Step-wise Emotion Recognition Using Concatenated-HMM

Derya Ozkan
Institute for Creative Technologies
University of Southern California
12015 Waterfron DR
Playa Vista, 90094
ozkan@ict.usc.edu

Stefan Scherer
Institute for Creative Technologies
University of Southern California
12015 Waterfron DR
Playa Vista, 90094
scherer@ict.usc.edu

Louis-Philippe Morency
Institute for Creative Technologies
University of Southern California
12015 Waterfron DR
Playa Vista, 90094
morency@ict.usc.edu

ABSTRACT

Human emotion is an important part of human-human communication, since the emotional state of an individual often affects the way that he/she reacts to others. In this paper, we present a method based on concatenated Hidden Markov Model (co-HMM) to infer the dimensional and continuous emotion labels from audio-visual cues. Our method is based on the assumption that continuous emotion levels can be modeled by a set of discrete values. Based on this, we represent each emotional dimension by step-wise label classes, and learn the intrinsic and extrinsic dynamics using our co-HMM model. We evaluate our approach on the Audio-Visual Emotion Challenge (AVEC 2012) dataset. Our results show considerable improvement over the baseline regression model presented with the AVEC 2012.

Categories and Subject Descriptors

I.2 [**Artificial Intelligence**]: Natural Language Processing

General Terms

Algorithms, Performance

Keywords

Emotion recognition

1. INTRODUCTION

Emotion is one of the fundamental elements of human-to-human interaction [37]. The emotional state of a person effects the way he/she communicates with others [8, 27]. Similarly, emotional state of others affect the way the person reacts to them. Therefore, the area of affective computing attracted a lot of attention from diverse research fields such as computer science, psychology, and cognitive science [38]. Affective computing aims at building systems that can recognize, interpret and produce human emotions. Such emotion sensitive systems can improve the way that the machines communicate with humans [28]. For instance, a virtual human can adjust its behavior based on the emotional state of a user. Other application areas of affect computing include commercial fields such as customer services, call centers, game and entertainment [41].

Earlier work on automated emotion recognition [41] has often focused on analysis of the six discrete basic emotions [12] (happiness, sadness, surprise, fear, anger and disgust), even though in everyday interactions people exhibit non-basic and recognizable mental/affective states such as interest, boredom, and confusion [31]. These emotions are often imprecise. In other words, the emotional state is not always one or the other, but has a level that indicates how strong the expressed feeling is. Furthermore, a single label might not describe the complexity of an affective state well. Therefore, there has been a move to analyze audio and video recordings along a set of small number of emotional dimensions. Examples of such affective dimensions are power (sense of control), valence (pleasant vs. unpleasant), activation (relaxed vs. aroused), and expectancy (anticipation). Fontaine *et al.* [15] argue that these four dimensions account for most of the distinctions between everyday emotion categories, and hence form a good set for analysis.

In this paper, we present an approach based on concatenated Hidden Markov Model (co-HMM) to infer the dimensional and continuous emotion labels from multiple high level audio and visual cues. This approach has the advantage of explicitly learning the temporal relationships among the audio-visual data and the emotional labels. The first step of our approach involves generating a step-wise representation of the continuous emotion dimensions, in which we model the distribution of each emotion dimension by a set of discrete labels (see Figure 2). In the second step, we build a generative model, co-HMM, that can estimate the most likely label at each sample. Using the co-HMM model allows us to learn both the intrinsic dynamics within the same class label and extrinsic dynamics among different classes. The affective dimensions analyzed in our work are arousal, expectancy, power, and valence. Our model is evaluated on the Second International Audio/Visual Emotion Challenge (AVEC 2012) dataset. A complete description of the challenge and the dataset can be found in Schuller *et al.* [3].

We evaluate our method on both set of labels: word-level (WLSC), and fully continuous (FCSC). We see an improvement in performance over existing approaches (Support Vector Machine Regression, and uni-modal co-HMM)

ICMI'12, October 22–26, 2012, Santa Monica, California, USA.

