AFFECT ANALYSIS IN NATURAL HUMAN INTERACTION USING JOINT HIDDEN CONDITIONAL RANDOM FIELDS

Behjat Siddiquie   Saad Khan   Ajay Divakaran   Harpreet Sawhney

SRI International Sarnoff, Princeton, NJ, 08540
{behjat.siddiquie, saad.khan, ajay.divakaran, harpreet.sawhney}@sri.com

ABSTRACT

We present a novel approach for multi-modal affect analysis in human interactions that is capable of integrating data from multiple modalities while also taking into account temporal dynamics. Our fusion approach, Joint Hidden Conditional Random Fields (JHRCFs), combines the advantages of purely feature level (early fusion) fusion approaches with late fusion (CRFs on individual modalities) to simultaneously learn the correlations between features from multiple modalities as well as their temporal dynamics. Our approach addresses major shortcomings of other fusion approaches such as the domination of other modalities by a single modality with early fusion and the loss of cross-modal information with late fusion. Extensive results on the AVEC 2011 dataset show that we outperform the state-of-the-art on the *Audio Sub-Challenge*, while achieving competitive performance on the *Video Sub-Challenge* and the *Audiovisual Sub-Challenge*.

Index Terms— Affect Recognition, Multimodal Fusion

1. INTRODUCTION

The affective state of a human is a good predictor of his or her intrinsic motivation level and actual performance over a variety of tasks [1]. This has led to increased interest in tracking the physical and affective states of humans for richer, more sophisticated human-computer interfaces. Such tracking presents the challenge of state estimation through accurate and unobtrusive detection of a large number of multi-modal behaviors in real time. Human emotions are inherently subtle and complex. They span multiple modalities such as paralinguistics, facial expressions, eye gaze, various hand gestures, head motion and posture. Each modality contains useful information on its own and humans employ a complex combination of cues from each of these modalities to fully interpret the emotional state of a person. The interactions between multiple modalities combined with the distinctive temporal variations of each modality make automated human emotion recognition an extremely challenging problem.

There has been extensive work on human emotion recognition in recent years [2, 3, 4, 5]. Recognizing that human emotion varies dynamically, several works have used techniques such as HMMs [3] and CRFs (and its variations) [2] for analyzing human emotions. However, in the case of multiple modalities, the majority of the work in automated human affect sensing has focussed on analyzing each different modality in isolation rather than studying the inherent dependencies and relationships across modalities [2, 3]. This is partly due to the limited availability of suitably labeled multi-modal datasets and the difficulty of fusion itself, as the optimal level at which the features should be fused is still an open research question. The work by Ramirez et al. [2] is a prime example in this area. They present Latent-Dynamic Conditional Random Field (LDCRF) [6] based models to infer the dimensional emotional labels from multiple high level visual cues and a set of auditory features [2] and then combine them using late fusion. However, this approach has the disadvantage of losing cross modal correlations due to late fusion i.e. correlations across modalities are explored only after inference on class labels has already been made on the basis of individual modalities.

In this paper we propose a novel sequence labeling approach, Joint Hidden Conditional Random Fields (JHCRFs), that are capable of fusing data from multi-modal observation sequences. We also explore a novel combination of class aware dimensionality reduction techniques followed by Hidden Conditional Random Fields (HCRFs), in case of unimodal data and JHCRFs for the case of multi-modal data. We analyze four different affective dimensions - *Activation*, *Expectancy*, *Power* and *Valence* [7]. We evaluate our approach on the first Audio Visual Emotion Challenge (AVEC 2011) dataset [7], which includes three sub-challenges: *Audio Sub-Challenge*, *Video Sub-Challenge* and *Audiovisual Sub-Challenge*. This enables us to examine the suitability of our approach for analyzing unimodal audio and visual data as well as multi-modal audio-visual data. We show that JHCRFs outperform all of the selected baselines (SVMs, CRFs and HCRFs) and are competitive with the state-of-the-art. Furthermore, we demonstrate that JHCRFs outperform late and early fusion methods with a number of classifiers (SVMs, CRFs, and HCRFs).

The rest of the paper is organized as follows. In Section 2, we describe the approach which includes the dataset, feature extraction, the Partial Least Squares (PLS) based dimensionality reduction (subsection 2.3) and JHCRFs (subsection 2.4). Next, we describe the experiments in Section 3 followed by the conclusion.

