A HIERARCHICAL FRAMEWORK FOR MODELING MULTIMODALITY AND EMOTIONAL EVOLUTION IN AFFECTIVE DIALOGS

Angeliki Metallinou, Athanasios Katsamanis, Shrikanth Narayanan

Signal Analysis and Interpretation Lab (SAIL), University of Southern California, Los Angeles, CA

metallin@usc.edu, nkatsam@sipi.usc.edu, shri@sipi.usc.edu

ABSTRACT

Incorporating multimodal information and temporal context from speakers during an emotional dialog can contribute to improving performance of automatic emotion recognition systems. Motivated by these issues, we propose a hierarchical framework which models emotional evolution within and between emotional utterances, i.e., at the utterance and dialog level respectively. Our approach can incorporate a variety of generative or discriminative classifiers at each level and provides flexibility and extensibility in terms of multimodal fusion; facial, vocal, head and hand movement cues can be included and fused according to the modality and the emotion classification task. Our results using the multimodal, multi-speaker IEMOCAP database indicate that this framework is well-suited for cases where emotions are expressed multimodally and in context, as in many real-life situations.

Index Terms— hierarchical HMM, multimodality, dialog modeling, discriminative training, emotion recognition

1. INTRODUCTION

Automatic emotion recognition is an emerging research area with various applications, e.g., in the development of educational software [1] and in behavioral informatics [2]. The challenges in recognizing emotional expressions often stem from the fact that emotions are complex dynamic processes that are expressed by multiple modalities, e.g., via facial expressions and speech prosody, which may be carrying complementary, or even conflicting information [3]. Moreover, most real-life emotions are expressed in a particular context, which can usually be very informative about the psychological state of the people involved. Therefore, handling multimodality and incorporating context awareness in emotion recognition systems are key issues. This paper proposes a flexible, hierarchical framework which can both exploit multiple modalities by fusing the corresponding cues appropriately, and also consider temporal context by modeling the emotional evolution in a dialog.

Researchers have recognized the importance of multimodality in order to obtain a more complete description of the expressed emotion [4]. Recent works have combined vocal, facial and body orientation cues for recognizing emotions [5] or social behaviors such as approach-avoidance[2]. In addition, taking into account some form of contextual information, such as temporal emotional evolution [6] or general conversational context [1], is an emerging topic in the emotion recognition literature. In [7] authors used speech cues from the past utterance of a speaker and his interlocutor to inform emotion recognition of the current utterance, while in our previous work[6], we modeled the evolution of emotions of a single speaker and exploited this (somewhat constrained) temporal context when classifying the currently expressed emotion.

Building upon this past work, we propose a hierarchical framework that enables modeling emotional dynamics at both the utterance level, i.e., within an emotion, and at the dialog level, i.e., between the emotions of a speaker or of both speakers that are expressed during a dyadic conversation. Through this high level modeling we incorporate temporal emotional context information into our system. The framework is flexible as to the classification approaches that can be applied to model the affective content of multiple modalities, such as face, voice, head and hand movement, and can be extended to include more modalities if they become available. In this work, we utilize Hidden Markov Model (HMM) classifiers to model utterance and dialog emotional evolution, therefore our approach could be seen as a two-level context-sensitive HMM.

As a testbed for evaluating the proposed approach, we use the Interactive Emotional Dyadic Motion Capture (IEMOCAP) database which is a multimodal and multispeaker database of improvised dyadic interactions [8]. Our experiments are organized in a dyad-independent manner to simulate real-life scenarios where no prior knowledge of the specific dyads is available. Our results indicate that our approach can successfully accomodate a variety of classifiers and fusion strategies, and can handle cases where a different amount of multimodal information is available from each speaker. These results are generally superior to the ones reported in our previous work for valence and activation classification tasks [6].

