Context-Sensitive Learning for Enhanced Audiovisual Emotion Classification (Extended Abstract)

Angeliki Metallinou
Amazon.com, Inc.

Martin Wöllmer, Florian Eyben, and Björn Schuller
audEERING UG
Landsberger Strasse 46d, 82205 Gilching, Germany

Athanasios Katsamanis
School of E.C.E., National Technical University of Athens,
Zografou campus, Athens 15773, Greece

Shrikanth Narayanan
Department of Electrical Engineering,
University of Southern California,
Los Angeles, CA, 90089 USA

***Abstract*—Human emotional expression tends to evolve in a structured manner in the sense that certain emotional evolution patterns, i.e., anger to anger, are more probable than others, e.g., anger to happiness. Furthermore the perception of an emotional display can be affected by recent emotional displays. Therefore, the emotional content of past and future observations could offer relevant temporal context when classifying the emotional content of an observation. In this work, we focus on audio-visual recognition of the emotional content of improvised emotional interactions at the utterance level. We examine context-sensitive schemes for emotion recognition within a multimodal, hierarchical approach: bidirectional Long Short-Term Memory (BLSTM) neural networks, hierarchical Hidden Markov Model classifiers (HMMs) and hybrid HMM/BLSTM classifiers are considered for modeling emotion evolution within an utterance and between utterances over the course of a dialog. Overall, our experimental results indicate that incorporating long-term temporal context is beneficial for emotion recognition systems that encounter a variety of emotional manifestations.**

I. Introduction

Human emotional expression is a complex process where a variety of multimodal cues interact to create an emotional display. Furthermore, emotions are usually slowly varying during a conversation and the perception of an emotional display is affected, among others, by recently perceived past emotional displays, which place the expressed emotion into context. Taking into account such contextual information may prove to be advantageous for real-life automatic emotion recognition systems that can process a great variety of complex, vague or ambiguous emotional displays [1], [2], [3], [4], [5], [6], [7]. The focus of this paper is to investigate learning frameworks for automatic, multimodal emotion recognition that allow the use of information of the structure of past and future evolution of an emotional interaction. The study also considers the flow of emotional expression by examining emotional transitions in a variety of improvised affective interactions.

Following our previous work [8], we define context to be information about the emotional content of audio-visual displays that happen before or after the observation that we examine. We focus on emotion recognition at the utterance level. Here an utterance is loosely defined as a chunk of speech where the speaker utters a thought or idea. The phrases that we examine have been manually segmented from longer dyadic conversations and usually last a few seconds. In addition to the current utterance's audio-visual cues, we exploit information from an arbitrary number of neighboring utterances that could range from one past or future utterance to all the utterances of the conversation. Apart from this definition of context which is our primary focus, we could also interpret the use of audio-visual cues as another form of context where the interplay within the multimodal streams provides context for one another and offers a fuller picture of the expressed emotion.

We investigate three alternative multimodal and hierarchical schemes for incorporating contextual information in emotion recognition systems, by modeling emotional evolution at two levels: within an emotional utterance and between emotional utterances of a conversation. Specifically, we examine the use of hierarchical Hidden Markov Model (HMM) classifiers [9], of Recurrent Neural Networks (RNNs), and specifically BLSTM neural networks [10], [11], [12], [13], [14], [15], [16] as well as the use of a hybrid BLSTM/HMM approach. The HMM-based classification is inspired by the Automatic Speech Recognition (ASR) literature, where algorithms exploit context at multiple levels within a Markov model structure: from phonetic details including coarticulation in speech production to word transitions reflecting language based statistics [17], [18]. We hypothesize that similar within and across model transitions can be advantageously used to capture the dynamics in the evolution of emotional states, including within and across emotional categories. Alternatively, RNN architectures are a powerful, discriminative framework that enables modeling the emotional flow of a conversation

without making Markov assuptions about emotional transitions. Here, we apply BLSTM neural networks which overcome the vanishing gradient problem of conventional RNNs and are able to learn from an arbitrarily large amount of past and future contextual information [19].

For our experiments we use a large multimodal and multisubject database of dyadic interactions between actors, namely the IEMOCAP database [20], which contains detailed facial information, obtained from facial Motion Capture (MoCap) as well as speech information. The IEMOCAP database consists of dyadic conversations that are elicited so as to contain emotional manifestations that are non-prototypical and resemble real-life emotional expression. Our goal is to obtain a realistic assessment of emotion recognition performance, when our system is required to make a decision about the emotional content of all possible input utterances, including those containing subtle or ambiguous emotions.

