Multimodal Fusion using Dynamic Hybrid Models

Mohamed R. Amer *
SRI International
amerm@onid.orst.edu

Behjat Siddiquie
SRI International
behjat.siddiquie@sri.com

Saad Khan
SRI International
saad.khan@sri.com

Ajay Divakaran
SRI International
ajay.divakaran@sri.com

Harpreet Sawhney
SRI International
harpreet.sawhney@sri.com

Abstract

We propose a novel hybrid model that exploits the strength of discriminative classifiers along with the representational power of generative models. Our focus is on detecting multimodal events in time varying sequences. Discriminative classifiers have been shown to achieve higher performances than the corresponding generative likelihood-based classifiers. On the other hand, generative models learn a rich informative space which allows for data generation and joint feature representation that discriminative models lack. We employ a deep temporal generative model for unsupervised learning of a shared representation across multiple modalities with time varying data. The temporal generative model takes into account short term temporal phenomena and allows for filling in missing data by generating data within or across modalities. The hybrid model involves augmenting the temporal generative model with a temporal discriminative model for event detection, and classification, which enables modeling long range temporal dynamics. We evaluate our approach on audio-visual datasets (AVEC, AVLetters, and CUAVE) and demonstrate its superiority compared to the state-of-the-art.

1. Introduction

Many events in real life are inherently multimodal with each modality containing information useful for detecting or recognizing the event. Despite this, most work [11, 12] focuses on modeling and recognizing events using a single modality, neglecting other sources of information. While this might be sufficient for certain problems, it is inadequate when the events to be detected are complex and subtle (e.g. human emotions). Humans are capable of combining cues from multiple modalities to reason about specific events. Therefore when multiple, information rich, modalities are present, it becomes important to jointly interpret and reason about the information from each modality. While jointly modeling multiple modalities, the temporal information within and across modalities also needs to be accounted for. Following the human cognitive system, we propose to solve the multimodal fusion using a neuro-inspired model namely Conditional Restricted Boltzmann Machines (CRBMs) [27]. The CRBM is a non-linear generative model for time series data that uses an undirected model with binary latent variables connected to a number of visible variables. To the best of our knowledge, we are the first to propose multimodal fusion of temporal data using temporal deep networks. A CRBM based generative model enables modeling short-term multimodal phenomenon and also allows us to deal with missing data by generating it within or across modalities. Furthermore, we propose a hybrid model to acquire the benefits of a discriminative classifier. The hybrid model involves enhancing the CRBM with a Conditional Random Field (CRF) based discriminative model, leading to a superior classification performance, while also allowing us to model long-term temporal dynamics. We evaluate our approach on multiple audio-visual datasets and show how our results are comparable/superior to the state-of-the-art approaches.

Our contributions are two fold. First, we propose a new general hybrid model that consists of a generative model that is capable of learning a homogeneous joint feature representation that captures low level concepts from multiple heterogeneous data sources and a discriminative model for high level reasoning. This hybrid model **combines the advantages of temporal generative and discriminative models** forming an extendable formal fusion framework for classifying multimodal data at multiple time scales. Furthermore, it can deal with missing data both within and across modalities. Second, we provide an extensive evaluation of the hybrid model on multimodal time-varying data

*The author is a student at Oregon State University and did this work while being an intern at SRI International.

sequences, **systematically justifying each component** of our model.

Paper organization: In sec. 2 we discuss prior work. In sec. 3 we give a brief background of similar models that motivate our approach, followed by a description of our hybrid model. In sec. 4 we describe the inference and learning. In sec. 5 we show quantitative results of our approach, followed by the conclusion in sec. 6.

2. Prior work

Representative work on multimodal (Audio-Video) fusion includes the Hidden Markov Models (HMMs) based methods [5] and Conditional Random Fields (CRFs) [18]. However, these approaches lack the advantages of generative models, which include, the ability to learn a joint representation and the ability to deal with missing data. Recently, deep networks have been used for multimodal fusion [23] (tags/image) and [13] (audio/images). While all prior work on multimodal deep learning ignores the temporal aspect of the data, our preliminary experiments showed that jointly modeling the temporal content from different modalities helps in substantially improving both classification and generation performance. In this paper, we focus on the joint modeling of multimodal data using a hybrid model that comprises of temporal generative and discriminative models. We first review prior work on hybrid models followed by deep networks and conditional random fields.

Hybrid Models: consist of a generative model, which usually learns a feature representation of low level input, and a discriminative model for higher level reasoning. Recent work has empirically shown that generative models which learn a rich feature representation tend to outperform discriminative models that rely solely on hand-crafted features [16]. Hybrid models can be divided into three groups, joint methods [9, 3], iterative methods[4], and staged methods [16]. Joint methods optimize a single objective function which consists of both the generative and discriminative energies. Iterative methods consist of a generative and a discriminative model that are trained in an iterative manner, influencing each other. In staged methods, both models are trained separately, with the discriminative model being trained on representations learned by the generative model. Classification is performed after projecting the samples into a fixed-dimensional space induced by the generative model. Staged methods are currently the most popular in the community, primarily because they are computationally easier to manage. Our proposed approach follows the staged method. For the purpose of solving our problem we choose a CRBM based generative model, and a CRF based discriminative model. Next we briefly go over recent literature of deep networks.