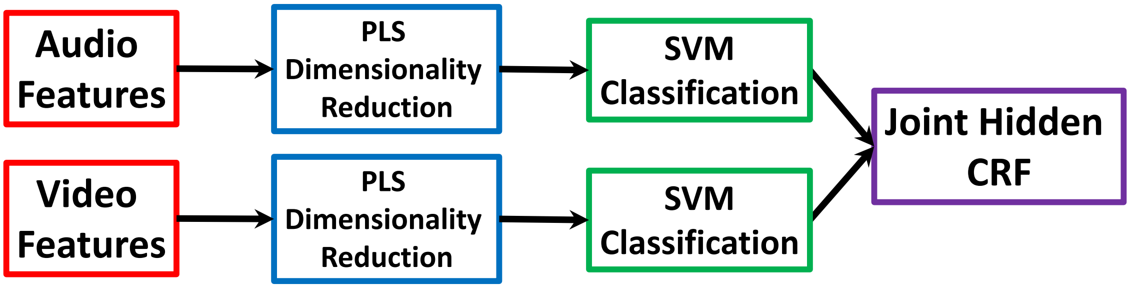


Fig. 1. **Overview of our approach:** We first extract features from the audio and visual streams, followed by PLS based dimensionality reduction. SVMs are used for framewise classification of the low dimensional features, outputs of which are used by JHCRFs for affect recognition.

Fig. 2. Sample frames from the AVEC 2011 dataset.

2. APPROACH

Fig. 1 shows an overview of our approach. The data consists of audio and visual streams captured from human interactions. We first extract features from the audio and visual data, followed by Partial Least Squares (PLS) based dimensionality reduction to help keep the problem tractable. SVMs are used for frame-wise classification of the low dimensional features, outputs of which are used by JHCRFs for affect recognition. In the following we present details of each of these components starting with a description of the dataset used to train and test the approach.

2.1. Dataset

We test our approach on the dataset provided by the Audio/Visual Emotion Challenge and Workshop (AVEC 2011) [7]. The dataset involves users interacting with emotionally stereotyped characters operated by a human. The dataset consists of audio and visual data. The visual data contains the mainly the face of the user interacting with the character, sample frames from the videos are shown in 2.1. The Audio data consists of recordings of utterances of the user and are synchronized with the video. The dataset has been annotated with binary labels for four different affective dimensions *Activation*, *Expectation*, *Power* and *Valence*. The dataset also comes with precomputed audio and video features, which we briefly describe below. See [7] for further details of the dataset.

2.2. Features

2.2.1. Audio Features

The audio feature set consists of 1941 dimensions. This includes energy and spectral related low level descriptors, voicing related low level descriptors and delta coefficients (derivatives) of energy/spectral features. A variety of functionals are computed over each of these audio features over segments corresponding to automatically determined word boundaries.

2.2.2. Video Features

The video features consist of the locations of the face and eye coordinates extracted using the OpenCV implementation of Viola-Jones face/eye detectors. The detected face region is then divided into 10×10 sub-blocks and uniform Local Binary Pattern (LBP) features are extracted from each sub-block. The dimensionality of the video feature vector is 5908.

2.3. PLS based Dimensionality Reduction

In case of both audio and video, the features have a high dimensionality. While it is possible to use the raw features directly in frame-wise approaches such as SVMs, it is not feasible to use them in dynamic methods such as CRFs. In order to reduce the dimensionality of the raw features, we apply a statistical technique known as Partial Least Squares (PLS) [8]. In contrast to other commonly used dimensionality reduction methods such as Principal Components Analysis (PCA), PLS is a supervised dimensionality reduction method and takes into account the class discriminability during dimensionality reduction. Its effectiveness has been demonstrated for dimensionality reduction of high dimensional HOG features for the purpose of human detection [8].