We focus on the recognition of dimensional emotional descriptions, i.e., valence and activation levels, instead of categorical emotional tags, such as 'anger' or 'happiness'. Valence describes how positive vs. negative and activation how calm vs. excited is the expressed emotion. We derive a dimensional label for all available utterances by averaging the decisions of multiple annotators. In addition to classifying the degree of valence and activation separately, we also investigate their joint modeling by classifying among clusters in the two-dimensional valence-activation space [21].

II. Context-Sensitive Frameworks

Our problem can be posed as a two-level modeling problem of an emotional conversation. At the higher level, an emotional conversation is modeled as a sequence of emotional utterances, while at the lower level, each such utterance is modeled as a sequence of audiovisual observations. We assume that an emotional utterance can be described by a single emotional label, e.g., a single level of activation, valence or a single cluster in the valence-activation space. However, an emotional conversation may contain arbitrary emotional transitions between utterances and may consist of a variety of emotional manifestations. Therefore utterance modeling captures the dynamics within emotional categories while conversation modeling captures the dynamics across emotional categories.

At the utterance level, we examine dynamic modeling by using fully-connected HMMs, which captures feature statistics and underlying emotional characteristics in the audio-visual feature streams. The intuition for using fully-connected HMMs is that there is no apparent left-to-right property in the dynamic evolution of the facial or vocal characteristics during emotional expression (as opposed to the evolution of phonemes during speech that is exploited in phoneme-specific left-to-right HMMs in ASR). The use of coupled- instead of simple, multistream HMMs enables us to model asynchrony between the audio-visual streams. Alternatively, we model the emotional utterance by estimating static, utterance-level, statistical features, through the use of statistical functionals over the low-level frame sequence. Such an approach implicitly captures some of the observation dynamics while it makes fewer modeling assumptions compared to the HMM (no Markovian property, conditional independence or synchronicity assumptions of the underlying audio-visual sequences). At the dialog level, we examine the use of HMM and discriminatively trained neural network classifiers (RNN, BLSTM). The latter make fewer assumptions on the underlying sequence of emotional utterances and may potentially capture more complex patterns of emotional flow.

Our approach of combining first and second layer HMMs for dialog and utterance modeling leads to a two-level structure, along the lines of multi-level HMMs [22] and Hierarchical HMMs [9]. Alternatively, we examine the performance of discriminatively trained neural network classifiers for conversational modeling when statistical functionals are extracted at the utterance level. In the hybrid HMM/BLSTM approach, we keep the probabilistic dynamic modeling of HMMs at the utterance level and use the emotion-specific HMM log-likelihoods to form an utterance-level feature vector, which is the input of the BLSTM at the conversation-level (for more details, see [23]).

We examine a combination of the HMM and BLSTM classifiers that takes advantage of both the explicit dynamic utterance modeling of the HMM framework and the ability of the BLSTM to learn an arbitrarily long amount of bidirectional context. We utilize the BLSTM network as a second layer of computation over the HMM classifiers as an alternative to Viterbi decoding. This combination has the advantages of a two-layer classification structure; therefore there is transparency as to the performance improvement we can gain from context modeling. Furthermore, the HMM+BLSTM combination may potentially capture more complex structure in the underlying emotional flow than the one captured by an HMM.

III. Database Description

The database used in this work is the Interactive Emotional Dyadic Motion Capture (IEMOCAP) database which contains approximately 12 hours of audio-visual data from five mixed gender pairs of actors [20]. Each recorded session lasts approximately 5 minutes and consists of two actors interacting with each other in scenarios that encourage emotional expression. During each recording, both actors wore microphones and one of them had face Motion Capture (MoCap) markers. In this study, we examine the emotions expressed when the actors wear the markers, so that there is audio-visual information available.

The dyadic sessions were later manually segmented into utterances, where consecutive utterances of a speaker may or may not belong to the same turn. We examine the sequence of utterances of a certain speaker during a

recording, and we make no distinction between utterances that are separated by one or more utterances of the other speaker and utterances that belong to the same turn. The emotional content of each utterance was annotated by human annotators in categorical labels (annotators had to choose between the following emotions: angry, happy, excited, sad, frustrated, fearful, surprised, disgusted, neutral and "other-please specify") and in dimensional descriptions of valence and activation. Valence describes how positive vs. negative and activation how calm vs excited is the expressed emotion. Value 1 denotes very low activation and very negative valence and 5 denotes very high activation and very positive valence. Those properties are rated on scales 1-5 and are averaged across 2 annotators (or for some few utterances across 3 annotators).

IV. Emotions and Emotion Transitions

The first emotion classification task that we consider in this work consists of the classification of three levels of valence and activation: level 1 contains ratings in the range [1,2], level 2 contains ratings in the range (2,4) and level 3 contains ratings in the range [4,5]. These levels intuitively correspond to low, medium and high activation, and to negative, neutral and positive valence respectively.