Deep Networks: are able to learn rich features in an unsupervised manner, this is what makes deep learning very powerful. They have been successfully applied to many problems [1]. Restricted Boltzmann Machines (RBMs) form the building blocks in deep networks models [19]. In [19], the networks are trained using the Contrastive Divergence (CD) algorithm [7], which demonstrated the ability of deep networks to capture the distributions over the features efficiently and to learn complex representations. RBMs can be stacked together to form deeper networks known as Deep Boltzmann Machines (DBMs), which capture more complex representations. Recently, deep networks based temporal models, capable of modeling a more temporally rich set of problems have been proposed. These include Conditional RBMs (CRBMs) [27] and Temporal RBMs (TRBMs) [25, 24]. CRBMs have been successfully used in both visual and audio domains. They have been used for modeling human motion [27], tracking 3D human pose [26] and phone recognition [12]. We now briefly go over recent literature on CRFs.

Conditional Random Fields: have been shown to be effective for labeling sequential data. CRFs [8] are able to utilize arbitrary features and model non-stationarities. Hidden Conditional Random Fields (HCRFs) have been proposed as an extension of CRFs with hidden states [17, 28]. CRFs with hidden states, showed an increase in modeling power and have been shown to improve the classification performance.

In the following section we formulate our hybrid model, and specify its generative and discriminative components.

3. Multimodal Hybrid Model

The hybrid model allows us to take advantage of the benefits of the generative models (filling in missing data, inferring joint representation), as well as the benefits of a discriminative model leading to a stronger classifier compared to purely generative models. We propose a general model that reduces to a purely discriminative model, if we marginalize over the hidden nodes of the generative part $\mathbf{h}$. This is equivalent to inferring the class labels $\mathbf{y}$ directly based on the features. Also it reduces to a purely generative model, if we marginalize over $\mathbf{y}$. This is equivalent to learning the features for a specific class. We first define the model's variables.

Our multimodal fusion hybrid model $p(\mathbf{y}_t, \mathbf{v}_t, \mathbf{h}_t|\mathbf{v}_{<t})$ shown in (1) is decomposed into two terms, a generative component $p(\mathbf{v}_t, \mathbf{h}_t|\mathbf{v}_{<t})$, and a discriminative component $p(\mathbf{y}_t|\mathbf{h}_t, \mathbf{v}_t)$[1]. We define $\mathbf{y}_t$ to be a multi-class label vector at time t, $\mathbf{v}_t$ is the vector of raw features at time t, and $\mathbf{h}_t$ is a vector of the hidden variables. $\mathbf{v}_{<t}$ is the concatenated

[1]In our model $\mathbf{y}_t$ is independent of $\mathbf{v}_t$ given $\mathbf{h}_t$, i.e. $p(\mathbf{y}_t|\mathbf{v}_t, \mathbf{h}_t) = p(\mathbf{y}_t|\mathbf{h}_t)$, however, for generality we decided to start from the more general formulation that allows us to explain our hybrid model.

history vector of the visible variables (raw features).

$$\underbrace{p(\mathbf{y}_t, \mathbf{v}_t, \mathbf{h}_t|\mathbf{v}_{<t})}_{\text{Hybrid}} = \underbrace{p_D(\mathbf{y}_t|\mathbf{v}_t, \mathbf{h}_t)}_{\text{Discriminative}} \cdot \underbrace{p_G(\mathbf{v}_t, \mathbf{h}_t|\mathbf{v}_{<t})}_{\text{Generative}} \quad (1)$$

In the following subsections, we first go over the background of our model by describing RBMs, followed by their extension to CRBMs, which are the main building blocks of our generative component. Then we define our discriminative component.

3.1. Background

We briefly define Restricted Boltzmann Machines and Conditional Restricted Boltzmann Machines since our hybrid model uses them for the generative part. We decided to use these models to enable efficient on-line inference and learning.

The Restricted Boltzmann Machines: An RBM defines a probability distribution $p_{\text{R}}(\mathbf{v}, \mathbf{h}; \boldsymbol{\theta}_{\text{R}})$ as a Gibbs distribution (2), where $\mathbf{v}$ is a vector of visible nodes, $\mathbf{h}$ is a vector of hidden nodes. $E_{\text{R}}(\mathbf{v}, \mathbf{h}; \boldsymbol{\theta}_{\text{R}})$ is the energy function and Z is the partition function which ensures that the distribution is valid. The parameters $\boldsymbol{\theta}_{\text{R}}$ to be learned are $\mathbf{a}$ and $\mathbf{b}$ the biases for $\mathbf{v}$ and $\mathbf{h}$ respectively and the weights W. The RBM architecture is defined as fully connected between layers, with no lateral connections. This architecture implies that $\mathbf{v}$ and $\mathbf{h}$ are factorial given one of the two vectors. This allows for the exact computation of $p(\mathbf{v}|\mathbf{h})$ and $p(\mathbf{h}|\mathbf{v})$.

$$p_{\text{R}}(\mathbf{v}, \mathbf{h}; \boldsymbol{\theta}_{\text{R}}) = \exp[-E_{\text{R}}(\mathbf{v}, \mathbf{h}; \boldsymbol{\theta}_{\text{R}})]/Z(\boldsymbol{\theta}_{\text{R}}),$$

$$Z(\boldsymbol{\theta}_{\text{R}}) = \textstyle\sum_{\mathbf{v},\mathbf{h}} \exp[-E_{\text{R}}(\mathbf{v}, \mathbf{h}; \boldsymbol{\theta}_{\text{R}})], \quad (2)$$

$$\boldsymbol{\theta}_{\text{R}} = \{\mathbf{a}, \mathbf{b}, W\}$$

In case of binary valued data $v_i \in \{0, 1\}$ and a binary valued hidden layer $h_j \in \{0, 1\}$, a logistic function[2] is defined for each of the two conditionals $p(v_i = 1|\mathbf{h}) = \sigma(a_i + \sum_j h_j w_{ij})$ and $p(h_j = 1|\mathbf{v}) = \sigma(b_j + \sum_i v_i w_{ij})$. The energy function is defined as in (3).