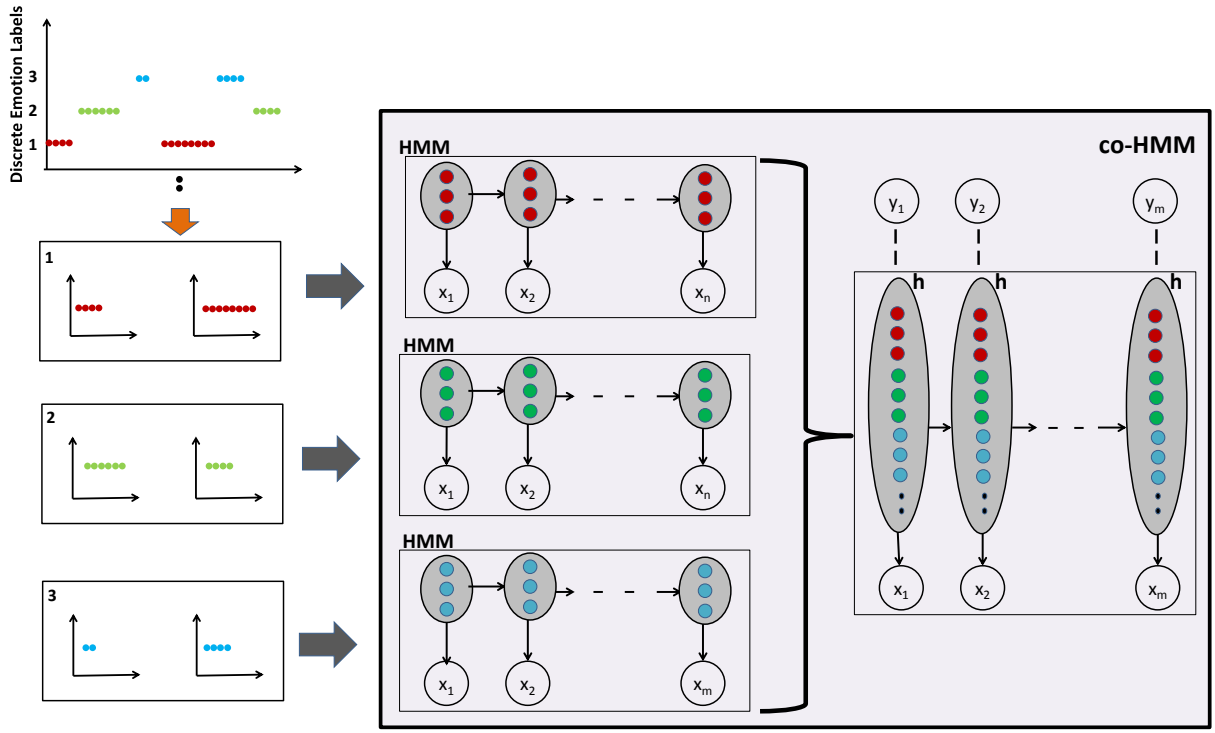

Figure 1: Overview of training procedure for our co-HMM model. First, continuous emotion levels are represented by step-wise labels (see Figure 2). Then, we create sub-datasets containing only sub-sequences with the same label, and one HMM model is learned for each step-wise class label. Finally, the hidden states of these one-label HMMs are concatenated together to form the hidden states of the final co-HMM model. This overview represents our training procedure for one emotional dimension only (i.e. valence) and this same procedure is applied independently on all four dimensions in our dataset.

when evaluating our approach on the development set. Furthermore, when evaluated on the test set our approach considerably improves the baseline results presented from [3].

In the rest of this paper, we first present related works in Section 2. Then, we describe our approach in Section 3. Experimental setup and results are presented in Section 4 and Section 5 respectively. Finally, we conclude in Section 6.

2. RELATED WORK

Several researchers used prosody (i.e pitch, speaking rate, etc.) for speech based emotion recognition [29, 39]. Some studies analyzed visual cues, such as facial expressions and body movements [4, 30, 13]. De Silva et al. [10] and Chen et al. [6] presented one of the early works that integrate both audio and visual information for emotion recognition. We refer readers to following publications on emotion recognition for an extensive survey: [41, 28, 38].

Of special relevance to our work is the work done by Nicolaou *et al.* [25] that presents experiments for classification of spontaneous affect based on Audio-Visual features using coupled Hidden Markov Models. Using coupled-HMMs allow them to model temporal correlations between different cues and modalities. They also show the benefits of using the likelihoods produced from separate coupled-HMMs as input to another classifier, rather than picking the label with a maximum likelihood for audio-visual classification of affective data. Interestingly, their experiments show that visual features contribute more in spontaneous affect classification in the valence dimension.

Wöllmer *et al.* [40] uses Conditional Random Fields (CRF) for discrete emotion recognition by quantising the continuous labels for valance and arousal based on a selection of acoustic features. In addition, they use Long Short-Term Memory Recurrent Neural Networks to perform regression analysis on these two dimensions. Both of these approaches demonstrate the benefits of including temporal information when approaching emotion recognition in dimensional space.

Most of the previous work on emotion recognition (including the studies mentioned above) have focused on classifying human emotional states into discrete labels such as angry, happy or surprised. However, in real life scenarios, human often show imprecise emotions. In other words, emotional have a level of intensity. In this paper, our goal is to not just recognize the binary emotion labels but the strength of the emotion, which we refer to as continuous labels.

Nicolaou et al. [26] recently presented a regression framework for dimensional and continuous emotion recognition. In [22], short term context is used that takes into account the past speech cues. Liscombe et al. [23] showed that using contextual features, such as the structure of spoken dialog and track user state, along with the standard lexical and prosodic features increases the classification accuracy.

