Deep architectures for Human Computer Interaction
Noulas, A.; Kröse, B.J.A.

Published in:
Proceedings of the Workshop on Affective Interaction in Natural Environments (AFFINE)

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (http://dare.uva.nl)
ABSTRACT
In this work we present the application of Conditional Restricted Boltzmann Machines in Human Computer Interaction. These provide a well suited framework to model the complex temporal patterns produced from humans in the audio and video modalities. They can be trained in a semi-supervised fashion and produce high accuracy, on-line inference results, while dealing with multiple concept groups. We present the results acquired in the task of speaker detection.

Categories and Subject Descriptors
H.1.2 [Information Systems Applications]: Models and Principles, User/machine Systems

Keywords
conditional Restricted Boltzmann Machines, deep architectures, audio-visual fusion

1. INTRODUCTION
If we want to create social robots and agents we have to equip them with the ability to analyse multi-modal information streams. More specifically, if we want to achieve natural Human Computer Interaction (HCI) we need machines that can analyse the audio-visual streams generated from humans and extract high order concepts, such as emotions or facial expressions for example. These concepts exhibit high variability and map to very noisy features, and the most promising results come from approaches that try to model this mapping using probabilistic methods.

Probabilistic modelling of this kind of information streams is a formidable task. First, the audio and video modalities produce complex temporal patterns whose dynamics cannot be analytically estimated, but rather should be learnt from the data. Second, complex models that can capture the underlying process are often unusable in practice, since on-line inference is computationally extremely expensive. Finally, we need large quantities of training data to learn the parameters of such a complicated pattern, but the models tend to generalise very badly in tasks involving the same data but different objectives.

Different approaches have been applied in audiovisual information streams in HCI. Some researchers try to capture arising patterns using a complex generative model, in order to represent the complex physical mechanisms that generated the data. Others prefer a discriminative approach in order to boost their classification accuracy, on a specific domain. In this work we adapt the model introduced in the work of Sutskever et. al, [7], called the Conditional Restricted Boltzmann Machine (cRBM), which is a much more suitable model for the task at hand.

A cRBM is an extension of the Restricted Boltzmann Machine (RBM). In short, the RBM is an energy based model, which can model complex relationships between observations and higher order concepts. Multiple RBMs can be stack on top of each other and learnt one at a time. The final deep architecture, will convey all the available information through different layers but capture a different view of this information in each one of the layers [1]. We can now tune the parameters of the whole architecture to optimise the task at hand. When the temporal dimension of the data is taken into consideration, by conditioning inference at each point on previous observations, we move to a cRBM.

There exists an efficient learning algorithm for all RBM-based models, called Contrastive Divergence [4], which is a local approximation of the gradient of the data likelihood function and performs very well in practice. Furthermore, the first part of the learning procedure, where the structure of the observed data is extracted, is completely unsupervised. The model can generate data following the observations distribution after this point. We can then extend it to semi-supervised learning using some training data to tune the parameters towards the desired objective (or objectives). In our case we model the audio-visual streams generated by interacting persons. Successful HCI should allow similar interaction between humans and computers. We demonstrate the ability of the system to determine whether a person is speaking or not, but we can extend our objective to emotion detection or gesture recognition by tweaking the output of the unsupervised phased using properly labelled data examples.

A further advantage of such models is the fact that exact
The Restricted Boltzmann Machine

A conditional RBM

Figure 1: On the left we can see the structure of the Restricted Boltzmann Machine. There are undirected connections from all the nodes of hidden layer $h$ to all the nodes of visible layer $x$. On the right side we see a conditional RBM where the hidden and visible layers of time slice $t$ receive incoming connections from the visible layers of the previous time slices.

on-line inference is fast, because all the variables we want to infer are conditionally independent given the observations. During inference the behaviour of a deep architecture is very similar to that of a neural network. Thus, real time performance of an agent utilising a deep architecture is feasible.

In the next section we present the proposed cRBM framework. In section 3 we discuss the application of such a model on multi-modal data coming from the face region of a person. Section 4 contains the details of our experimental setup and results compared to those achieved by the methods proposed in [6] and [2]. We conclude the paper with a section discussing future directions in the application of cRBMs on multi-modal information streams.

2. BUILDING DEEP ARCHITECTURES

In this section we briefly review the process of building a deep architecture for temporal data modelling. Firstly, we describe the probabilistic interpretation of the RBM as well as the way CD is applied for learning its parameters. Secondly, we present how we extend from the RBM to the cRBM. Thirdly, we describe how we can build deep architectures using a cRBM as building block. We conclude this section describing how we can use deep architectures to perform task specific inference.