$$E_{\text{R}}(\mathbf{v}, \mathbf{h}; \boldsymbol{\theta}_{\text{R}}) = -\sum_i a_i v_i - \sum_j b_j h_j - \sum_{i,j} v_i w_{i,j} h_j \quad (3)$$

In case of real valued data, $p(v_i|\mathbf{h})$ is defined as a multivariate Gaussian distribution with zero mean and unit covariance $p(v_i|\mathbf{h}) = \mathcal{N}(a_i + \sum_j h_j w_{ij}, 1)$. The conditional $p(h_j = 1|\mathbf{v})$ stays the same, since we want the hidden layer to be binary (empirically proven to be better [27, 25]). The energy function $E_{\text{RBM}}(\mathbf{v}, \mathbf{h}; \boldsymbol{\theta}_{\text{RBM}})$ is slightly modified to allow for the real valued $\mathbf{v}$ as shown in (4).

$$E_{\text{R}}(\mathbf{v}, \mathbf{h}; \boldsymbol{\theta}_{\text{R}}) = -\sum_i (a_i - v_i)^2/2 - \sum_j b_j h_j - \sum_{i,j} v_i w_{i,j} h_j \quad (4)$$

[2]The logistic function $\sigma(\cdot)$ for a variable x is defined as $\sigma(x) = (1 + exp(-x))^{-1}$.

The Conditional Restricted Boltzmann Machines: CRBMs are a natural extension of RBMs for modeling short term temporal dependencies. In simple terms, a CRBM is an RBM model which takes into account history from the previous time instances $[(t - N), \ldots, (t - 1)]$ at time (t). This is done by treating the previous time instances as additional inputs. Doing so does not complicate inference[3]. A CRBM defines a probability distribution $p_{\text{C}}(\mathbf{v}_t, \mathbf{h}_t|\mathbf{v}_{<t})$ as a Gibbs distribution (5).

$$p_{\text{C}}(\mathbf{v}_t, \mathbf{h}_t|\mathbf{v}_{<t}; \boldsymbol{\theta}_{\text{C}}) = \exp[-E_{\text{C}}(\mathbf{v}_t, \mathbf{h}_t|\mathbf{v}_{<t}; \boldsymbol{\theta}_{\text{C}})]/Z(\boldsymbol{\theta}_{\text{C}}),$$

$$Z(\boldsymbol{\theta}_{\text{C}}) = \textstyle\sum_{\mathbf{v},\mathbf{h}} \exp[-E_{\text{C}}(\mathbf{v}_t, \mathbf{h}_t|\mathbf{v}_{<t}; \boldsymbol{\theta}_{\text{C}})] \quad (5)$$

$$\boldsymbol{\theta}_{\text{C}} = \{\mathbf{a}, \mathbf{b}, A, B, W\}$$

The additional inputs from previous time instances are modeled as directed autoregressive edges from the past N visible nodes and the past M hidden layers, where, N does not have to be equal to M. The concatenated history vector is defined as $\mathbf{v}_{<t}$. The new energy function $E_{\text{C}}(\mathbf{v}_t, \mathbf{h}_t|\mathbf{v}_{<t}; \boldsymbol{\theta}_{\text{C}})$ is defined similar to (4) in (6), where $c_{i,t} = a_{i,t} + \sum_n A_{n,i} v_{n,<t}$ and $d_{j,t} = b_{j,t} + \sum_m B_{m,j} v_{m,<t}$ and A and B are matrices of concatenated vectors of previous time instances of $\mathbf{a}$ and $\mathbf{b}$.

$$\begin{aligned} E_{\text{C}}(\mathbf{v}_t, \mathbf{h}_t|\mathbf{v}_{<t}; \boldsymbol{\theta}_{\text{C}}) = &-\textstyle\sum_i (c_{i,t} - v_{i,t})^2/2 \\ &- \textstyle\sum_j d_{j,t} h_{j,t} - \sum_{i,j} v_{i,t} w_{i,j} h_{j,t} \end{aligned} \quad (6)$$

3.2. Generative Model

For the generative component, $p_{\text{G}}(\mathbf{v}_t, \mathbf{h}_t|\mathbf{v}_{<t})$, we define a Gibbs distribution over a multimodal network of stacked CRBMs (7). This is similar to the approach proposed in [23] and [13] except that we use CRBMs as our main building block instead of RBMs. This enables us to model the temporal nature of the audio-visual data.

$$p_{\text{G}}(\mathbf{v}_t, \mathbf{h}_t|\mathbf{v}_{<t}) = \exp[-E_{\text{G}}(\mathbf{v}_t, \mathbf{h}_t|\mathbf{v}_{<t})]/Z(\boldsymbol{\theta}_{\text{G}}),$$

$$Z(\boldsymbol{\theta}_{\text{G}}) = \textstyle\sum_{\mathbf{v},\mathbf{h}} \exp[-E_{\text{G}}(\mathbf{v}_t, \mathbf{h}_t|\mathbf{v}_{<t})], \quad (7)$$

$$\boldsymbol{\theta}_{\text{G}} = \{\mathbf{a}, \mathbf{b}, A, B, W\}$$

The multimodal energy $E_{\text{G}}(\mathbf{v}_t, \mathbf{h}_t|\mathbf{v}_{<t}; \boldsymbol{\theta}_{\text{G}})$ is decomposed into two parts as shown in (8).

$$\begin{aligned} E_{\text{G}}(\mathbf{v}_t, \mathbf{h}_t|\mathbf{v}_{<t}; \boldsymbol{\theta}_{\text{G}}) = &\textstyle\sum_m E_{\text{S}}(\mathbf{v}_t^m, \mathbf{h}_t^m|\mathbf{v}_{<t}^m) \\ &+ E_{\text{F}}(\mathbf{h}_t^{1,\ldots,M}, \mathbf{h}_t^F|\mathbf{h}_{<t}^{1,\ldots,M}) \end{aligned} \quad (8)$$

[3]Some approximations have been made to facilitate efficient training and inference, more details are available in [27].