In this paper, we present a method for emotion recognition that first builds a step-wise representation of the emo-

tion dimension, and then uses concatenated Hidden Markov Model to learn the dynamics among different class labels of the stepwise representation. Concatenated models, such as concatenated HMMs, have been extensively used in speech recognition and handwriting recognition [19]. For instance, Hu et al. [18] models subcharacter stroke types modeled by HMM's. Then, these HMM models are concatenated together to form letter models for handwriting recognition.

3. APPROACH

In this paper, we propose to use a variant of the Hidden Markov Model (HMM), called concatenated HMM (co-HMM), to recognize affective dimensions in un-segmented video and audio sequences.

We hypothesize that although emotion dimensions are *continuous*, the distribution of these levels of emotion can effectively be modeled by a set of discrete classes for each emotional dimension independently. For example, in Figure 2, originally continuous emotion levels are represented by 6 class labels for arousal. The correlation between original emotion levels and the discrete labels are about 0.8983 on average. Based on that assumption, we use a step-wise label representation of each emotional dimension, in which each continuous emotion level is assigned to a discrete value that can be seen as a class label. This step-wise representation, in practice, provides the most relevant subset of label ranges and helps reducing the effect of noise.

The main idea behind co-HMM is to build a generative model that can estimate the most likely label at each sample (i.e. recognition on unsegmented sequences). We achieve this goal by creating a concatenated HMM model, in which each hidden state is directly associated with a specific label. Figure 1 shows this label-hidden state association through different colors for one emotional dimension. Using the co-HMM model allows us to learn both the intrinsic dynamics within the same class label and extrinsic dynamics among different classes. Intrinsic dynamics are learned through the hidden states of individual HMMs trained on one single class label. By concatenating these individual HMMs, we are able to model the extrinsic dynamics (temporal relationships) among different class labels.

The following section describes the co-HMM model for classification and sub-section 3.2 explains our step-wise label representation process.

3.1 Concatenated HMM

Our concatenated HMM model learns a step-wise representation of continuous emotion dimensions. In the co-HMM, one HMM model is trained for each label class of the step-wise representation and concatenate these one label HMMs. In other words, we train each HMM with segmented subsequences where the frames of each subsequence all belonged to the same label class. For a better understanding of our co-HMM model, we will first briefly introduce Hidden Markov Models and then describe our co-HMM model.

3.1.1 HMMs

Hidden Markov Models are one of the most widely used machine learning technique for modeling sequential data, such as in speech recognition, computational molecular biology, image sequence modeling, and other areas of artificial intelligence and pattern recognition [17]. A Hidden Markov Model learns a probability distribution over a sequence of

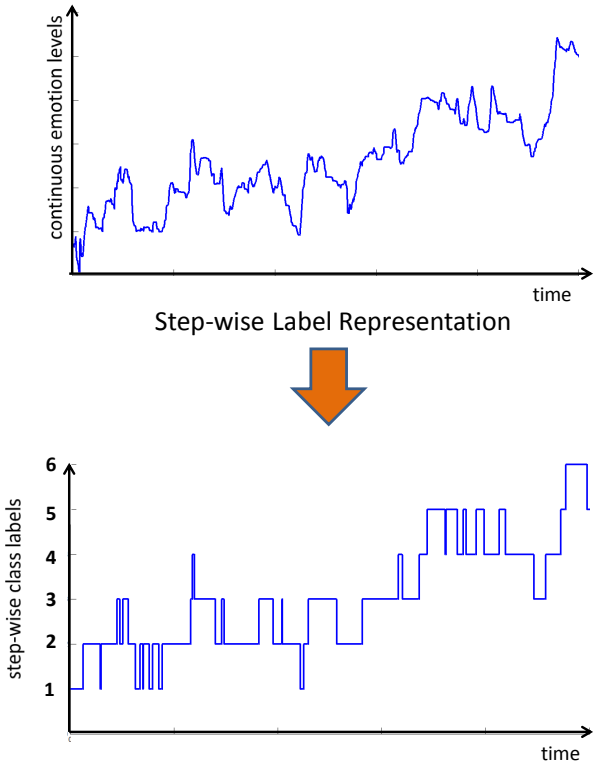

Figure 2: An example of step-wise label representation. The original continuous emotion levels are mapped into 6 label classes for one emotional dimension (i.e. arousal).

observations $\mathbf{x} = \{x_1, x_2, ..., x_m\}$, where each frame observation x_j is represented by a feature vector $\in \mathbf{R}^d$, for example, the audio features at each sample. In a first order HMM, these observations are associated with a set of hidden states $\mathbf{h} = \{h_1, h_2, ..., h_m\}$, where the state of h_t at time t depends only on the previous state t_{t-1} at time $t-1$ and the observation x_t at time t. Graphical representations of HMM models are shown in Figure 1 (middle column).