2. DATABASE

The IEMOCAP database [8] contains approximately 12 hours of audio-visual data from five mixed gender pairs of actors. It comprises recordings of improvised affective dialogs, where scripts and hypothetical scenarios were used to elicit emotions that resemble natural emotion expression and are generated in context. Apart from speech, the recorded streams include detailed face information, head and hand movement cues, obtained by a Motion Capture (MoCap) system for one of the speakers in each dialog. The speakers in each pair take turns in wearing the markers across the dialogs, so that we have a similar amount of audio-visual recordings for each speaker. Audio information is available for both participants in every dialog through two shotgun microphones directed at each one of them.

Dyadic sessions of approximately five-minute duration were recorded and were later manually segmented into utterances. Each utterance was annotated using categorical emotional tags as well as dimensional ratings of valence and activation by human annotators, two or three at least per utterance. The dimensional label of an utterance is on a scale from one to five and the final dimensional score

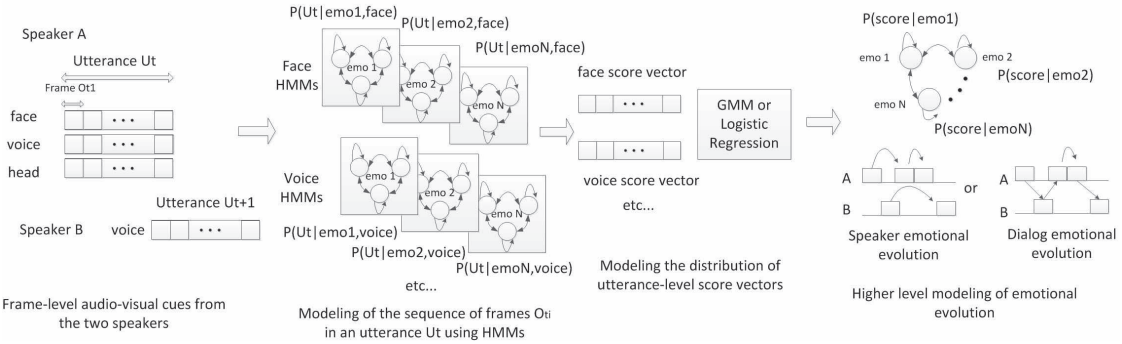

Fig. 1. *An overview of the proposed hierarchical framework.*

is the average over the multiple available (two or three) annotator scores. We perform classification into three levels of valence and activation: level 1 (emo_1) contains ratings in the range [1,2], level 2 (emo_2) contains ratings in the range (2,4) and level 3 (emo_3) contains ratings in the range [4,5]. These intuitively correspond to low, medium, high activation, and to negative, neutral and positive valence respectively. We focus on the classification of dimensional labels, which enables us to make use of all the available data, even of utterances for which there was no categorical inter-annotator agreement, and thus no categorical label exists.

3. HIERARCHICAL MODELING APPROACH

Our problem can be posed as a two-level one: modeling the sequence of audio-visual observations $\{O_{t1}, O_{t2}, ..., O_{tn}\}$ belonging to an emotional utterance U_t and on top of that modeling the sequence of utterances $\{U_1, U_2, ..., U_T\}$ belonging to an emotional conversation C. We assume that an emotional utterance can be described by a single emotional label while an emotional conversation may contain arbitrary emotional transitions between utterances. Our proposed approach, illustrated in Fig. 1, can be seen as a two-level HMM, and shares similarities with the (unimodal) multi-level HMM proposed in [9]. At the lower level, the system processes multimodal cues during each of the speaker's utterances which are modeled using emotion specific HMMs to capture the dynamics within emotional categories. At the higher level, which represents the temporal emotional context, the emotional flow between utterances is modeled by a conversation-level HMM which transitions between emotions. More specifically, the overall system comprises:

Emotional Utterance Modeling: The system takes as input audio-visual cues from each speaker during his utterance. Here, we consider vocal cues for both interlocutors, and for one of the speakers we also consider visual cues: facial expressions, head movement and hand movement. Emotion-specific, 3-state, fully-connected HMMs are trained for each modality, using the HTK Toolbox [10].