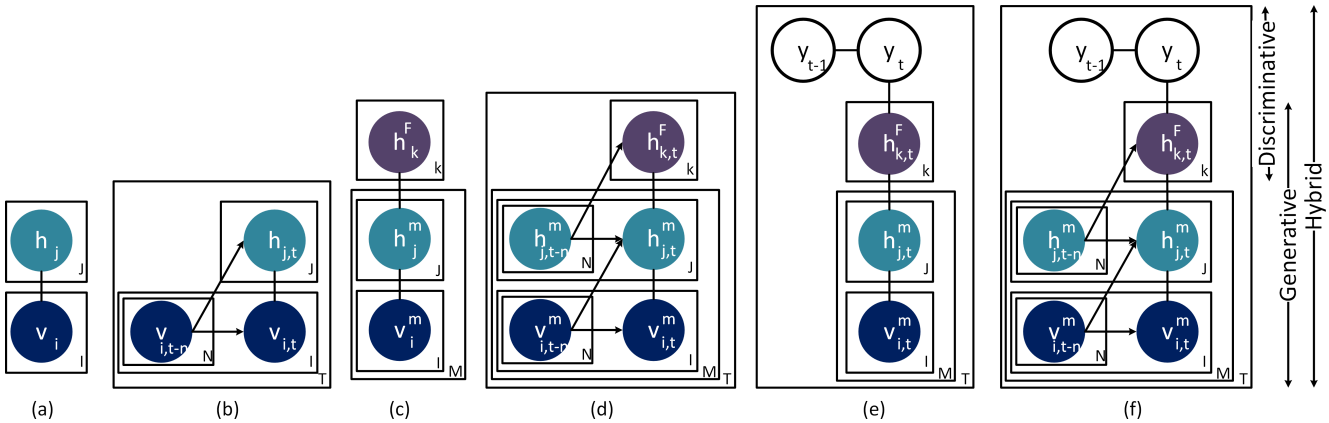

Figure 1. This figure shows a progression of models in increasing order of sophistication from (a) to (f) ((a) being the simplest). (a) Non temporal generative deep network RBM [19], (b) Temporal generative deep network CRBM[27], (c) Multimodal non temporal generative deep network RBM[13], (d) Multimodal dynamic generative deep network CRBM, (e) Hybrid Multimodal CRF-RBM, and (f) Hybrid dynamic CRF-CRBM.

The first part is the single modality energy E_{S}, which is defined over a CRBM of a single modality m[4]. It consists of unary terms representing the bias of each layer, and a pairwise term which relates the nodes of two layers (9).

$$E_{\text{S}}(\mathbf{v}_t^m, \mathbf{h}_t^m | \mathbf{v}_{<t}^m) = -\sum_i (c_{i,t}^m - v_{i,t}^m)^2/2 \\ -\sum_j d_{j,t}^m h_{j,t}^m - \sum_{i,j} v_{i,t}^m w_{i,j}^m h_{j,t}^m \tag{9}$$

The second part of (8) is the fusion energy E_{F} for the joint representation, where $\mathbf{h}_t^F$ is the fusion hidden layer. For a multimodal CRBM, we define the joint representation (i.e fusion) layer to be the top layer as shown in (10).

$$E_{\text{F}}(\mathbf{h}_t^{1,\ldots,M}, \mathbf{h}_t^F | \mathbf{h}_{<t}^{1,\ldots,M}) = -\sum_{i,m} c_{i,t}^m h_{i,t}^m \\ -\sum_j d_{j,t}^F h_{j,t}^F - \sum_{i,j,m} h_{i,t}^m w_{i,j}^F h_{j,t}^F \tag{10}$$

Next, we specify the discriminative component of our hybrid model.

3.3. Discriminative Model

We now describe our discriminative model for classification. While the multimodal CRBMs are very effective for learning and representing short term temporal phenomena, we also need to model long range temporal dynamics. With this requirement in mind, we choose Conditional Random Fields (CRFs) [8] as our discriminative model. Although hidden state CRFs (HCRFs) provide an increased modeling power compared to CRFs, in our case the classification results of CRFs and HCRFs were comparable, justifying the use of CRFs. Our discriminative component, modeled by $p_{\text{D}}(\mathbf{y}_t | \mathbf{h}_t^F; \boldsymbol{\theta}_{\text{D}})$ as a Gibbs distribution of a CRF model as shown in (11).

$$p_{\text{D}}(\mathbf{y}_t | \mathbf{h}_t^F; \boldsymbol{\theta}_{\text{D}}) = \exp[E_{\text{D}}(\mathbf{y}_t | \mathbf{h}_t^F; \boldsymbol{\theta}_{\text{D}})]/Z(\boldsymbol{\theta}_{\text{D}}), \\ Z(\boldsymbol{\theta}_{\text{D}}) = \sum_{\mathbf{y}} E_{\text{D}}(\mathbf{y} | \mathbf{h}_t^F; \boldsymbol{\theta}_{\text{D}}), \\ \boldsymbol{\theta}_{\text{D}} = \{\boldsymbol{\omega}^1, \boldsymbol{\omega}^2\} \tag{11}$$

We define $\mathbf{y}_t$ to be the label of the sequence at time t and $\mathbf{h}_t^F$ to be the output of the multimodal CRBM (8) which serves as an input to the CRF as shown in Fig. 1(e,f). Finally, we define Z to be the partition function to ensure the proper normalization of the model. The energy of the CRF, $E_{\text{D}}(\mathbf{y}_t | \mathbf{h}_t^F; \boldsymbol{\theta}_{\text{D}})$, is defined in (12). Note that the CRF model assigns a label to each node of the sequence.