2.1 Restricted Boltzmann Machine

A RBM is an energy-based model, which means that the probability distribution over the variables of interest is defined through an energy function. The undirected graphical model of the RBM is composed from a layer of observable variables $x = \{x(i)\}$ and a layer of hidden variables $h = \{h(j)\}$. The graphical model of the RBM is visible in figure 1. There are connections between different layers but no connections within a layer. The log likelihood of a probability of a given configuration is defined by:

$$-\log P(x,h) = -a^T x - b^T h - h^T W x + \text{const} \quad (1)$$

where the vectors $a$ and $b$ contain the biases of the corresponding nodes, and matrix $W$ the weights connecting hidden to visible nodes. Note that we didn’t write down the explicit form of the normalization term in order to express our inability to compute it in general.

The main advantage of such an architecture is that the posterior over the variables of a layer factorizes completely. Thus,

$$p(x|h) = \prod_i p(x_i|h) \quad \text{and} \quad p(h|x) = \prod_j p(h_j|h) \quad (2)$$

Typically, RBMs use binary variables for both the hidden and visible nodes. In our case we have real valued data and therefore we use continuous real valued nodes with Gaussian noise [10] for the visible layer. The conditional distributions are then given by:

$$p(x_i|h) = \mathcal{N}\left(c_i + \sum_j h_k w_{ij}, 1\right)$$

$$p(h_j|x) = \sigma \left(b_j + \sum_i x_i w_{ij}\right) \quad (3)$$

with $\sigma(\cdot)$ being the logistic function.

The parameters $W$, $b$ and $h$ of the Energy function can be obtained by maximum likelihood learning but that is very slow. Instead, we can apply CD learning and , use the following update rules:

$$\Delta w_{ij} \propto \langle x_i h_j \rangle_{\text{data}} - \langle x_i h_j \rangle_{\text{recon}}$$

$$\Delta a_i \propto \langle x_i \rangle_{\text{data}} - \langle x_i \rangle_{\text{recon}}$$

$$\Delta h_j \propto \langle h_j \rangle_{\text{data}} - \langle h_j \rangle_{\text{recon}} \quad (4)$$

We refer the interested reader to [4] regarding the origin of these rules and a thorough comparison of CD with ML learning. Intuitively, CD can be seen as an indirect way to make the model distribution identical to the data distribution. We obtain the model parameters that would distort the data reconstruction as little as possible. If we push this error to zero, in an infinite chain sampling operation, the model would produce the exact same data with our training distribution. Thus the model of the data and the learnt model would be identical.

2.2 Conditional Restricted Boltzmann Machine

The RBM could model in our case static frames of data. We can model temporal dependencies by treating the observa-
3. FEATURE EXTRACTION

The audio and video streams generated by a person are complex highly varying signals, making it very hard to detect correlations in the raw pixel and audio data. There are many different descriptors for such data, each one capturing the most informative part of the stream for a given task. Since we want to model the correlation between a person’s face movement and the corresponding audio stream, we use the features that produced the state-of-the-art result in lip-movement generation [3].

From the video stream we create an active appearance model of the person [9], and then detect the points of interest on each frame of the stream. Some example frames are visible in figure 2. There are 52 points tracked on each face throughout the stream. Some of these points correspond to less informative parts of the face in terms of speaking activity. However we expect the cRBM to automatically learn the informative subspace of the high-dimensional input data, and utilise only information relative to the inference objective.

In the audio stream we extract the 13 first MFCCs. In speaker diarization or recognition tasks, it is a common technique to concatenate the first and second order differences of these vectors. However, in the cRBM setting, consequent audio descriptors will be connected to common hidden variables. Thus, their weights will compete during the training procedure to explain the data, and therefore useful information regarding the descriptor differences can automatically be detected.

4. EXPERIMENTS

In this section we first describe the experimental setting and the objectives we wanted to meet. In the second subsection we present our experimental results along with the results acquired on the same data using the algorithms of [6] and [2].

4.1 Experimental Setup

We trained our cRBM using 3000 sequences from a discussion recording. Example frames of this data set are visible on the left side in Figure 2. We used the weights acquired during the cRBM training to initialise a sigmoid belief net.