Modeling of utterance-level score vectors: We estimate the log-likelihood $s_{it} = \log P(U_t|\lambda_i), i = 1, .., N$ of every emotion-specific HMM λ_i given each utterance. N is the number of emotional categories. We collect these likelihood scores to create an N-dimensional utterance-level score vector $[s_{it}]_{i=1}^{N}$. The joint distribution $P([s_{it}]_{i=1}^{N}|emo_i)$ of these scores is then modeled for each emotional category separately. Two potential models are examined, namely emotion-specific Gaussian Mixture Models (GMMs) and multinomial nominal or ordinal regression. The latter can be used when our classification task is of ordinal nature (e.g., levels of activation). The purpose of this joint score modeling step is to also exploit useful relations that may potentially exist between scores obtained at the first level. For example, we would like an utterance that scores high for emo_1 and low for the other two classes to more strongly qualify for the class emo_1 rather than one that has scored higher for emo_1 but also relatively high for other categories as well. Alternatively, this joint score modeling could be seen as a score normalization process that takes into account the scores from all utterance-level emotional models.

Speaker and Dialog emotional modeling: Speaker context can be included by modeling the evolution of a speaker's emotional state, while dialog context is included by modeling the evolution of emotional states in a dialog in a speaker independent manner. Speaker modeling incorporates temporal context of a speaker based on the assumption that his emotional state is slowly varying, while dialog modeling typically incorporates temporal context from both speakers based on the assumption that their emotional states influence each other. In both cases, context is modeled by a higher-level HMM, where each state corresponds to a different emotion. This higher-level HMM can be either first or second order, where each state is dependent upon the previous one or two states respectively. The transition probabilities between states are then approximated by bigrams or trigrams, estimated in the training set. So, in the case of a first-order higher-level HMM for example, $P(emo_2|emo_1)$ would be estimated as the ratio of the training emo_2 utterances that follow emo_1 utterances, divided by the total number of emo_1 utterances. The emission probability distribution per state is the joint score conditional distribution for the corresponding emotion, $P([s_{it}]_{i=1}^{N}|emo_i)$. In the case of dialog context modeling the second-order HMM allows us to relate the current state with the previous emotions of not only the current speaker but also the interlocutor. This happens because in most cases an emotional utterance of a speaker is followed by an emotional utterance from his interlocutor.

Multimodal Fusion: There is flexibility regarding the choice of a fusion strategy that can be used for the multimodal cues. Modalities could be fused at the feature level, e.g., by training audio-visual lower level HMMs or at the model-level by collecting the log-likelihood scores for each modality, e.g., $[P(U_t|\lambda_i)^f, P(U_t|\lambda_i)^v]_{i=1}^{N}$ to model multimodal score vector discribution (f denotes facial and v vocal modality). Alternatively, they could be fused at the score-level by adding the score log-likelihoods with appropriate weights, e.g., $w_f \cdot \log P([s_{it}^f]_{i=1}^{N}|emo_i) + w_v \cdot \log P([s_{it}^v]_{i=1}^{N}|emo_i)$.

4. FEATURE EXTRACTION

Facial feature extraction is based on the normalized (x,y,z) coordinates from 46 facial markers [11, 8]. In order to obtain a low-dimensional representation of the facial marker information, we use Principal Feature Analysis [12]. This method performs Principal Component Analysis as a first step and selects features (here marker coordinates) so as to minimize the correlations between them. We select 30 features (covering approximately 95% of the total variability) and append the first derivatives. For details, please see [11].

The head features consist of the head translation in the (x,y,z) directions as well as the head angles (yaw, pitch and roll). Translations are derived from the nose marker and head angles are computed from all the markers using a technique based on Singular Value Decomposition [8]. The hand features consist of the average (x,y,z) coordinates of three markers on each hand. For the head and hand features we include first derivatives in our feature vector. Speech features are z-normalized 12 MFCC coefficients, pitch, and energy, together with their first derivatives, extracted using Praat. Audio and visual features are extracted at 25 Hz, with a 50 ms window.

5. RESULTS AND DISCUSSION

Our experiments are organized in a five-fold leave-one-pair-out cross validation. The presented recognition results (unweighted F1 measures) are the pair-independent averages over the five folds. The number of test utterances per fold is 1954 ± 194 on average when we have audio cues and 988 ± 97 when we have audio-visual cues.

