities often manifest as strong edges in images. Conversely, audio and visual data for gesture recognition have correlations at a "mid-level", as phonemes and joints motions; it can be difficult to relate joint spatio-temporal information to audio waveforms or spectrograms. Learning from multimodal inputs is technically challenging because different modalities have different statistics and different kinds of representations. For instance, text is discrete and often represented by very large and sparse vectors, while images are represented by dense tensors that exhibit strong local correlations. Traditional multi-agent systems tend to adopt the late fusion scheme by normalizing the confident values from an individual modality for final prediction, ignoring the subtle intrinsic properties within different modalities. Fortunately, deep learning has the promise to learn adaptive representations from the input, potentially bridging the gap between these different modalities.

In this paper, a novel framework of bimodal dynamic networks is proposed for continuous gesture recognition given 3D joint positions and the audio utterance of the gesture tokens. We focus on data driven analysis of acyclic audio-

MM'14, November 3–7, 2014, Orlando, Florida, USA.

video sequence labeling problems.

Problem formulation: Given a multimodal input sequence $\mathbf{X^m} = \{\mathbf{x_1^m}, \mathbf{x_2^m}, \ldots, \mathbf{x_t^m}\}$, where $\mathbf{m}$ is the modal index (in our experimentm=$\mathbf{2}$ because we only use audio and skeleton inputs), instead of finding the global label $\mathbf{Y}$ directly, we dissect the problem into finding the individual $\{\hat{\mathbf{y_1}}, \hat{\mathbf{y_2}}, \ldots, \hat{\mathbf{y_t}}\}$ and reasoning with a higher level Markov field to obtain the most likely label $\hat{\mathbf{Y}}$.

Contributions:

i) Relying on a pure learning approach, all the knowledge in the model comes from the data without sophisticated preprocessing or dimensionality reduction via manifold learning methods. The proposed feed forward neural networks offer several potential advantages as a better estimator for emission probabilities of the Markov field over the traditional paradigms (*e.g.* Gaussian mixture models) because its estimation of the posteriori probabilities does not require detailed assumptions about the data distribution. An advantage of a fully-automatic learning-based method is that it incorporates the feature learning and classification procedures in a unified framework by minimizing the energy (*i.e.* optimizing the object function). Therefore, the proposed framework is more adaptable to different object functions or different input sensory modalities.

ii) Instead of brutally flattening a sequence as in [7] where a fixed number of input frames is required, we employ a dynamic time programming scheme, dissecting the problem into modeling frame-based emission probability. The system is scalable to various time length sequences and is easily adapted for simultaneously segmenting and recognizing video sequences, discovering anchor points.

iii) The multimodal framework learns the shared representation from multiple high level feature representations. Experimental results show the gains over late fusion methods and it opens the door for an early fusion fashion of generative time series modeling.

The model has been designed with bi-modal gesture recognition in mind, but should lend itself well to other multimodal high-dimensional time series.

2. ARCHITECTURE

The overall architecture of our proposed model is shown in Figure 1. The individual emission probability estimators are based on the state-of-the-art architectures as in [5, 12]. Specifically, a Deep Belief Network is deployed for each modality to estimate the output emission probability. Because both feature modalities $\mathbf{X^m}$ are continuous instead of binomial features, the first visible layer is a Gaussian Restricted Boltzmann Machine to model the energy term:

$$E(v,h;\theta) = -\sum_{i=1}^{D} \frac{(v_i - b_i)^2}{2\sigma_i^2} - \sum_{i=1}^{D}\sum_{j=1}^{F} W_{ij} h_j \frac{v_i}{\sigma_i} - \sum_{j=1}^{F} a_j h_j \tag{1}$$

where model parameters $\theta = \{W, b, a, \sigma\}$ with W_{ij} representing the symmetric interaction term between visible unit i and hidden unit j while b_i and a_j are their bias terms with visible unit variance σ. D and F are the numbers of visible and hidden units.

The outputs of the neural net are the hidden states learned by force alignment during the supervised training process. Once each individual modality is trained, the penultimate layer is extracted and fused for the shared representation. Then, the standard backpropagation can be adopted for adjusting the weight W^m for each modality m:

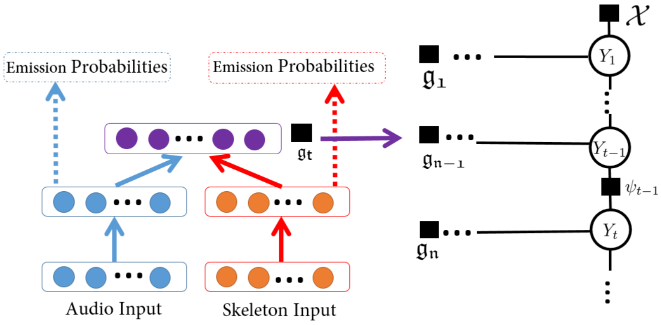

Figure 1: Architecture of the multimodal dynamic networks: each modality (audio or skeleton input) is first pre-trained by a Deep Belief Network, and their penultimate layers are fused together to generate a shared representation for dual modalities. The outputs are the emission probabilities $\mathfrak{g}_t$ for temporal dynamic modeling. In our experiments, we assume each of the conditional distributions per frame is independent of all previous observations except the most recent, hence the higher level is specified as a Hidden Markov Model.

$$W^m = W^m - \alpha \frac{\partial}{\partial W^m} J(W) \tag{2}$$

where α is the annealed learning rate and $J(W)$ is the cost function (cross-entropy) of the last layer perceptron by feed forwarding the fused penultimate layers from mutli-modalities. The output of the network is $p(X_t|Y_t)$ denoting the fused emission probability and is denoted by $\mathfrak{g}_t$ in Figure 1.

Assuming each of the conditional distributions is independent of all previous observations except for the most recent, the full probability model is now specified as HMM:

$$p(Y_{1:T}, X_{1:T}) = p(Y_1)p(X_1|Y_1)\prod_{t=2}^{T} p(X_t|Y_t)p(Y_t|Y_{t-1}), \tag{3}$$

where $p(Y_1)$ is the prior on the first hidden state and in all our experiments, we have a uniform prior, and $p(H_t|H_{t-1})$ is the transition dynamic model. We can infer the action presence in a new sequence by Viterbi decoding as:

$$V_{t,\mathcal{Y}} = P(Y_t|X_t) + \log(\max_{\mathcal{Y}\in\mathcal{Y}_a}(V_{t-1,\mathcal{Y}})) \tag{4}$$

where initial state $V_{1,\mathcal{Y}} = \log(P(Y_1|X_1))$. From the inference results, we define the probability of an action $a \in \mathcal{A}$ as $p(y_t = a|x_{1:t}) = V_{T,\mathcal{Y}}$.

Related Work

The proposed framework is mostly related to the works of [7, 10] in that, instead of the traditional late fusion, all resort to an early fusion scheme. In [7], the input sequences are treated as a holistic entity, hence, the method is not adaptable to various time length input. A multimodal Deep Boltzmann Machine is introduced in [10] to learn a good generative model of the joint space of image and text inputs for information retrieval.

In order to model the time series data, the unimodal of our architecture is built upon the framework of [5, 12] which

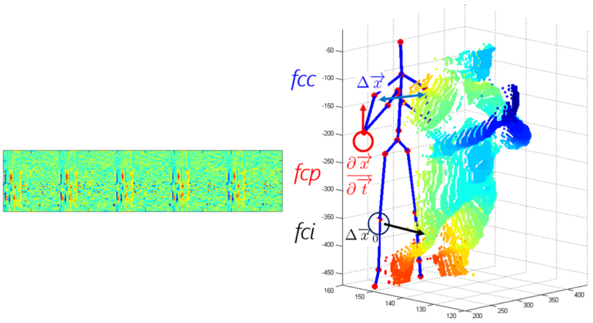

Figure 2: Input modules: left–audio input in the form of MFCC, and in order to conform to the 20 fps, 5 frames are concatenated together (10-ms fixed frame rate); right–skeleton 3D positional features.

deploy Deep Belief Networks in the place of Gaussian Mixture Models to model the emission probabilities for Hidden Markov Models. However, our proposed framework is the first work to learn the shared representation for modeling multimodal dynamic time series inputs.

3. EXPERIMENTS

ChaLearn Italian Gesture Recognition

This dataset is on "multiple instance, user independent learning" [2] of gestures. We focus on the skeletal modality and audio modality. There are 20 Italian cultural/anthropological signs, *i.e.,vattene, vieniqui, perfetto, furbo, cheduepalle, chevuoi, daccordo, seipazzo, combinato, freganiente , ok, cosatifarei, basta, prendere, noncenepiu, fame, tantotempo, buonissimo, messidaccordo, sonostufo.* We use the subset where the label data are provided during our evaluation process. The set contains 393 labeled sequences with a total of 7754 gestures. We used 350 sequences for training and the rest 43 sequences for testing, where each sequence contains 20 unique gestures. In the training set, there are in total 339,700 frames (20 fps). Note that a large number of frames is advantageous in our model settings over other nonparametric models for estimating skeletal human poses.