We also examine the joint classification of the emotional dimensions by building three and four clusters in the valence-activation emotional space. The motivation for clustering the valence-activation space is to build classifiers that provide richer and more complete emotional information by combining valence and activation information. We apply data-driven clustering through K-means to automatically select clusters that fit the distribution of the emotional manifestations of our database in the emotional space (similar approaches are also followed in [21],[24]). The ground truth of every utterance is assigned to one of the clusters using the minimum Euclidean distance between its annotation and the cluster midpoints.

When abstracting our emotion classes into clusters of the valence-activation space, we also study their relation to the corresponding categorical annotations and investigate which categorical emotional manifestations tend to fall into each cluster. Specifically, we examine how each cluster breaks down in terms of categorical labels. For example, in Fig. 1(a) and 1(b) we illustate the 3 clusters in the emotional space and the histogram of the categorical emotional tags of the utterances they contain. Note that the utterances that belong to each cluster depend on the training set of each fold. Thus, the bar graphs in Fig. 1(b) represent the mean over the 10 folds of our experiment (see also the experimental setup in section VI-A) and the error bars represent the standard deviation over the 10 folds. The plot of Fig. 1(a) corresponds to the first fold of our experiment, but the differences across folds are relatively small (the average standard deviation of the cluster centroid coordinates across the ten folds is as low as 0.05).

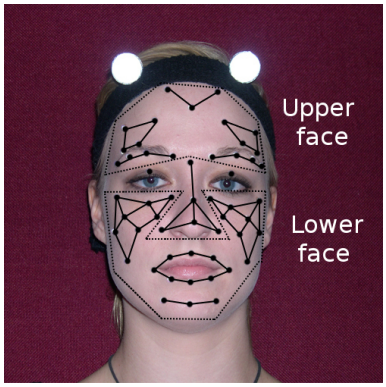


Fig. 2. *Positions of the MOCAP face markers and separation of the face into lower and upper facial regions.*

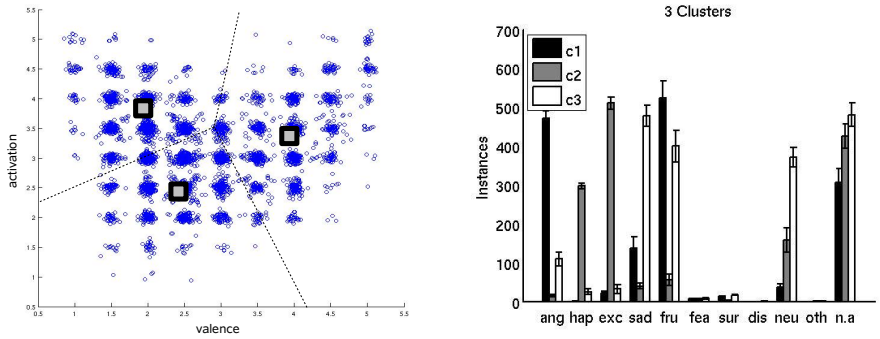
V. Feature Extraction and Fusion

A. Audio-Visual Frame-level Feature Extraction

The IEMOCAP data contain detailed MoCap facial marker coordinates. The positions of the facial markers can be seen in figure 2. The markers were normalized for head rotation and translation and the nose marker tip is defined as the local coordinate center of each frame. In total, information from 46 facial markers is used; namely their (x,y,z) coordinates. This results in a 138-dimensional facial representation, which tends to be redundant because it does not exploit the correlations of neighboring marker movements and the structure of the human face.

In order to obtain a lower-dimensional representation of the facial marker information, we use Principal Feature Analysis (PFA) [25]. This method performs Principal Component Analysis (PCA) as a first step and selects features so as to minimize the correlations between them. In contrast to PCA, PFA selects features in the original feature space (here marker coordinates) instead of linear combinations of features, which makes selection results interpretable. We select 30 features, since the PCA transformation explains more than 95% of the total variability. To these we append the first derivatives, resulting in a 60-dimensional representation. The facial features are normalized per speaker to smooth out individual facial characteristics that are unrelated to emotional expressions. Our speaker normalization approach consists of finding a mapping from the individual average face to the general average face. This is achieved by shifting the mean value of each marker coordinate of each subject to the mean value of that marker coordinate across all subjects. The feature selection and normalization framework is described in our previous work [26].