Based on the above definitions, the joint distribution of state variables and the sequence observations can be found by the following:

$$p(\mathbf{h}, \mathbf{x}) = p(h_1)p(x_1|h_1)\prod_t p(h_t|h_{t-1})p(x_t|h_t) \qquad (1)$$

The first two terms are the model priors (distribution over the initial state). Training of an HMM involves finding the probability distribution over the $K * K$ *state transition matrix* that defines $p(h_t|h_{t-1})$, and the output model that defines $p(x_t|h_t)$, which can be modeled in different ways. In our experiments, we use mixtures of gaussians, since our observations–$\mathbf{x}$– are real values. Note that K is the total number hidden states used in the model. In summary, HMMs are represented by 5 types of parameters: priors, transition matrix, mean and variance of gaussian distribution, and mixing values.

Table 1: Correlations between the original and step-wise labels for different number of label classes (3,4,5,and 6) on the training set.

Discretization Level	Emotion Labels				
	arousal	**expectacy**	**power**	**valence**	**average**
3	0.79092	0.83127	0.75666	0.77554	0.7886
4	0.85866	0.88130	0.82747	0.83778	0.8513
5	0.89305	0.90759	0.85160	0.88595	0.8845
6	0.90992	0.92438	0.85192	0.90691	0.8983

3.1.2 co-HMM

As shown in Figure 1, our co-HMM model is concatenation of multiple HMMs trained on individual labels. In the final co-HMM, the number of hidden states is equal to the sum of hidden states in the one label HMMs. The first step for creating co-HMM model is to divide the dataset into N different sub-datset, where N is the number of label classes (i.e. discrete labels in the step-wise representation). Each sub-datset contains only the sub-sequences of the same class. For example, the third sub-dataset will only contain subsequences from class label 3. Then, for each of these N sub-datsets, we learn a one-label HMM.

In the second step, we concatenate parameters of the one label HMMs. The priors, mean and variance of gaussian, mixture parameters of co-HMM are obtained by simple vector and matrix concatenation. For the transition matrix of the concatenated HMM, we create it in 3 steps. For the first step, we copy the transition matrices of the one label HMMs in a large transition matrix, where the block-wise diagonal of this matrix is the one-label HMM transition matrices. Then, we compute the Viterbi path of each training subsequence using the appropriate one-label HMM, and then count the number of transitions between class labels. In the final step, the counts are inserted in the co-HMM transition matrix,and normalized so that its rows sum to one.

At testing, we apply the forward-backward algorithm on the new sequence, and then sum at each frame the hidden state marginal probabilities associated with each class label y_t.

3.2 Step-Wise Label Representation

In this section, we describe our procedure to create the step-wise label representation and show that these discrete labels correlate with the original continuous emotion levels. As exemplified in Figure 2, we perform label discretization using a percentile approach. For each emotion dimension, we use the labels from the training set to determine the range of continuous emotion levels falling into each discrete label class. This decision depends on the percentage of the continuous levels that have similar (or close) values. For instance, if we want to represent the data with n discrete step-wise labels, then we automatically find 5 thresholds such that each label class contain $1/n$ samples from the training set that have similar continuous levels. These thresholds can then be used to determine the class label for new data (i.e. development set).

The discrete step-wise labels are used to train our co-HMM model. Figure 2 shows an example where the continuous labels are mapped into 6 label classes. To study the effect of the discretization on the accuracy of emotion levels, we computed the correlation between the original continuous emotion levels and the new step-wise labels. Table 1 shows the correlation for different number of discrete class labels. We can see that even with only 3 labels, the correlation is still higher than 0.756 for all dimensions. Note that the baseline models in [3] achieve around 0.112 correlation on average. Therefore, this correlation of 0.756 between step-wise class labels and continuous emotion levels can be considered as high.

In the original data, 66060 unique labels are required to represent the continuous levels. Therefore, this correlation of 0.756 is quiet high considering the amount of labels used in step-wise representation.

4. EXPERIMENTAL SETUP

In this section, we first present our dataset and the audio-visual features used in our experiments. We then describe the training and validation methodology. Finally, we explain how we obtain the word level labels from fully-continuous emotion data.

4.1 Dataset

For all our experiments we used the dataset provided by Schuller *et al.* [3]. The dataset consist of 95 video and audio recorded dyadic interaction sessions between human participants and a virtual agent operated by a human. The dataset consists of upper body video segments with per frame and audio and audio-visual segments with per word binary labels along the four affective dimensions (activation, expectation, power and valence).

The dataset contains two sets of labels: the fully-continuous levels which are sampled at 50Hz and the word-level labels which are only defined during spoken words (see [3] for details).