$$E_{\text{D}}(\mathbf{y}_t | \mathbf{h}_t^F; \boldsymbol{\theta}_{\text{D}}) = \sum_j \omega_j^1 f_j^1(\mathbf{y}_{t-1}, \mathbf{y}_t, \mathbf{h}_t^F) \\ + \sum_k \omega_k^2 f_k^2(\mathbf{y}_t, \mathbf{h}_t^F) \tag{12}$$

where $\mathbf{f}^1$ is a transition feature function and $\mathbf{f}^2$ is a state feature function, with $\boldsymbol{\omega}_t^1$ the transition component of the parameters and $\boldsymbol{\omega}_t^2$ the state component of the parameters. In the following section we specify the inference and learning algorithms for our model.

4. Inference and Learning

Inference is done in a layer-wise manner by activating a hidden layer given the visible layer using the conditional independence advantage of the CRBM model $p(h_j = 1 | \mathbf{v})$. Fig. 2(a) shows the feature representation activated using unimodal CRBMs for audio and video on the AVEC dataset and Fig. 2(b) shows the fused feature representation from the multimodal CRBM using the activations of unimodal CRBMs on the AVEC dataset. Given the fused feature representation $\mathbf{h}_t^F$, we use it as an input to the CRF and then

[4]Extension to multiple hidden layers is straightforward, where the higher layers are binary CRBMs, with inputs from the previous hidden layer.

we get the predicted label sequence $\mathbf{y}_t$ by maximizing (13). We summarize our inference in Algorithm 1.

$$\mathbf{y}_t = \arg\max_{\mathbf{y}} p_{\mathrm{D}}(\mathbf{y}_t|\mathbf{h}_t^F;\boldsymbol{\theta}_{\mathrm{D}}) \tag{13}$$

Algorithm 1: Inference

Input: Multimodal data, network parameters θ_G, and θ_D.
Output: Activity class label per frame y_t

for $m = 1$:*Number of modalities* M **do**
  for $j = 1$:*Number of nodes in* $\mathbf{h}^m$ **do**
    Activate the modality's hidden layer:
    $p_G(h_j^m = 1|\mathbf{v}^m, \mathbf{v}_{<t}^m) \sim \sigma(c_j^m + \sum_i v_i^m w_{ij}^m)$;
  end
end
for $k = 1$:*Number of nodes in* $\mathbf{h}^F$ **do**
  Activate the Fusion hidden layer:
  $p_G(h_k^F = 1|\mathbf{h}^{1,\ldots,M}, \mathbf{h}_{<t}^{1,\ldots,M}) \sim$
  $\sigma(c_k^F + \sum_j h_j^{1,\ldots,M} w_{jk}^F)$;
end
Classify the frame label: $\mathbf{y}_t = \arg\max_{\mathbf{y}} p_{\mathrm{D}}(\mathbf{y}_t|\mathbf{h}_t^F;\boldsymbol{\theta}_{\mathrm{D}})$

Learning our hybrid model is performed by separately learning the parameters for the generative $\boldsymbol{\theta}_{\mathrm{G}}$ and discriminative part $\boldsymbol{\theta}_{\mathrm{D}}$. The parameters of the generative model (CRBM)[5] are learned using Contrastive Divergence (CD) [7] which produces the learning rules in (14). The update equations of the dynamically changing bases $\Delta\mathbf{c}$ and $\Delta\mathbf{d}$ are obtained by first updating ΔA and ΔB as in the case of the real valued RBM (4) and then combining them with Δa and Δb.

$$\begin{aligned}
\Delta w_{i,j} &\propto \langle v_i h_j\rangle_{data} - \langle v_i h_j\rangle_{recon} \\
\Delta a_i &\propto \langle v_i\rangle_{data} - \langle v_i\rangle_{recon} \\
\Delta b_j &\propto \langle h_j\rangle_{data} - \langle h_j\rangle_{recon} \\
\Delta A_{k,i,t-n} &\propto v_{k,t-n}(\langle v_{i,t}\rangle_{data} - \langle v_{i,t}\rangle_{recon}), \\
\Delta B_{i,j,t-m} &\propto v_{i,t-m}(\langle h_{j,t}\rangle_{data} - \langle h_{j,t}\rangle_{recon})
\end{aligned} \tag{14}$$

Where $\langle\cdot\rangle_{data}$ is the expectation with respect to the data distribution and $\langle\cdot\rangle_{recon}$ is the expectation with respect to the reconstructed data. The reconstruction is generated by first sampling $p(h_j = 1|\mathbf{v})$ for all the hidden nodes in parallel. The visible nodes are then generated by sampling $p(v_i|\mathbf{h})$ for all the visible nodes in parallel. The discriminative learning is done by maximum likelihood estimation of $\boldsymbol{\theta}_{\mathrm{D}}$ in (11) using [20]. In the following section we evaluate our model on standard benchmarks against the state-of-the-art approaches.

[5]Maximum likelihood learning is slow in learning RBM parameters, however, learning still works if we approximately follow the gradient of another function, in this case the other function is CD.

5. Experiments

In order to evaluate the performance of our proposed approach, we compare the different hybrid models shown in Fig. 1 (c), (d), (e), and (f), which differ in the type and configuration of discriminative and generative models. We show the improvement resulting from each additional enhancement. We also compare against the state-of-the-art generative, discriminative, and hybrid models and report the classification results in multiple settings which include - unimodal data (Tab 1), multimodal data (Tab 1), missing data within a modality (Tab 2) and missing data across modalities (Tab 3).