Let X be an $(n \times D)$ matrix containing the features from the training data and Y be an $(n \times C)$ matrix containing the labels of the corresponding instances. Here D is the dimensionality of the features and C is the number of classes. PLS decomposes X and Y such that:

$$\begin{aligned} X &= TP^T + E \\ Y &= UQ^T + F \end{aligned} \tag{1}$$

where T and U are $(n \times p)$ matrices containing the p extracted latent vectors, the $(D \times p)$ matrix P and the $(C \times p)$ matrix Q represent the loadings and the $(n \times D)$ matrix E and the $(n \times C)$ matrix F are the residuals. The PLS method iteratively constructs projection vectors $W_x = \{w_{x1}, w_{x2}, \ldots, w_{xp}\}$ and $W_y = \{w_{y1}, w_{y2}, \ldots, w_{yp}\}$ in a greedy manner. Each stage of the iterative process, involves computing:

$$[\text{cov}(t_i, u_i)]^2 = \max_{\|w_{xi}\|=1, \|w_{yi}\|=1} [\text{cov}(X_{w_{xi}}, Y_{w_{yi}})]^2 \quad (2)$$

where t_i and u_i are the ith columns of the matrices T and U respectively, and $\text{cov}(t_i, u_i)$ is the sample covariance between latent vectors t_i and u_i. This process is repeated until the desired number of latent vectors p, have been determined. PLS produces the projection matrix W_x which is used to project the features to a low dimensional subspace. We employ PLS to learn the projection matrices for audio and video features.

2.4. Joint Hidden Conditional Random Fields

In this subsection, we describe our Joint Hidden Conditional Random Field (JHCRF) technique for discriminative sequence labeling based on fusing temporal data from multiple modalities. Conditional Random Fields (CRFs) have proved to be extremely effective for labeling sequential and temporal data as they offer several advantages compared to earlier approaches for sequence labeling like Hidden Markov Models (HMMs), including the benefits of discriminative learning, the ability to utilize arbitrary features and the ability to model non-stationarity [9]. Furthermore, augmenting CRFs with hidden states [10, 11, 12], increases their representation and modeling power leading to further improvements in performance. However, when presented with temporal data from multiple modalities representing the same sequence, most approaches deal with this by performing either **Early Fusion** which involves fusing the data from multiple modalities and using that as an input for a single CRF or **Late Fusion** which consists of applying multiple CRFs for each modality and then fusing the label probabilities obtained from the individual sequences. However, both these approaches have their disadvantages - in early fusion one modality can dominate the others, while late fusion tends to lose cross-modal information. We address these problems by proposing Joint Hidden Conditional Random Fields (JHCRFs) which are capable of effectively fusing data from multiple modalities while also simultaneously modeling the temporal dynamics of the data.

We now formally describe JHCRFs in detail. Without loss of generality, we assume two modalities since the extension to more than two modalities is straightforward. We are given a set of n sequences, for which we have data from two different modalities $\mathcal{X} = \{X_i\}$ and $\mathcal{Y} = \{Y_i\}$, where $i = 1, 2, \ldots, n$ and each X_i is a sequence $X_i = \{x_1^i, x_2^i, \ldots, x_T^i\}$ of length T, similarly $Y_i = \{y_1^i, y_2^i, \ldots, y_T^i\}$. Here $x_t^i \in \mathbb{R}^{D_x}$ and $y_t^i \in \mathbb{R}^{D_y}$, where D_x and D_y are the dimensionalities of the data from modalities $\mathcal{X}$ and $\mathcal{Y}$ respectively. Corresponding to each sequence $X_i(Y_i)$, we have a sequence of labels $W_i = \{w_1^i, w_2^i, \ldots, w_T^i\}$, with $w_t^i \in \mathcal{C}$, where $\mathcal{C}$ is the set of labels. Let us first describe CRFs and HCRFs, which will help us compare and contrast them against JHCRFs.

Conditional Random Fields (CRFs): CRFs (Fig. 3a) model the conditional distribution over the label sequence given the data as:

$$p(W|X, \theta) = \frac{1}{Z(X, \theta)} \exp(\Psi(X, W; \theta)) \quad (3)$$

here θ are the model parameters, Ψ is the potential function and $Z(X, W, \theta)$ is the partition function that ensures that the model is properly normalized and is defined by $Z(X, \theta) = \sum_W \Psi(X, W; \theta)$. The potential function Ψ is defined as:

$$\Psi(X, W; \theta) = \sum_j \theta_j^t T_j(w_{i-1}, w_i, X, i) + \sum_k \theta_k^s S_k(w_i, X, i) \quad (4)$$

where T_j is a transition feature function and S_k is a state feature function and θ^t and θ^s are the transition and state components of the parameters respectively. Given a new observation sequence X and model parameters θ obtained during training, the predicted label sequence W can be computed as $W = \arg\max_W p(W|X; \theta)$.