4.2 Audio-Visual Features

When approaching the challenging problem of recognizing affective dimensions in un-segmented video and audio sequences, one valid approach is to experiment with an extensive set of visual or audio features, where each feature is a low-level representation of the instantaneous appearance of the face or a low level descriptor of the audio signal. The problem with this approach is that the feature space will end up extremely large (5908 dimensions of visual and 1841 of audio features in the case of Schuller *et al.* [3]). This high dimensionality issue can be partially solved by performing dimensionality reduction or feature selection.

In this paper, we approach the problem by using a smaller set of features inspired from previous literature on emotion and human communication. The following three subsections describe the audio, video and time features used in our experiments.

Table 2: Experimental results on the test and devel sets for fully-continuous and word-level datasets.

Sampling	Dataset	Model	Arousal	Expectancy	Power	Valence	Mean
fully-continuous	devel	co-HMM	0.3964	0.2464	0.4755	0.2348	**0.3383**
		SVR [3]	0.181	0.148	0.084	0.215	0.157
	test	co-HMM	0.3248	0.3107	0.4506	0.1825	**0.3171**
		SVR [3]	0.141	0.101	0.072	0.072	0.112
word-level	devel	co-HMM	0.2092	0.2397	0.2893	0.2079	**0.2365**
		SVR [3]	0.018	0.009	0.001	0.002	0.007
	test	co-HMM	0.1431	0.2874	0.2874	0.1699	**0.2003**
		SVR [3]	0.021	0.028	0.009	0.004	0.015

4.2.1 Audio Features

Our audio features include the following measures:

- **Energy (in dB)** is a measure of the intensity of the speech signal. Higher values indicate louder speech.
- **Articulation rate** is calculated by identifying the number of syllables per second. The syllables are detected by identifying vowels in the speech. Articulation rate is extracted following the algorithm in [9].
- **Fundamental frequency (f_0)** is the base frequency of the speech signal. It is the frequency the vocal folds are vibrating at during voiced speech segments. f_0 was extracted following the algorithm in [11].
- **Peak slope** is a measure suitable for the identification of breathy to tense voice qualities. Values closer to zero are considered as more breathy. Peak slope parameters were extracted as explained in [20].
- **Spectral stationarity** is a value that captures the fluctuations and changes in the voice signal. High values indicate a stable vocal tract and little change in the speech (e.g. during a hesitation or sustained elongated vowels). It is a measure of the speech monotonicity and is extracted as explained in [14].

The selection and choice of features is motivated mainly by related work and previous research [32, 7, 33, 34]. Further, they have proven to be robust representatives of the targeted prosodic phenomena. In particular, variations in speech energy have been associated with varying emotional states (e.g. higher energy is related to high activation emotional states and lower energy is associated with low activation and low power emotional states), as mentioned in [32]. Similarly to speech energy, articulation rate, syllable durations and pause variations are related to emotional states [32]. Further, a multitude of emotion recognition and affective computing studies successfully used fundamental frequency (f_0) and its variations in their approaches and evaluations [32, 7]. In [24] and [7] the importance of voice qualities for emotion recognition are investigated and reported. We chose the peak slope parameter for the representation of breathy to tense voice qualities as it has proven to be very robust and successful in voice quality classification tasks [34]. Lastly, the spectral stationarity measure is used as an indicator for monotonicity in speech which is associated with low activity and negative valence [32].

4.2.2 Video Features

We selected a subset of visual communicative signals which were shown to be useful when analyzing dyadic interactions [2, 21, 1] and could be estimated robustly by an off-the-shelf sensing software. In our experiments, we processed each video sequence with the Omron OKAO Vision software library [36] to automatically extract the following four facial features: horizontal eye gaze direction (degrees), vertical eye gaze direction (degrees), smile intensity (from 0-100) and head tilt (degrees). We reason that eye gaze and head movements can help to better recognize emotion [41, 35]. In addition, by using low dimensional visual features, we can better learn the temporal relationships among these features.

4.2.3 Time Feature

It is often observed that the emotional state of a participant is ambiguous at the beginning of a conversation. As the participants perceive the context of the conversation, they start expressing their emotions. The more the participants get engaged to the conversation, the more intense their emotions might get. To take this observation into account, we use the time information as a feature in our models. This feature is in the form of a scalar value that represents the frame number for each single frame observation (i.e. number of frames since the beginning of the video).

4.3 Baseline Models

We compare our co-HMM approach to two baseline approaches as follows:

SVR: The baseline proposed by Shuller et al. [3] uses Support Vector Machine for regression (SVR). Different than the SVR model used there, we trained our baseline SVRs with the 3 sets of features described above (audio, video and time). In our experiments, we use the libSVM library [5].

Uni-modal co-HMMs: We train one co-HMM using uni-modal feature sets. In other words, we learn 3 co-HMM models, where each model is trained using either only the audio, video or the time features. The main purpose of selecting these baseline models is to analyze the effect of combining multimodal information for emotion recognition.