Variants: To fully evaluate our approach we propose a set of variants that consists of different combinations of discriminative and generative models. The discriminative set consists of {SVM, CRF}, and the generative set consists of {RBM, CRBM}. Using the four variants created from the two sets in addition to running the discriminative set on raw features (which we refer to as RAW) as our baseline, we can evaluate the value of each additional component of our hybrid model. We compare against [13] using the variant SVM-RBM, which is equivalent to the one they used. The training of the generative part is completely unsupervised, while the discriminative part is trained in a supervised manner.

Datasets and Implementation Details: After examining the literature we found three datasets suitable for evaluating our hybrid model. All the three datasets consist of temporal data from two modalities – Audio and Video. In our experiments, we explore the different combinations of hand-crafted features versus the feature representations learned by deep learning. Our experiments allow us to use the CRBMs for learning feature representations from raw pixels/audio spectrograms as well from hand-crafted features.

AVEC [21] is an audio-visual dataset for single person affect analysis. The dataset involves users interacting with emotionally stereotyped virtual characters operated by a human. The visual data contains mainly the face of the user interacting with the character. The Audio data consists of recordings of utterances of the user and is synchronized with the video. The dataset has been annotated with binary labels for four different affective dimensions - Activation, Expectation, Power and Valence. We use the AVEC dataset to compare against [18, 5]. The dataset is divided into two sets, 31 sequences for training[6] and 32 sequences for testing. The dataset comes with pre-computed audio and video features; refer to [21] for details. We apply PCA on the extracted features and reduce each of the audio and video features to 100 dimensions. For each modality we choose a

[6]We did not use the complete sequences because it caused over fitting of our models, rather we have dropped 30% of the frames per sequence and used the rest for training. This was empirically evaluated for achieving the best classification/generation results.

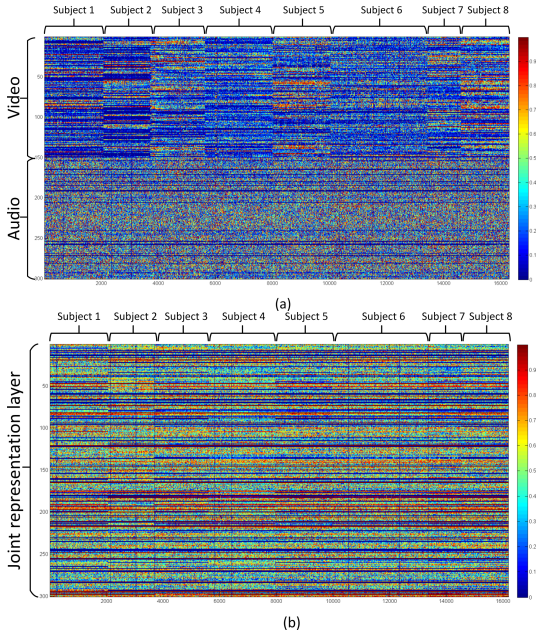

Figure 2. **Inference in Multimodal Fusion:** (a) Feature representation activated by unimodal CRBMs for audio/video on AVEC dataset. (b) Fused feature representation activated by a multimodal CRBM based on the unimodal activations on AVEC dataset. The figures show that multimodal CRBMs can learn a homogeneous joint representation from heterogeneous modalities enabling cross-modal generation of missing data.

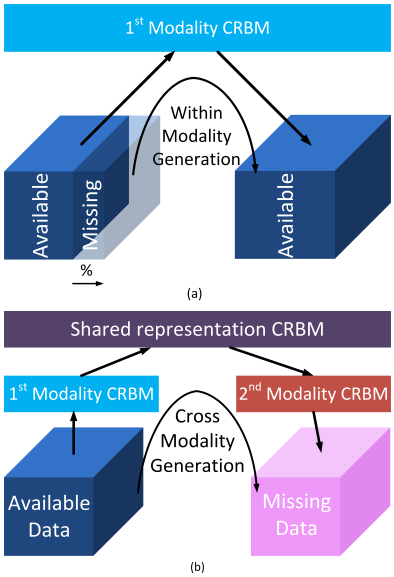

Figure 3. **(a) Within modality generation:** We evaluate the ability of our approach to handle varying degrees of missing data in a single modality (Tab. 2). **(b) Cross modality generation:** We evaluate the ability of our approach to generate missing data across modalities – only one of the modalities is available (Tab. 3).

CRBM with a temporal order $N = 8$, with the first hidden layer being over-complete consisting of 150 nodes, and the multimodal fusion layer consisting of 300 nodes.

AVLetters[11], consists of 10 speakers uttering the letters A to Z, three times each. The dataset also provides pre-extracted 60×80 patches of lip regions along with audio features (MFCC features of 483 dimensions[7]). The dataset is divided into two sets, $2/3$ of the sequences for training and $1/3$ for testing. Following the same setup as in [13], we reduce the dimensionality of the audio features to 100 dimensions using PCA whitening and the video features (lip region) to 32 dimensions. For each modality we choose a CRBM with order $N = 3$ with the first hidden layer being over-complete with 150 nodes and the multimodal fusion layer consisting of 300 nodes.

CUAVE [15], consists of 36 speakers uttering the digits 0 to 9. The dataset provides the aligned face of each speaker of size 75×50, as well as the audio spectrogram and MFCC features of dimensionality 534. The dataset is divided into two sets, $1/2$ for training and $1/2$ for testing. We follow the same experimental setup as in [13]. As with the AVLetters dataset, we reduce the dimensionality of the audio features to 100 dimensions using PCA whitening and the video features (lip region) to 32 dimensions. For each modality we choose a CRBM with order $N = 3$ with the first hidden layer being over-complete with 150 nodes and the multimodal fusion layer consisting of 300 nodes.