Hidden Conditional Random Fields (HCRFs): An approach to improving the performance of CRFs involves augmenting them by introducing hidden variables. The hidden variables model the latent structure increasing the representation power of the model, resulting in an improved discriminative performance. As shown in Fig. 3b, corresponding to each sequence of observations X_i and labels W_i, we introduce a sequence of hidden variables H_i, defined as $H_i = \{h_1^i, h_2^i, \ldots, h_T^i\}$. The hidden CRF is now defined as:

$$p(W|X, \theta) = \frac{1}{Z(X, \theta)} \sum_H \exp(\Psi(X, H, W; \theta)) \quad (5)$$

with the partition function now defined as:

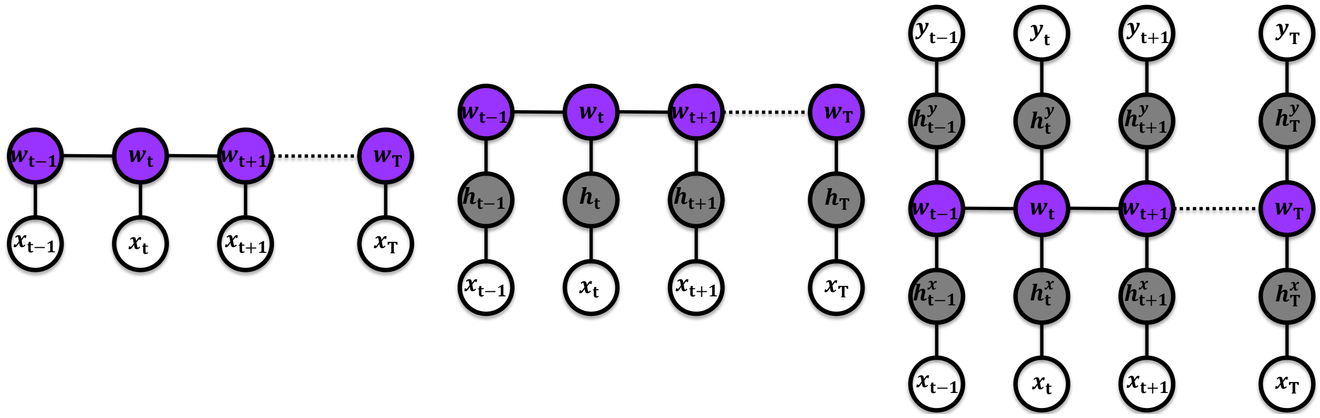
$$Z(X, \theta) = \sum_W \sum_H \Psi(X, H, W; \theta) \quad (6)$$

and the potential function is modified to include state and transition functions for the hidden variables:

$$\Psi(X, H, W; \theta) = \sum_j \theta^{t_i^1} T_j^1(w_{i-1}, w_i, X, i) \quad (7)$$
$$+ \sum_j \theta^{t_j^2} T_j^2(h_i, w_i, X, i) + \sum_k \theta_k^s S_k(h_i, X, i)$$

Training and inference are performed by marginalizing over the hidden variables.

Our HCRF model assigns a label to each node of the sequence and closely resembles Hidden CRF models proposed in [12] and [11] and these are quite different from the Hidden CRF model proposed in [10], which assigns a single label to the entire sequence. Also, note that both CRFs and



a) Standard CRF. b) Hidden CRF. b) Joint Hidden CRF.

Fig. 3. Variations of Conditional Random Fields.

HCRFs can only be applied over data from a single modality. In case of multi-modal data, the most common approach is to use CRFs and HCRFs by resorting to either early or late fusion.

Joint Hidden Conditional Random Fields (JHCRFs): We propose Joint Hidden Conditional Random Fields for fusing temporal data from multiple modalities. We now have two observation sequences X_i and Y_i corresponding to two different modalities. Corresponding to each observation sequence $X_i(Y_i)$, we introduce a sequence of hidden variables $H_i^x(H_i^y)$ as shown in Fig. 3c. The Joint Hidden CRF is defined as:

$$p(W|X,\theta) = \frac{1}{Z(X,\theta)} \sum_H \exp(\Psi(X,H,W;\theta)) \quad (8)$$

where H includes both H^x and H^y. The partition function $Z(X,\theta)$ remains the same as in Eq. 6, while the potential function is modified as follows:

$$\Psi(X,H,W;\theta) = \sum_j \theta_i^{t^1} T_j^1(w_{i-1}, w_i, X, Y, i) \quad (9)$$
$$+ \sum_j \theta_j^{t^2} T_j^2(h_i^x, w_i, X, i) + \sum_j \theta_j^{t^3} T_j^3(h_i^y, w_i, Y, i)$$
$$+ \sum_k \theta_k^{s^1} S_k^1(h_i^x, X, i) + \sum_k \theta_k^{s^2} S_k^2(h_i^y, Y, i)$$