4.4 Methodology

For all our experiments we use the data provided by Schuller *et al.* [3]. The data is divided into 3 subsets: training, development and testing. The training set consists of 31 sessions, while the development set consists of 32 sessions that were used for validation of the model parameters. The test set consists of 32 audio-visual sequences. The test sequences did not have any publicly available labels. The training was

Table 3: Comparison with different models on the development set for fully-continuous data.

Model	Features	Correlations				
		arousal	expectancy	power	valence	average
SVR	audio-visual-time	0.017299	0.0021259	0.0057934	0.020038	0.0113
co-HMM	audio only	0.25846	0.081497	0.34215	0.16722	0.2123
co-HMM	video only	0.11698	076478	0.06298	0.20049	0.1142
co-HMM	time only	0.18634	0	0.048736	0.084662	0.0799
co-HMM	audio-visual-time	0.39640	0.24641	0.47551	0.23482	**0.3383**

Table 4: Comparison with different models on the development set for word-level.

Model	Features	Correlations				
		arousal	expectancy	power	valence	average
SVR	audio-visual-time	0.020329	0.022021	0.0081474	-0.02148	0.0073
co-HMM	audio only	0.14304	0.077095	0.093099	0.13166	0.1112
co-HMM	video only	0.080739	0.087916	0.023905	0.18289	0.0939
co-HMM	time only	0.13474	0	0	0.087367	0.0555
co-HMM	audio-visual-time	0.2092	0.2397	0.2893	0.2079	**0.2365**

performed on the training dataset and validated on development set. We automatically validated the following model parameters: number of hidden states for the co-HMM models was validated with values (2-7), and the number of gaussian mixtures was validated with values (1-4). We also validate the number of label classes (3-6) in the step-wise label representation using the development set.

The co-HMM model is implemented using MATLAB and the uni-modal HMMs were trained using Kevin Murphy Toolbox [16]. The input audio-visual features were computed at 50 frames per second. We first down sample all these input observations to 25 frames per second before training our models to reduce training time and memory requirements. Given the continuous labeling nature of our concatenated-HMM model, prediction outputs are also computed at 25Hz. Once we get these output labels, we up sample them back to the original frame rate of 50fps.

4.5 Word-level Labels

Labels for word-level can be obtained in two ways. In the first approach, we all co-HMM models are trained on the fully-continuous emotion levels. This provides us labels at 50 frames per second. To get the word-level labels, we use the word timings to locate the start and end frames of each word in the output co-HMM labels. Then, the average of these frames are used to assign the corresponding world label. In the second approach, we use word-level co-HMM to directly to learn the word-level labels through word level features and compare it to the first approach in the experiments.

5. RESULTS

In this section, we present an discuss our experimental results on both word-level and fully-continuous dataset in the following two subsections.

5.1 Fully-continuous Data

In our first set of experiments, we compare the co-HMM model with the regression model (SVR) presented in [3]. The correlation values for both test and devel sets are given in Table 2. Remark that the SVR model in [3] uses 1188 video and 1841 audio features that are different than the 11 features that we use in our co-HMM approach. We achieve 0.3171 correlation on average for the test set, whereas the baseline SVR model gives 0.112.

Our second set of experiments on the fully-continuous dataset consist of comparison with our baseline models (see Section 4.3). The main goal of these experiments is to study the influence of different modalities on our results. In these experiments, we learn one regression model (SVR) using the same set of multi-modal features (audio, video and time) as in our co-HMM model. We also train separate co-HMMs using only one of the three modalities. Correlation result for each of these models along with the co-HMM model using all the multi-modal features are given in Table 3. Labels for AVEC 2012 test set are not provided with the dataset; and participants of the challenge had a limited number of 5 submissions to evaluate their model on the test set. Therefore, all these models are tested on the development set.

The co-HMM model outperforms the regression model (SVR) in all emotion dimension even if only the features from one modality is used. This implies that our co-HMM model is able to learn the temporal relationships among input features. Furthermore, our step-wise feature representation allows us to model the most relevant subset of label ranges by removing noise.

Another interesting result is that the performance of our co-HMM approach is the best when all three modalities are used together. This indicates that these features contain complementary information relevant to human emotion recognition.

5.2 Word-level Data

For word-level data, we first compare our co-HMM model with the SVR and uni-modal co-HMM models similar to fully-continuous data. The results are shown in Table 4. These results also implies the importance of modeling temporal relationships among features and combining all 3 different multi-modal information for emotion recognition.

Performance of our co-HMM model and the baseline SVR model in [3] is listed in Table 2. We get 0.2003 correlation by using the co-HMM approach, whereas the baseline is 0.015 on the test set.

Note that all these word-level labels are computed from the fully-continuous output labels as described in Section 4.5.