5.1. Unimodal Classification and Joint Score Modeling

In this section we present the HMM unimodal recognition results for discriminating three levels of valence and activation, for facial, vocal, head and hand movement modalities. For the valence task, we have discriminatively re-trained the emotion-specific HMMs, using the Minimum Phone Error criterion [10], which led to an improvement in the average F1 measure around 1-2% absolute. Due to class imbalance for activation, where most of the instances are of medium activation, discriminative training did not improve performance and has not been used. We also present the classification results when we perform joint score distribution modeling on top of the lower-level HMM classification. We examine two such modeling approaches: through a GMM and through logistic regression, either nominal for the valence task, or ordinal for the activation task. The GMMs or regression models have been trained using the likelihood scores in the trainset, to which we added low random gaussian noise to improve generalization. The results are presented in Table 1.

We notice that facial cues seem to be more informative for valence, while vocal cues more informative for activation. The head and hand movement cues generally carry less emotional information, although they seem informative for activation level discrimination. Recognition performance based on voice tends to be higher when the speaker is wearing the markers. Analysis of the microphone signals suggests that this is possibly an artifact of the database due to the placement of the microphones. Joint score distribution modeling generally gives an improvement over low-level HMMs; the GMM and logistic regression models perform comparably, with ordinal regression being the best performing for the ordinal activation task, but nominal regression performing lower than the GMM for the valence task. For the rest of our experiments, due to lack of space, we only present the results of the GMM method which gives a consistent improvement for both classification tasks.

Table 1. *Classification performances (F1 measures) for three levels of valence and activation using face (f), voice (v), head (h) and hand (ha) features: mean and standard deviation of F1-measure across the 5 folds (5 dyadic sessions).*

Lower-level HMM Classification: F1 (mean & std.dev)		
classifier	valence	activation
HMM (v)	54.03 ± 3.09	55.39 ± 1.38
with markers	54.91 ± 2.82	57.53 ± 2.10
without markers	52.45 ± 4.05	52.75 ± 2.11
HMM (f)	60.26 ± 3.71	47.71 ± 4.49
HMM (h)	43.53 ± 2.59	51.51 ± 2.05
HMM (ha)	42.34 ± 3.14	44.69 ± 1.01
HMM and Score Vector modeling using GMM		
classifier	valence	activation
HMM+GMM (v)	54.95 ± 1.86	56.61 ± 2.88
with markers	56.46 ± 2.38	58.88 ± 4.26
without markers	52.99 ± 2.41	54.09 ± 1.91
HMM+GMM (f)	59.62 ± 3.71	48.57 ± 4.44
HMM+GMM (h)	45.48 ± 1.75	52.74 ± 1.83
HMM+GMM (ha)	43.09 ± 2.20	48.79 ± 3.31
HMM and Score Vector modeling using Multinomial Logistic Regression (MLR)		
classifier	valence(nominal)	activation(ordinal)
HMM+MLR (v)	54.52 ± 2.24	57.74 ± 1.11
with markers	55.35 ± 2.95	59.45 ± 2.57
without markers	53.17 ± 2.48	55.79 ± 1.13
HMM+MLR (f)	58.91 ± 2.57	49.47 ± 5.38
HMM+MLR (h)	44.41 ± 1.89	52.52 ± 1.32
HMM+MLR (ha)	43.57 ± 2.35	48.18 ± 2.93

5.2. Multimodal Fusion at Multiple Levels

In Table 2, we present the multimodal fusion results using face and voice (fv), or face,voice and head (fvh) modalities (corresponding only to the cases where the speakers are wearing markers). We examine fusion at multiple levels; at the feature-level we train multistream HMMs where the steam weights for audio and facial modalities are optimized on the train set. At the model level, we train GMMs to model multimodal score vectors, and at the score level we perform a weighted average of the score-vector GMM log-likelihoods, where the weights for each modality have been optimized on the train set. The results using the hand modality are omitted since they did not show a significant performance increase.