The potential function includes state features S^1 and S^2 corresponding to both sets of hidden states, as well as transition functions T^1 for transitions among the predicted states and T^2 and T^3 for transitions from the hidden states to the predicted states. Therefore JHCRFs simultaneously model and learn the correlations between different modalities as well as the temporal dynamics of sequence labels. Learning and inference are performed by marginalizing over the hidden variables.

2.5. Implementation Details

Our implementations of JHCRFs, HCRFs and CRFs are based on the "Undirected Graphical Models" (UGM) software of [13]. For learning and inference on these models, we use max-product based Loopy Belief Propagation. In case of PLS based dimensionality reduction, we use the first 10 components of the low dimensional subspace, both in case of audio as well as video features. While the low dimensional features obtained by PLS can be used as inputs to the CRF/HCRF/JHCRF, we instead train SVM classifiers for each emotion, on the low dimensional features and use the SVM outputs as inputs to the JHCRF as this was empirically found to give the best results.

3. EXPERIMENTAL RESULTS

We now present the experimental results to demonstrate the effectiveness of our approach. We compare our JHCRF model against several baselines as well as other state-of-the-art techniques. We also compare against several different fusion approaches and show that JHCRF is an effective technique for combining temporal data from multiple modalities. We compare against three baseline methods, which are as follows:

PLS-SVM: This consists of training a non-linear (Radial Basis Function) SVM on the PLS induced low dimensional features. The challenge baseline proposed by Schuller *et al.* [7] is similar and it involves training an SVM over the statistics of raw features.

CRF: This consists of training a Conditional Random Field (CRF) based discriminative classifier over the SVM outputs. Unlike the SVM which looks at each frame in isolation, the CRF takes into account the temporal dynamics of the sequence of features.

HCRF: This involves training a Hidden Conditional Random Field (HCRF) over the SVM outputs. A HCRF differs from a CRF in that it has hidden nodes, which provides it increased representation and modeling power.

The dataset consists of three different sub-challenges - *Audio Sub-Challenge*, *Video Sub-Challenge* and the *Audiovisual Sub-Challenge*. In each case, we use the training set for learning the classifier and report results on the development set.

We use weighted average accuracy as the performance measure, as it is also used in the AVEC 2011 challenge [7]. In the *Audiovisual Sub-Challenge* we demonstrate the effectiveness of fusing multi-modal temporal data. In case of the *Audio Sub-Challenge* and the *Video Sub-Challenge*, the JHCRF model is not applicable, since there is just a single modality present, however we still show that our PLS based dimensionality reduction followed by a standard HCRF model is competitive with the state-of-the-art approach.

3.1. Audio Sub-Challenge

Here we compare our audio only approach against the baselines as well as three state-of-the-art approaches [3, 5, 2]. The results are shown in Table 1. We can see that our approach HCRF, outperforms all the other approaches in terms of mean performance over all the four affective dimensions and thus advances the state-of-the-art. The results also provide us with two important insights. Firstly, note that training an SVM on the PLS induced low dimensional subspace outperforms training an SVM on raw features, while also being computationally much more efficient. Secondly, it can be seen that approaches such as CRFs which model the temporal dynamics of affect, outperform static methods such as SVMs (see Fig. 4). Finally, we can see that Hidden CRFs further improve upon the performance of CRFs, demonstrating the importance of hidden states. These results show that each component of our system - PLS based dimensionality reduction, CRF and HCRF - leads to an increase in performance, thus systematically justifying our design choices.