Another option for word-level label computation would be to use one feature per word and model the word-level labels using the same co-HMM model for fully-continuous data. For this approach, we average all the audio-visual features that fall into the time period where a word is spoken. Then, we use these average audio-visual features to train a co-HMM.

Using this later approach for word-level labels, we get the following correlations for arousal, expectancy, power and valence respectively on the test set: 0.1216, 0.2162, 0.1861, 0.0122. The average of these values is 0.1340, which is not as good as our original word-level correlation of 0.2003. This results shows us that the co-HMM model is able to better learn the temporal relationships on frame level than word level.

6. CONCLUSIONS

In this paper, we proposed an approach that models dimensional and continuous human emotions using concatenated Hidden Markov Models. Our approach relies on the assumption that continuous emotion levels can be model by a set of discrete class labels. Based on this, we first represent our data by some step-wise labels and use them to train our co-HMM model. By using the co-HMM model, we are able to learn both the intrinsic dynamics within each class label, and the temporal relationships among these labels.

We evaluated our approach on the Audio-Visual Emotion Challenge (AVEC 2012) dataset. Our results show considerable improvement over the regression model presented in [3] on the test set. We achieve an average of 0.3171 correlation on the test set, which can be considered as high taking into account that continuous emotion recognition is a difficult problem. Using a step-wise representation, we are able to discover the range of the most relevant subset of label ranges helps reducing the effect of noise. Therefore, we believe that our proposed approach can scale to more realistic settings.

We also compared our model to uni-model co-HMM, and have seen that using all three modalities (audio, video and time) improves the overall recognition process. This shows that our co-HMM model is able to learn the temporal relationships among input features, and these features contain complementary information relevant to human emotion recognition.

7. ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant No. 0917321 and the U.S. Army Research, Development, and Engineering Command. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or the Government.

8. REFERENCES

[1] M. Argyle and J. Dean. Eye-contact, distance and affiliation. *Sociometry*, 28(3):233–304, September 1965.

[2] J. B. Bavelas, L. Coates, and T. Johnson. Listeners as co-narrators. *Journal of Personality and Social Psychology*, 79(6):941–952, December 2000.

[3] Florian Eyben Roddy Cowie Maja Pantic Björn Schuller, Michel Valstar. Avec 2012 Ű- the continuous audio/visual emotion challenge. In *to appear in Proc. of Second International Audio/Visual Emotion Challenge and Workshop (AVEC 2012), Grand Challenge and Satellite of ACM ICMI 2012*. ACM, October 2012.

[4] A. J. Calder, A. M. Burton, P. Miller, A. W. Young, and S. Akamatsu. A principal component analysis of facial expressions. *Vision research*, 41(9):1179–1208, April 2001.

[5] Chih-Chung Chang and Chih-Jen Lin. Libsvm : a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):1–27, April 2011.

[6] L. S. Chen, T. S. Huang, T. Miyasato, and R. Nakatsu. Multimodal human emotion/expression recognition. In *Proceedings of the 3rd. International Conference on Face & Gesture Recognition*, pages 366–371, Washington, DC, USA, 1998. IEEE Computer Society.

[7] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J.G. Taylor. Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine*, 18(1):32–80, January 2001.

[8] C. Darwin. *The expression of the emotions in man and animals*. University of Chicago Press.

[9] N. H. De Jong and T. Wempe. Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior research methods*, 41(2):385–390, May 2009.

[10] L C De Silva, T Miyasato, and R Nakatsu. Facial emotion recognition using multi-modal information. 1:397–401, September 1997.

[11] T. Drugman and A. Abeer. Joint robust voicing detection and pitch estimation based on residual harmonics. In *Proceedings of 12th Annual Conference of the International Speech Communication Association (Interspeech)*, pages 1973–1976. ISCA, Italy, August 2011.

[12] Paul Ekman. An argument for basic emotions. *Cognition & Emotion*, 6(3):169–200, May 1992.

[13] I.A. Essa and A.P. Pentland. Coding, analysis, interpretation, and recognition of facial expressions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 19(7):757 –763, July 1997.

[14] S. Finkelstein, S. Scherer, A. Ogan, L.-P. Morency, and J. Cassell. Investigating the influence of virtual peers as dialect models on students' prosodic inventory. In *Workshop on Child, Computer and Interaction (WOCCI'12)*. ISCA, Oregon, September, 2012.

[15] J. R. Fontaine, K. R. Scherer, E. B. Roesch, and P. Ellsworth. The world of emotion is not two-dimensional. *Psychological Science*, 18:1050–1057, 2007.

[16] Hidden Markov Model (HMM) Toolbox for Matlab. software. `http://www.cs.ubc.ca/~murphyk/Software/HMM/hmm.html`.

[17] Zoubin Ghahramani. An introduction to hidden markov models and bayesian networks. *IJPRAI*, 15(1):9–42, February 2001.