2.3 Building a Deep Architecture

Once we have learnt the parameters of the cRBM, we can add layers as in a Deep Belief Network (DBN) [5]. The previously learnt cRBM is used to get a sequence of hidden layer vectors from the training data. This sequence can now be used as visible data for a new hidden layer. We expect higher levels of the model to capture higher concepts appearing in the data. This greedy learning can be justified using a variational bound [5]. At this point, the final structure can be seen as an autoregressor with Gaussian noise, which takes into account the specific temporal relations of the data.

2.4 Inference using the Deep Architecture

There is a different number of ways to infer the quantity in question from the deep architecture. For example, we can train one deep architecture per class, and classify a novel example as the class that creates a more accurate reconstruction. Alternatively we can use any classification method on the informative, low dimensional hidden layer(s). Different methods are discussed thoroughly in [4].

In this work we tune the final architecture as a sigmoid belief net. In this way, we expect to alter slightly only the weights that will perform optimal classification in the subspace learnt during CD training [1].

\begin{equation}
\Delta d_{t+q}^{i-j} \propto x_{t+q}^{i-j} \left( h_{t+q}^j \right)_{\text{data}} - \left( h_{t+q}^j \right)_{\text{recon}}
\end{equation}

\begin{equation}
\Delta a_{t+q}^{i-j} \propto a_{t+q}^{i-j} \left( x_{t+q}^j \right)_{\text{data}} - \left( x_{t+q}^j \right)_{\text{recon}}
\end{equation}

Figure 2: Some example frames of our data. The two left columns come from the labelled streams, while the rest come from the test streams (left to right, speakers 1, 2, 3). The top row contains a typical frame of the stream while the bottom one some extreme cases.
with the same structure, and performed speaker diarization on novel recording. The objective of our task, is to infer whether each person is speaking or not, on each frame of the novel recording.

The novel recording comes from a meeting setting with three participants. The first participant is dominant, speaking 70% of the time, while the two other participants vocalise significantly less. The second and third participant are moving their lips, or in different ways occludes their face, in order to make audio-visual synchronisation detection harder. We note here, that none of the participants in the training recording exhibits such extreme behaviour.

We evaluate our model in terms of the percentage of frames where the speaker status was correctly classified. In pattern recognition terms the total of true positive and true negative examples divided by the total number of classifications. Furthermore we compare our results to those acquired using MI based MI based MI based estimates the mutual information between the pixel value variations of a person and the corresponding audio script [2]. The speaker selected is the person most correlated to the stream. The second method, labelled as D.B.N., uses this information as well as information about the voice and appearance of the speakers. All this information is fused in a dynamic Bayesian network whose parameters are acquired directly from the test data [6].

### 4.2 Results & Discussion

In table 1 we can see the results acquired by our approach and the methods presented in [6] and [2]. Our model achieved an overall classification accuracy of 68%, which is close to the 76% achieved by the dynamic Bayesian network presented of [6].

Furthermore we observe that the acquired results are much more robust when it comes to speakers 2 and 3 which vocalise less. The D.B.N., outperforms our proposed method when it comes to speaker 1, because he provides enough data to acquire a reliable model of his voice. Furthermore, our cRBM is outperforming both models when it comes to a speaker with behaviour very different from that of the training set.

![Image](image.png)

<table>
<thead>
<tr>
<th>Speaker</th>
<th>Speaker 1</th>
<th>Speaker 2</th>
<th>Speaker 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>MI Based</td>
<td>88%</td>
<td>62%</td>
<td>61%</td>
</tr>
<tr>
<td>D.B.N.</td>
<td>76%</td>
<td>70%</td>
<td>76%</td>
</tr>
</tbody>
</table>

Table 1: The accuracy results in speaker diarization on different speakers of the meeting.

This data set, although limited, can help drawn some important conclusions. Firstly, the cRBM manages to generalise well, using a limited training set to classify a novel test set in a very sparse domain. Secondly, adaptation to the new data set is really important. However, the first training phase of a cRBM is computationally intensive and novel data is not easily incorporated. Finally, a possible incorporation of RBM-like observational models to a dynamic probabilistic framework can account well for complex multimodal patterns, as in this example the audio and video streams generated by a person.

### 5. Discussion of Core Contribution

This paper, to our best knowledge is the first time that a cRBM is applied on audio visual data. Similar deep probabilistic architectures gain increased attention in modern machine learning research, and since their theoretical advantages are well fitted to multimodal fusion we expect them to be widely used in coming years.

The first advantage of these models regards their representational power. cRBMs belong to the family of product of experts. Therefore they are able to capture complex, high- 

#### 7. REFERENCES