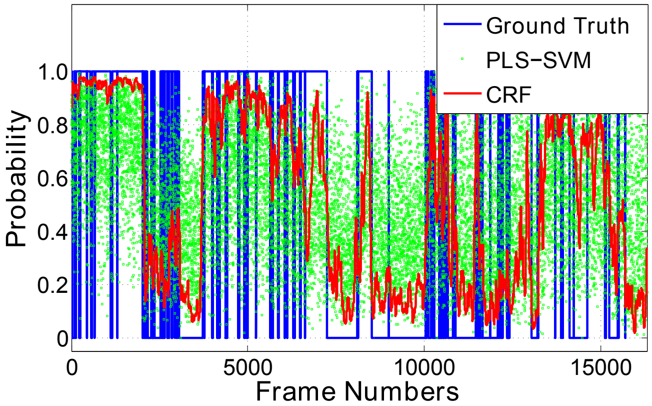


Fig. 4. **SVM vs CRFs:** The blue lines indicate the ground-truth affect (arousal) labels for the *Audio Sub-Challenge*, the green dots denote the frame-wise PLS-SVM probability outputs for arousal and the red curve denotes the probability outputs from a CRF. The plot illustrates that CRFs which analyze the sequence as a whole, outperform static(frame-wise) methods such as SVMs, which can often be very noisy.

WA(%)	**A**	**E**	**P**	**V**	**mean**
baseline [7]	63.7	63.2	65.6	58.1	62.65
[3]	66.9	62.9	63.2	65.7	64.67
[5]	65.1	54.2	61.3	61.8	60.57
CRF [2]	62.9	67.3	67.0	44.6	60.45
LDCRF [2]	74.9	**68.4**	67.0	63.7	68.50
PLS-SVM (ours)	64.6	66.6	66.2	61.9	64.81
CRF (ours)	**76.9**	65.5	68.7	61.7	68.20
HCRF (ours)	73.4	65.5	**68.7**	**70.0**	**69.42**

Table 1. Audio Sub-Challenge Results

3.2. Video Sub-Challenge

In the *Video Sub-Challenge*, we again compare our results against the baselines and three different state-of-the-art approaches [3, 4, 2]. Again, one can see that the PLS dimensionality reduction helps in improving the results over the baseline which is an SVM trained on the raw features. Also, dynamic methods (CRFs and HCRFs) outperform static methods (SVMs). In case of the *Video Sub-Challenge*, while our results are competitive, they are marginally below [2] and [4]. We conjecture that this is due to the additional features used by them. For example, [2] employs additional features such as eye gaze, smile intensity and head tilt, while we use only the LBP features that are provided by the challenge [7] and we believe that the lack of these additional features results in our performance being marginally inferior compared to [2].

WA(%)	**A**	**E**	**P**	**V**	**mean**
baseline [7]	60.2	58.3	56.0	63.6	59.53
CRF [2]	72.3	53.8	46.2	69.5	60.45
LDCRF [2]	**74.5**	60.0	60.3	**72.9**	**66.93**
[4]	69.3	**65.6**	59.9	67.8	65.64
[3]	58.2	53.5	53.7	53.2	54.65
PLS-SVM (ours)	68.1	57.3	55.4	68.9	62.43
CRF (ours)	69.5	59.1	55.3	68.8	63.17
HCRF (ours)	70.1	59.5	55.4	68.8	63.45

Table 2. Video Sub-Challenge Results

3.3. AudioVisual Sub-Challenge

In the *Audiovisual Sub-Challenge*, we compare our results against the state-of-the-art results [3] and [2], as well as early and late fusion over different classifiers. The results in Table 3 show that JHCRF outperforms SVMs, CRFs and HCRFs, demonstrating that JHCRFs are an effective technique for sequence labeling tasks over multiple modalities. Our results are also competitive against the state-of-the-art and we outperform [3] and two different classifier-fusion combinations in [2]. While the results of LDCRF (late fusion) [2] are superior to ours, we believe that this is due to the augmented set of video features employed by them.

We also compare the performance of JHCRFs against different fusion methodologies over multiple kinds of classifiers.

The results are shown in Table 4. First of all, comparing the results in Table 4 against Tables 1 and 2, we can see that with each classifier type, fusion helps in improving results over the individual modalities. The results also demonstrate that JHCRF outperforms both early and late fusion over HCRFs as well as other classifiers, thereby making it a superior alternative to early and late fusion for multi-modal sequence labeling tasks. Finally, we can also observe that late fusion tends to perform better than early fusion does over multiple classifier types.