Audio Features

The speech was analyzed using a 25-ms Hamming window with a 10-ms fixed frame rate. We represented the speech using 12th-order Mel frequency cepstral coefficients (MFCCs) and energy, along with their first and second temporal derivatives. In order to conform to the 20 fps, 5 frames are concatenated together (10-ms fixed frame rate). Hence one audio frame will be of dimensionality $39 * 5 = 195$. The data were normalized so that, averaged over the training cases, each coefficient or first derivative or second derivative had zero mean and unit variance.

Skeleton Features

Only upper body joints are relevant to our discriminative gesture recognition tasks. Therefor, we consider only the 9 upper body joints for our task (full body joints have been compared, but as expected, led to inferior results compared to upper body 9 joints). The 9 upper body joints used are *"ShoulderCenter, ShoulderLeft, ElbowLeft, WristLeft, HandLeft, ShoulderRight, ElbowRight, WristRight, HandRight".*

Modality and Method	Classification rate
Audio Only, DBN+HMM [5]	0.554
Skeleton Only, DBN+HMM [12]	0.586
Audio + Skeleton, Model averaging	0.668
Multimodal DBN+HMM	***0.701***

Table 1: Recognition accuracy compared to the individual modal method and the multimodal confident score averaging scheme (late fusion).

The 3D coordinates of N joints of current frame c are given as: $X_c = \{x_1^c, x_2^c, \ldots, x_N^c\}$. We deploy 3D positional pairwise differences of joints [13] for observation domain $\mathcal{X}$. They capture posture features, motion features and offset features by direction concatenation: $\mathcal{X} = [f_{cc}, f_{cp}, f_{ci}]$ as demonstrated in Figure 2.

$$f_{cc} = \{x_i^c - x_j^c | i, j = 1, 2, \ldots, N; i \neq j\}$$
$$f_{cp} = \{x_i^c - x_j^p | x_i^c \in X_c; x_j^p \in X_p\}$$
$$f_{ci} = \{x_i^c - x_j^I | x_i^c \in X_c; x_j^I \in X_I\}$$

This results in a raw dimension of $N_{\mathcal{X}} = N_{joints} * (N_{joints} - 1)/2 + N_{joints}^2 + N_{joints}^2) * 3$ where N_{joints} is the number of joints used. Hence, in our experiment, $N_{joints} = 9, N_{\mathcal{X}} = 594$. Note that before extracting any features, all the 3D joint coordinates are transformed from the world coordinate system to a person centric coordinate system by placing the HipCenter (or ShoulderCenter if applied) at the origin. By including temporal differences f_{cp}, f_{ci} partially overcomes the very strong conditional independence assumption of HMMs, *i.e.*, successive frames are independent given the hidden states of the HMM.

Admittedly, we do not completely neglect human prior knowledge about information extraction for relevant static postures, velocity and offset overall dynamics of motion data. Nevertheless, the aforementioned three attributes are all very crude pairwise features without any tweak into the dataset or handpicking the most relevant pairwise, triple wise, *etc.* , designed features [1, 6, 8, 9]. A similar data driven approach has been adopted in [3] where random forest classifiers were adapted to the problem of recognizing gestures using a bundle of 35 frames. These sets of feature extraction processes resemble the *Mel Frequency Cepstral Coefficients (MFCCs)* for the speech recognition community [5].

Dynamic Networks Setup

All DBNs were pre-trained with a fixed recipe using stochastic gradient decent with a mini-batch size of 128 training cases. For Gaussian-binary RBMs, we ran 225 epochs with a fixed learning rate of 0.002 while for binary-binary RBMs we used 75 epochs with a learning rate of 0.02 and with a mini-batch size of 100 training cases. Unsupervised initializations tend to avoid local minima and increase the network's performance stability. For fine-tuning, the learning rate starts at 0.1 with 0.998 scaling after each epoch. To prevent complex co-adaptations in which a feature detector is only helpful in the context of several other specific feature detectors, we dropout [4] half of the feature detectors.

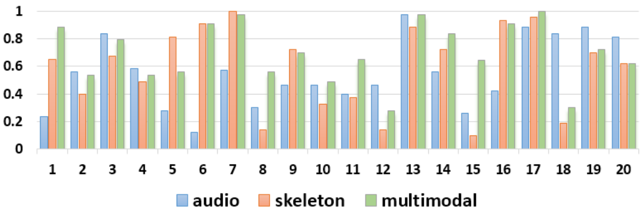

Figure 3: Comparison of individual gesture class classification rates among different modalities.

For high level feature extraction, we fix the network architecture as $[N_{\mathcal{X}}, N_{2}, 1000, 1000, 1000, 1000, N_{\mathcal{TC}}]$ where $N_{\mathcal{X}}$ is the observation domain dimension and N_{2} is the number of hidden nodes in *GRBM*. For audio modality, $N_{\mathcal{X}} = 195, N_{2} = 1000$ and for skeletal modality, $N_{\mathcal{X}} = 594, N_{2} = 2000$; $N_{\mathcal{TC}}$ is the output target class. And, in all our experiments, the number of states associated with an individual action $N_{\mathcal{H}_a}$ is chosen as 10 for modeling the states of an action class.