[7]The raw audio was not provided for AVLetters dataset.

Quantitative Results: For evaluating our hybrid model, we present two different kinds of evaluations 1) Unimodal Classification Tab. 1 and Multimodal Classification Tab. 1 and 2) Unimodal generation Tab. 2 Fig. 3(a) and Multimodal Generation Tab. 3 Fig. 3(b). Below we go into more details regarding each of the experiments.

Unimodal classification, (Tab. 1) given features from a single modality, the task is to detect events, as previously described. We report the average classification accuracy. As we can see, in each dataset AVEC (AVEC-A, AVEC-V), AVLetters (AVLetters-A, AVLetters-V) and CUAVE (CUAVE-A, CUAVE-V), the CRBM tends to be better than both the RBM and raw features. We can further observe that the CRFs tend to outperform SVMs. This happens due to the temporal component of the data that was explicitly modeled by using CRFs, thereby justifying our choice of CRFs as a discriminative model. Also, notice that CRBMs tend to outperform RBMs and RAW, justifying the use of CRBMs for learning temporal representations of the data.

Unimodal generation (Tab. 2), this experiment allows us to evaluate the ability of the unimodal CRBM to deal with missing data as shown in Fig. 3(a). We remove the last

(10%, 30%) of the data from each sequence and have the model generate the missing data. Subsequently, we predict the class label using both the available and generated data. As seen in Tab 2, our CRBM outperforms the RBM model. As you can see, the CRBM was able to learn a good representation of the data, that allows for dealing with missing data. The drop in accuracy becomes significant when more data is missing.

Multimodal classification (Tab. 6) shows the average accuracy on the AVEC dataset. We can observe that our CRF-CRBM and CRF-RBM models perform comparable to the state-of-the-art [21, 18, 22]. However, a key advantage of our hybrid approach is its ability to generate missing data, enabling it to handle missing unimodal or cross-modal data. None of the other approaches [21, 18, 22] are capable of dealing with missing data. Table 5 shows the classification performance for visual speech recognition on the AVLetters dataset [2]. Our hybrid model shows a substantial improvement over the state-of-the-art which include the hand-engineered features [11, 29] as well as the SVM-RBM model of [13]. Table 4 shows the classification performance for visual speech recognition on the CUAVE dataset [15]. Please note that the models [6, 10, 14], use a visual front-end system that is substantially more complex than ours. In our case, we use the same front end as in [13] which extracts bounding boxes ignoring orientation and perspective changes. In multimodal classification (Tab. 1), we detect events given both modalities. The multimodal classification outperforms unimodal classification, which proves that our model is able to learn a good joint representation of the data.

Multimodal generation (Tab. 3), we evaluate the ability of our model to generate data in a cross modal setup as shown in Fig. 3(b). As seen in Tab 3, our CRBM is able to outperform the RBM model in most cases. Note that chance performance for CUAVE dataset is 10% [13], and for AVLetters is 4%.

6. Conclusion

We have proposed a hybrid model comprising of temporal generative and discriminative models for classifying sequential data from multiple heterogeneous modalities. We employ a deep networks based temporal generative model which enables us to learn a rich feature representation to model the short-term temporal characteristics, while also allowing us to handle missing data. The discriminative component of our model consists of a CRF, which enables modeling long range temporal dependencies leading to a superior classification performance. An extensive experimental evaluation on three different datasets demonstrates the superiority of our approach over the state-of-the-art. In future, we plan to explore a joint framework for simultaneously learning the generative and discriminative components of the hybrid model.

Model	Accuracy
Discrete Cosine Transform [6]	64
Fused Holistic+Patch [10]	77.08
Visemic AAM [14]	83
SVM-RBM [13]	66.7
CRF-RBM	68.6
CRF-CRBM	69.1

Table 4. Classification accuracy using multimodal data on the CUAVE dataset.

Model	Accuracy
Multiscale Spatial Analysis [11]	44.6
LBP [29]	58.85
SVM-RBM [13]	59.2
CRF-RBM	63.8
CRF-CRBM	67.1

Table 5. Classification accuracy using multimodal data on the AVLetters dataset.

Model	mean Accuracy
Baseline-RAW [21]	65.27
SVM-RAW (Late Fusion)[18]	70.55
LDCRF-RAW (Late Fusion)[18]	75.40
PLS-SVM (Late Fusion)[22]	67.37
CRF-RAW (Late Fusion)[22]	69.97
HCRF-RAW (Late Fusion)[22]	69.90
JHCRF-RAW [22]	71.85
CRF-RBM	68.4
CRF-CRBM	70.8

Table 6. Average classification accuracy using multimodal data on the AVEC dataset [21].

Acknowledgments

This work is supported by DARPA W911NF-12-C-0001. The views, opinions, and/or conclusions contained in this paper are those of the author and should not be interpreted as representing the official views or policies, either expressed or implied of the DARPA or the DoD.

References

[1] Y. Bengio. Learning deep architectures for ai. In *FTML*, 2009.

[2] S. Cox, R. Harvey, Y. Lan, and J. Newman. The challenge of multispeaker lip-reading. In *AVSP*, 2008.

[3] G. Druck and A. McCallum. High-performance semi-supervised learning using discriminatively constrained generative models. In *ICML*, 2010.

[4] A. Fujino, N. Ueda, and K. Saito. Semi-supervised learning for a hybrid generative/discriminative classifier based on the maximum entropy principle. In *TPAMI*, 2008.

[5] M. Glodek and et al. Multiple classifier systems for the classification of audio-visual emotional states. In *ACII*, 2011.