WA(%)	**A**	**E**	**P**	**V**	**mean**
[3]	69.3	61.7	61.3	68.8	65.27
LDCRF (Early) [2]	79.3	63.4	66.9	62.8	68.10
LDCRF (Late) [2]	81.7	73.1	73.3	73.5	75.40
SVM (Late) [2]	75.4	69.4	65.3	72.1	70.55
PLS-SVM (Late)	67.5	65.8	65.8	70.4	67.37
CRF (Late)	70.9	66.6	65.3	77.1	69.97
HCRF (Late)	70.5	66.5	65.6	77.1	69.90
JHCRF	75.7	66.3	69.1	76.3	71.85

Table 3. Audio-Visual Sub-Challenge Results

WA(%)	**A**	**E**	**P**	**V**	**mean**
PLS-SVM (Early)	68.5	61.9	63.1	70.2	65.91
PLS-SVM (Late)	67.5	65.8	65.8	70.4	67.37
CRF (Early)	73.6	56.2	66.3	69.1	66.27
CRF (Late)	70.9	66.6	65.3	77.1	69.97
HCRF (Early)	70.8	57.6	66.2	74.6	67.29
HCRF (Late)	70.5	66.5	65.6	77.1	69.90
JHCRF	75.7	66.3	69.1	76.3	71.85

Table 4. Audio-Visual Sub-Challenge Fusion Experiments

4. CONCLUSION

We have proposed an approach for emotion recognition based on audio and visual cues. The key novelty of our work is an effective approach for fusing temporal data from multiple modalities. We first perform a PLS based dimensionality reduction on the raw audio and video features. For unimodal data, audio or visual, we apply a Hidden Conditional Random Field on the low dimensional features for emotion recognition. For multi-modal data, we propose a Joint Hidden Conditional Random Filed (JHRCF) model for fusing temporal data from multiple modalities as an alternative to early and late fusion. Extensive results on the AVEC 2011 dataset show that we outperform the state-of-the-art on the *Audio Sub-Challenge*, while achieving competitive performance on the *Video Sub-Challenge* and the *Audiovisual Sub-Challenge*.

Acknowledgement: This work was supported by The Defense Advanced Research Projects Agency under Army Research Office Contract Number W911NF-12-C-0001. The views, opinions, and/or findings 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.

5. REFERENCES

[1] Robert A. Sottilare, “Using student mood and task performance to train classifier algorithms to select effective coaching strategies within intelligent tutoring systems (its),” 2009.

[2] Geovany Ramirez, Tadas Baltrusaitis, and Louis-Philippe Morency, “Modeling latent discriminative dynamic of multi-dimensional affective signals,” in *ACII 2011*, 2011.

[3] Michael Glodek and et al., “Multiple classifier systems for the classification of audio-visual emotional states,” in *ACII 2011*, 2011.

[4] Albert Cruz, Bir Bhanu, and Songfan Yang, “A psychologically inspired match-score fusion model for video-based facial expression recognition,” in *ACII 2011*, 2011.

[5] Jonathan C. Kim, Hrishikesh Rao, and Mark A. Clements, “Investigating the use of formant based features for detection of affective dimensions in speech,” in *ACII 2011*, 2011.

[6] Louis-Philippe Morency, Ariadna Quattoni, and Trevor Darrell, “Latent-dynamic discriminative models for continuous gesture recognition,” in *CVPR*, 2007.

[7] Bjorn Schuller and et al., “Avec 2011 -the first international audio visual emotion challenge,” in *ACII 2011*, 2011.

[8] William Robson Schwartz, Aniruddha Kembhavi, David Harwood, and Larry S. Davis, “Human detection using partial least squares analysis,” in *ICCV*, 2009.

[9] John Lafferty, Andrew McCallum, and Fernando Pereira, “Conditional random fields: Probabilistic models for segmenting and labeling sequence data,” in *ICML*, 2001.

[10] Ariadna Quattoni, Sybor Wang, Louis-Philippe Morency, Michael Collins, and Trevor Darrell, “Hidden conditional random fields,” in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2007.

[11] L.J.P. van der Maaten, M. Welling, and L.K. Saul, “Hidden-unit conditional random fields,” in *AISTATS*, 2011.

[12] Yun-Hsuan Sung and Dan Jurafsky, “Hidden conditional random fields for phone recognition,” in *IEEE workshop on Automatic Speech Recognition and Understanding*, 2009.

[13] Mark Schmidt, “Ugm: Matlab code for undirected graphical models,” 2012.