Once each individual modality's Deep Belief Network finishes fine tuning, we combine the multi-DBNs and extract their penultimate layers and further run 200 epochs to slightly adjust the weights for each individual modal DBN. Though we believe further carefully fine-tuned parameters would lead to more competitive results, in order to avoid "creeping overfitting", as algorithms over time become too adapted to the dataset, essentially memorizing all its idiosyncrasies and losing the ability to generalize [11], we would like to treat the model as the aforementioned more generic approach.

Results & Computational Complexity Analysis

We compare our model with the state-of-the-art methods using individual input modals and the baseline multimodal method by averaging the individual modal output confident scores in Table 1. It can be seen that both multimodal recognition rates are considerably higher than a single modal input. And the proposed framework of early fusion outperforms the confident score averaging scheme. We further plot the classification rate for each gesture class among different modalities in Figure 3. The bar plot shows the complementary information between two modalities: even when one modality achieves low recognition rate, the multimodal fusion achieves on par with another modality, *e.g.*, gesture 6; when both modalities generate noisy output, the shared multimodal scheme could learn the complementary representation and achieve a superior result, *e.g.*, gesture 15.

With a low inference cost, our framework can perform real-time gesture recognition. Specifically, a single feed forward neural network incurs trivial computation time, linearly in $\mathcal{O}(mT)$ and the complexity of Viterbi algorithm is $\mathcal{O}(T * |S|^2)$ with the number of modalities m, the number of frames T and the state number S.

4. CONCLUSION

Hand-engineered, task-specific features are often less adaptive and time-consuming to design. This difficulty is more pronounced with multimodal data as the features have to relate multiple data sources. In this paper, we presented a framework that utilizes Deep Belief Networks for modeling emission probabilities at frame-level, and introduced the shared representation for learning multimodal sensory inputs. The framework can be used to extract a unified representation that fuses various modalities together for modeling time series data. Our experimental results on bi-modal time series data, *i.e.*, audio and skeletal joints data, show that the multimodal DBN+HMM framework can learn a good model of the joint space of multiple sensory inputs, and is consistently as good as/better than the unimodal input. The proposed model also outperforms the traditional late fusion scheme, opening the door for exploring the complementary representation among multimodal inputs. It also suggests that learning features directly from data is a very important research direction and the learning-based methods are not only more generalizable to many domains, but also are powerful in combining with other well-studied probabilistic graphical models for modeling and reasoning dynamic sequences.

5. REFERENCES

[1] R. Chaudhry, F. Ofli, G. Kurillo, R. Bajcsy, and R. Vidal. Bio-inspired dynamic 3d discriminative skeletal features for human action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2013.

[2] S. Escalera, J. Gonzàlez, X. Baró, M. Reyes, O. Lopés, I. Guyon, V. Athitsos, and H. J. Escalante. Multi-modal gesture recognition challenge 2013: Dataset and results. In *ChaLearn Multi-Modal Gesture Recognition Grand Challenge and Workshop*. ACM, 2013.

[3] S. Fothergill, H. M. Mentis, P. Kohli, and S. Nowozin. Instructing people for training gestural interactive systems. In *CHI*. ACM, 2012.

[4] 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.

[5] A. Mohamed, G. E. Dahl, and G. Hinton. Acoustic modeling using deep belief networks. *Audio, Speech, and Language Processing, IEEE Transactions on*, 2012.

[6] M. Müller and T. Röder. Motion templates for automatic classification and retrieval of motion capture data. In *SIGGRAPH/Eurographics symposium on Computer animation*. Eurographics Association, 2006.

[7] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng. Multimodal deep learning. In *International Conference on Machine Learning*. IEEE, 2011.

[8] S. Nowozin and J. Shotton. Action points: A representation for low-latency online human action recognition. Technical report, Technical report, 2012.

[9] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy. Sequence of the most informative joints (smij): A new representation for human skeletal action recognition. *Journal of Visual Communication and Image Representation*, 2013.

[10] N. Srivastava and R. Salakhutdinov. Multimodal learning with deep boltzmann machines. In *Neural Information Processing Systems*, 2012.

[11] A. Torralba and A. A. Efros. Unbiased look at dataset bias. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[12] D. Wu and L. Shao. Leveraging hierarchical parametric networks for skeletal joints based action segmentation and recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[13] X. Yang and Y. Tian. Eigenjoints-based action recognition using naive-bayes-nearest-neighbor. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2012.