Model/Dataset	AVEC-A	AVEC-V	AVEC-AV	AVLetters-A	AVLetters-V	AVLetters-AV	CUAVE-A	CUAVE-V	CUAVE-AV
SVM-RAW	64.8	62.4	67.4	55.8	56.2	58.5	61.5	58.4	65.0
CRF-RAW	**68.1**	**63.5**	**69.9**	**58.4**	**59.3**	**60.0**	**64.3**	**62.0**	**66.8**
SVM-RBM	61.8	63.9	67.8	58.4	62.1	62.9	65.1	61.8	65.4
CRF-RBM	**67.6**	**65.4**	**68.3**	**62.6**	**64.6**	**63.8**	**67.6**	**65.2**	**68.6**
SVM-CRBM	65.8	66.9	68.2	61.2	62.6	64.8	65.3	64.6	66.7
CRF-CRBM	**69.2**	**70.1**	**70.8**	**66.9**	**64.8**	**67.1**	**67.9**	**66.3**	**69.1**

Table 1. Average classification accuracy on unimodal/multimodal data. As you can see our full model CRF-CRBM is able to achieve better results than the other models in most cases. Best performers are highlighted in green, and the better perfoming per type of features is bold.

Model/Dataset	AVEC-A	AVEC-V	AVLetters-A	AVLetters-V	CUAVE-A	CUAVE-V
SVM-RBM (0%)	61.8	63.9	58.4	62.1	65.1	61.8
SVM-CRBM (0%)	**65.8**	**66.9**	**61.2**	**62.6**	**65.3**	**64.6**
SVM-RBM (10%)	48.6	46.5	50.7	54.5	59.7	42.8
SVM-CRBM (10%)	**54.9**	**52.1**	**53.6**	**58.2**	**63.1**	**52.6**
SVM-RBM (30%)	35.5	31.2	39.2	32.1	36.1	31.9
SVM-CRBM (30%)	**42.7**	**40.2**	**45.8**	**41.6**	**43.7**	**41.2**

Table 2. Within modality generation classification accuracy (0%, 10%, and 30% missing), note that the percentage missing is at the end of the sequences as shown in Fig. 3(a).

Model/Dataset	AVEC-A||V	AVEC-V||A	AVLetters-A||V	AVLetters-V||A	CUAVE-A||V	CUAvE-V||A
SVM-RBM	31.2	28.2	27.3	25.1	23.1	19.5
SVM-CRBM	**40.4**	**32.1**	**29.6**	**26.5**	**30.7**	**24.4**

Table 3. Cross modality generation classification accuracy as shown in Fig. 3(b). For example, AVE-A||V, means generate Audio given Video.

[6] M. Gurban and J. P. Thiran. Information theoretic feature extraction for audio-visual speech recognition. In *IEEE Trans. on Sig. Proc.*, 2009.

[7] G. E. Hinton. Training products of experts by minimizing contrastive divergence. In *NC*, 2002.

[8] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*, 2001.

[9] H. Larochelle and Y. Bengio. Classification using discriminative restricted boltzmann machines. In *ICML*, 2008.

[10] P. Lucey and S. Sridharan. Patch based representation of visual speech. In *HCSnet Workshop on the use of Vision in Human-Computer Interaction*, 2006.

[11] I. Matthews and et. al. Extraction of visual features for lipreading. In *TPAMI*, 2002.

[12] A. R. Mohamed and G. E. Hinton. Phone recognition using restricted boltzmann machines. In *ICASSP*, 2009.

[13] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Ng. Multimodal deep learning. In *ICML*, 2011.

[14] G. Papandreou, A. Katsamanis, V. Pitsikalis, and P. Maragos. Adaptive multimodal fusion by uncertainty compensation with application to audiovisual speech recognition. In *TASLP*, 2009.

[15] E. Patterson and et. al. Cuave: A new audio-visual database for multimodal human-computer interface research. In *ICASSP*, 2002.

[16] A. Perina, M. Cristani, U. Castellani, V. Murino, and N. Jojic. Free energy score spaces: Using generative information in discriminative classifiers. In *TPAMI*, 2012.

[17] A. Quattoni, S. Wang, L.-P. Morency, M. Collins, and T. Darrell. Hidden conditional random fields. In *TPAMI*, 2007.

[18] G. Ramirez, T. Baltrusaitis, and L. P. Morency. Modeling latent discriminative dynamic of multi-dimensional affective signals. In *ACII*, 2011.

[19] R. Salakhutdinov and G. E. Hinton. Reducing the dimensionality of data with neural networks. In *Science*, 2006.

[20] M. Schmidt. Ugm: Matlab code for undirected graphical models. 2012.

[21] B. Schuller and et al. Avec 2011 -the first international audio visual emotion challenge. In *ACII*, 2011.

[22] B. Siddiquie, S. Khan, A. Divakaran, and H. Sawhney. Affect analysis in natural human interactions using joint hidden conditional random fields. In *ICME*, 2013.

[23] N. Srivastava and R. Salakhutdinov. Multimodal learning with deep boltzmann machines. In *NIPS*, 2012.

[24] I. Sutskever, G. Hinton, and G. Taylor. The recurrent temporal restricted boltzmann machine. In *NIPS*, 2008.

[25] I. Sutskever and G. E. Hinton. Learning multilevel distributed representations for high-dimensional sequences. In *AISTATS*, 2007.

[26] G. W. Taylor and et. al. Dynamical binary latent variable models for 3d human pose tracking. In *CVPR*, 2010.

[27] G. W. Taylor, G. Hinton, and S. Roweis. Two distributed-state models for generating high-dimentional time series. In *JMLR*, 2011.

[28] L. van der Maaten, M. Welling, and L. Saul. Hidden-unit conditional random fields. In *AISTATS*, 2011.

[29] G. Zhao and M. Barnard. Lipreading with local spatiotemporal descriptors. In *Transactions of Multimedia*, 2009.