Based on the intuition that face and voice cues are more correlated than head cues, we fuse the head movement modality at the same or later stage than the other two. For the valence task, where both facial and vocal cues have adequate discriminative power, fusing them at the feature-level leads to good performance, and including head cues at the score level gives a further small performance increase. For the activation task, where the vocal cues alone perform considerably better that the facial cues, it is preferable to combine the three modalities at the score level, after adjusting the modality weights on the train set. Finally, as expected, multimodal classifiers perform considerably better than unimodal ones.

5.3. Speaker and Dialog Modeling

Here, we present the results after context modeling. We examine two issues; firstly whether emotional context information from a speaker could benefit classification of his current emotion (speaker modeling), and secondly whether including context from his interlocutor further increases performance (dialog modeling). For speaker modeling we use a first-order HMM denoted as HMM_{sp}^{1st} while for dialog modeling we use a second-order HMM denoted as HMM_{d}^{2nd}.

Table 2. *Classification performances (F1 measures) for three levels of valence and activation by fusing using face (f), voice (v) and head (h) modalities at various levels: mean and standard deviation of F1-measure across the 5 folds.*

Fusion of face and voice: F1 (mean & std.dev)		
classifier & fusion approach	valence	activation
HMM(fv)+GMM feature	62.75 ± 4.43	57.43 ± 3.76
HMM+fuse(fv)+GMM(f) model	61.08 ± 4.40	57.94 ± 3.71
HMM+GMM+fuse(fv) score	62.22 ± 2.73	59.27 ± 3.92
Fusion of face, voice and head: F1 (mean & std.dev)		
classifier & fusion approach	valence	activation
HMM(fv)+fuse(h)+GMM fv:feature, h:model	61.92 ± 4.11	57.16 ± 3.79
HMM(fv)+GMM+fuse(h) fv:feature, h:score	63.26 ± 4.05	58.83 ± 3.66
HMM+fuse(fvh)+GMM fvh:model	60.47 ± 4.37	58.64 ± 2.74
HMM+fuse(fv)+GMM+fuse(h) fv:model, h:score	61.01 ± 3.80	59.30 ± 2.92
HMM+GMM+fuse(fvh) fvh:score	61.84 ± 3.26	61.15 ± 2.95

Higher order modeling for speaker context did not provide any additional benefits. In all cases, transition probabilities were estimated on the training set. The results for the valence classification task are presented in Table 3. We do not present results for activation, where using temporal emotional context does not significantly increase performance, as we found in our previous work [6].

Here, we make a distinction between emotion classification of the speaker wearing the markers, where audio-visual cues are available, and the speaker without markers where only audio cues are available. The temporal information that is beneficial for each case may vary as can be seen in Table 3. The difference in available modalities for each speaker makes emotion classification more reliable for the speaker wearing the markers, compared to the speaker without markers. Therefore, although temporal emotional context from the same speaker generally seems beneficial, context from the other speaker through dialog modeling is beneficial only if the other speaker has more available modalities, and therefore more reliable emotional estimates. This motivated us to try mixed modeling, that is speaker modeling for the speaker wearing markers and dialog modeling for the speaker without markers, which resulted in the highest classification performance, as can be seen in Table 3.

6. CONCLUSION AND FUTURE WORK

We have proposed and tested a hierarchical, multimodal framework that captures emotional dynamics within and between emotions in affective dialogs and incorporates multiple modalities at various levels, depending upon the classification task and the modality. According to our results multimodal classifiers outperform unimodal ones, especially for the case of activation where facial, vocal and head movement cues carry relevant emotional information. Valence classification benefits significantly from incorporating temporal context from the same speaker. Considering context from the other speaker it is only helpful when we have a reliable (multimodal) emotional estimate of that speaker. Our framework is flexible enough to handle varying characteristics of each emotional task or dialogs where a different amount of multimodal information is available per speaker. We have examined HMM, GMM and MLR classifiers for utterance and score-vector modeling, however other generative or discriminative approaches could be applied according to the problem in hand.