[18] Jianying Hu, Michael K. Brown, and William Turin. Hmm based on-line handwriting recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18:1039–1045, 1996.

[19] Daniel Jurafsky and James H. Martin. *Speech and Language Processing*. Prentice Hall, second edition, 2008.

[20] J. Kane and C. Gobl. Identifying regions of non-modal phonation using features of the wavelet transform. In *Proceedings of Interspeech 2011*, pages 177–180. ISCA, Italy, August, 2011.

[21] Krämer N. C. *Human behavior in military contexts*, chapter Nonverbal Communication, pages 150 – 188. Washington: The National Academies Press, July 2008.

[22] Chi-Chun Lee, Carlos Busso, Sungbok Lee, and Shrikanth S. Narayanan. Modeling mutual influence of interlocutor emotion states in dyadic spoken interactions. In *Proceedings of InterSpeech*, Brighton, UK, September 2009.

[23] Jackson Liscombe and et al. Using context to improve emotion detection in spoken dialog systems. In *Proceedings of Interspeech*, pages 1845–1848, Portugal, September, 2005.

[24] M. Lugger and B. Yang. Cascaded emotion classification via psychological emotion dimensions using a large set of voice quality parameters. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 4945–4948. IEEE, Las Vegas, February 2008.

[25] M.A. Nicolaou, H. Gunes, and M. Pantic. Audio-visual classification and fusion of spontaneous affective data in likelihood space. In *ICPR*, 2010.

[26] Mihalis A. Nicolaou, Hatice Gunes, and Maja Pantic. Output-associative rvm regression for dimensional and continuous emotion prediction. In *IEEE International Conference on Automatic Face & Gesture Recognition and Workshops, FG 2011*, pages 16–23, USA, March 2011. IEEE Communications Society.

[27] Keith Oatley, Dacher Keltner, and Jennifer M. Jenkins. *Understanding Emotions*. Wiley-Blackwell, March 2006.

[28] Maja Pantic and Leon J. M. Rothkrantz. Toward an affect-sensitive multimodal human-computer interaction. In *Proceedings of the IEEE*, volume 91, pages 1370–1390, September 2003.

[29] Thomas S. Polzin and Alex H. Waibel. Recognizing emotions in speech. In *The Fourth International Conference on Spoken Language Processing (ICSLP)*, USA, October 1996.

[30] M. Rosenblum, Y. Yacoob, and L.S. Davis. Human expression recognition from motion using a radial basis function network architecture. *Neural Networks, IEEE Transactions on*, 7(5):1121 –1138, September 1996.

[31] Paul Rozin and Adam B. Cohen. High frequency of facial expressions corresponding to confusion, concentration, and worry in an analysis of naturally occurring facial expressions of Americans. *Emotion*, 3(1):68 – 75, March 2003.

[32] K. R. Scherer, T. Johnstone, and G. Klasmeyer. *Handbook of Affective Sciences - Vocal expression of emotion*, chapter 23, pages 433–456. Affective Science. Oxford University Press, January 2003.

[33] S. Scherer. *Analyzing the User's State in HCI: From Crisp Emotions to Conversational Dispositions*. PhD thesis, Ulm University, 2011.

[34] S. Scherer, J. Kane, C. Gobl, and F. Schwenker. Investigating fuzzy-input fuzzy-output support vector machines for robust voice quality classification. *Computer Speech and Language*, June 2012.

[35] Nicu Sebe, T. Gevers, I. Cohen, and T. S. Huang. Multimodal approaches for emotion recognition: A survey. 5670:56–67, November 2004.

[36] OKAO Software. `http://www.omron.com/r_d/coretech/vision/okao.html`.

[37] Fussell SR. *The verbal communication of emotion: introduction and overview*. NJ: Lawrence Erlbaum associates Inc. Publishers, 2002.

[38] Jianhua Tao and Tieniu Tan. Affective computing: A review. In *Affective Computing and Intelligent Interaction*, volume 3784 of *Lecture Notes in Computer Science*, chapter 125, pages 981–995. Springer Berlin / Heidelberg, 2005.

[39] Raquel Tato, Rocío Santos, Ralf Kompe, and J. M. Pardo. Emotional space improves emotion recognition. In *7th International Conference on Spoken Language Processing, ICSLP*, pages 2029–2032, USA, September 2002.

[40] Martin Wöllmer, Florian Eyben, Stephan Reiter, Björn Schuller, Cate Cox, Ellen Douglas-Cowie, and Roddy Cowie. Abandoning emotion classes - towards continuous emotion recognition with modelling of long-range dependencies. In *9th Annual Conference of the International Speech Communication Association (Interspeech)*, pages 597–600. ISCA, Australia, September 2008.

[41] Zhihong Zeng, Maja Pantic, Glenn I. Roisman, and Thomas S. Huang. A survey of affect recognition methods: Audio, visual and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(1):39–58, January 2009.