Our future goals include extending this framework to perform recognition instead of classification, where we do not assume that test dialogs are presegmented, and performing video processing to obtain visual information for the speaker without markers. Further, we would like to include lexical information or information about the subject of the dialog, in order to incorporate higher-level dialog understanding in our emotion recognition system.

Table 3. *Classification performances (F1 measures) for three levels of valence for speaker and dialog modeling. We use the best multimodal fusion approach from the previous section.*

valence			
No higher level modeling			
speaker	classifier	modeling	F1 (mean & std.dev)
with markers	HMM(fv)+GMM+fuse(h)	–	63.26 ± 4.05
no markers	HMM+GMM (v)	–	52.75 ± 2.11
	in total		58.92 ± 2.65
Speaker modeling, 1st order HMM for each speaker separately			
speaker	classifier	modeling	F1 (mean & std.dev)
with markers	HMM(fv)+ HMM,fuse(h)	HMM_{sp}^{1st}	66.09 ± 3.39
no markers	HMM+HMM (v)	HMM_{sp}^{1st}	54.88 ± 3.52
	in total		61.30 ± 2.93
Dialog modeling, 2nd order HMM for total dialog			
speaker	classifier	modeling	F1 (mean & std.dev)
with markers	HMM(fv)+ HMM,fuse(h)	HMM_{d}^{2nd}	62.01 ± 3.14
no markers	HMM+HMM (v)	HMM_{d}^{2nd}	57.14 ± 2.80
	in total		59.98 ± 2.61
Mixed modeling, according to whether the speaker wears markers or not			
speaker	features	modeling	F1 (mean & std.dev)
with markers	HMM(fv)+HMM,fuse(h)	HMM_{sp}^{1st}	66.09 ± 3.39
no markers	HMM+HMM (v)	HMM_{d}^{2nd}	57.14 ± 2.80
	in total		62.31 ± 2.18

7. REFERENCES

[1] Cristina Conati, "Probabilistic assessment of user's emotions in educational games," *Applied Artif. Intel.*, vol. 16, pp. 555–575, 2002.

[2] V. Rozgic, B. Xiao, A. Katsamanis, B. Baucom, P. Georgiou, and S. Narayanan, "Estimation of ordinal approach-avoidance labels in dyadic interactions: ordinal logistic regression approach," in *ICASSP*, 2011.

[3] A. Mehrabian, "Communication without words," *Psychology today*, vol. 2, pp. 53–56, 1968.

[4] T.S. Huang N. Sebe, I. Cohen, *Multimodal Emotion Recognition*, Handbook of Pat. Rec. and Comp. Vision, World Scientific, 2005.

[5] G. Castellano, L. Kessous, and G. Caridakis, "Multimodal emotion recognition from expressive faces, body gestures and speech.," in *Doctoral Consortium of ACII*, 2007.

[6] M. Wöllmer, A. Metallinou, F. Eyben, B. Schuller, and S. Narayanan, "Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional lstm modeling," in *Proc. of Interspeech*, 2010.

[7] C.-C. Lee, C. Busso, S. Lee, and S. Narayanan, "Modeling mutual influence of interlocutor emotion states in dyadic spoken interactions," in *Proc. of Interspeech*, 2009.

[8] C. Busso, M. Bulut, C-C Lee, A.Kazemzadeh, E. Mower, S. Kim, J. Chang, S.Lee, and S.Narayanan, "IEMOCAP: interactive emotional dyadic motion capture database," *Lang. Res. and Eval.*, vol. 42, pp. 335–359, 2008.

[9] I. Cohen, A. Garg, and T. S. Huang, "Emotion recognition from facial expressions using multilevel hmm," *NIPS*, 2000.

[10] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book*, Entropic Cambridge Research Laboratory, Cambridge, England, 2006.

[11] A. Metallinou, C.Busso, S. Lee, and S. Narayanan, "Visual emotion recognition using compact facial representations and viseme information," in *Proc. of ICASSP*, 2010.

[12] Y. Lu, I. Cohen, X.S. Zhou, and Q. Tian, "Feature selection using principal feature analysis," in *Proc. of the 15th Intl. Conf. on Multimedia, Germany*, 2007.