In addition, we extract a variety of features from the speech waveform: 12 MFCC coefficients, 27 Mel Frequency Band coefficients (MFB), pitch and energy values. We also compute their first derivatives. All the audio features are computed using the Praat software [27] and are normalized using z-standardization. The audio and visual features are extracted at the same framerate of 25 ms, with



(a) 3 clusters in the valence-activation space (fold1)

(b) 3 clusters

Fig. 1. *Analysis of classes in the 3 cluster task in terms of categorical labels. The bars and the error bars correspond to the mean and standard deviation computed across the 10 folds. We notice that the data-driven clusters tend to contain different categorical emotional manifestations according to their position in the emotional space. Specifically, clusters 1,2 and 3 roughly contain categorical emotions of 'anger or frustration', 'happiness or excitement' and 'neutrality or sadness or frustration' respectively.*

a 50 ms window. The utterance-level audio HMMs were trained using the 27 MFBs, pitch and energy along with their first derivatives, while the visual HMMs were trained using the 30 PFA features with their first derivatives. For the audio-visual HMMs and coupled-HMMs we used both these voice and face features, fused at feature or at model-level.

B. Utterance-level Statistics of Audio-Visual Features

We use a set of 23 utterance-level statistical functionals that are computed from the low-level acoustic and visual features . Thus, we obtain $142 \times 23 = 3266$ utterance-level features. All functionals were calculated using the openSMILE toolkit [28].

In order to reduce the size of the resulting feature space, we conduct a cyclic Correlation based Feature Subset Selection (CFS) using the training set of each fold. The main idea of CFS is that useful feature subsets should contain features that are highly correlated with the target class while being uncorrelated with each other [29], [30].

C. Audio-Visual Feature Fusion

For the utterance-level HMM approaches where frame-level features are used, we apply multi-stream HMM classifiers (here denoted simply as HMMs). These assign different importance weights to the audio and visual modalities and assume synchronicity between them. When modeling the dynamics of high level attributes such as emotional descriptors of a whole utterance, allowing asynchrony in the dynamic evolution of the underlying audio-visual cues could be beneficial. We also apply model-level fusion through the use of coupled Hidden Markov Models (c-HMMs), which allow this type of asynchrony, and have been widely used in the literature [31], [32]. All models are trained using the HTK Toolkit [18]. HTK offers the functionality for defining and training a multi-stream HMM but it does not explicitly allow for coupling of multiple single stream HMMs. However, following the analysis presented in [32], [33], we can implement c-HMMs in HTK using a product HMM structure.

VI. Experiments and Results

A. Experimental Setup

Our experiments are organized in a cyclic leave-one-speaker-out cross validation. The feature extraction PCA transformations (for the face features) and the feature z-normalization constants are computed based on the respective training set of each fold. The mean and standard deviation of the number of test and training utterances across the folds is 498 ± 60 and 4475 ± 61, respectively. For each fold, we compute the F1-measure, which is the harmonic mean of unweighted precision and recall, as our primary performance measure. As a secondary measure we also report unweighted recall (unweighted accuracy). The presented recognition results are the subject-independent averages over the ten folds and the corresponding standard deviation.

In the next sections we present the results of the context-sensitive neural network and HMM frameworks for the various classification tasks. We trained 3-state ergodic HMMs and c-HMMs with observation probability distributions modeled by Gaussian mixture models. The stream weights and the number of mixtures per state (varying from 4 to 32 mixtures) have been experimentally optimized on a validation set. For the context-sensitive HMM approaches, the bigrams and the BLSTMs of each fold have been computed or trained on the corresponding training set. The LSTM networks consist of 128 memory blocks with one memory cell per block while the BLSTM networks consist of two LSTM layers with 128 memory blocks per input direction. To improve generalization, we add

TABLE I

Comparing context-free and context-sensitive classifiers for discriminating three levels of valence and activation, and three and four clusters in the valence-activation space, using face (f) and voice (v) features: mean and standard deviation of F1-measure and unweighted Accuracy across the 10 folds (10 speakers).

classifier	features	F1	Acc.(uw)
valence			
HMM	v	49.85 ± 3.18	49.99 ± 3.63
HMM	f	58.85 ± 3.86	60.98 ± 4.96
HMM	v+f	60.79 ± 2.53	62.50 ± 3.39
cHMM	v+f	60.42 ± 3.59	61.75 ± 4.66
HMM+HMM(w=2)	v+f	62.02 ± 2.25	63.16 ± 3.18
HMM+BLSTM	v+f	63.97 ± 3.03	62.78 ± 6.43
BLSTM	v+f	65.12 ± 5.13	64.67 ± 6.48
activation			
HMM	v	57.54 ± 3.33	61.92 ± 4.88
HMM	f	49.04 ± 4.40	51.36 ± 4.14
HMM	v+f	57.56 ± 4.27	60.00 ± 4.45
cHMM	v+f	57.39 ± 3.25	61.29 ± 5.16
HMM+HMM(w=4)	v+f	57.71 ± 4.23	60.02 ± 4.54
HMM+BLSTM	v+f	53.41 ± 5.99	46.93 ± 5.69
BLSTM	v+f	54.90 ± 5.02	52.28 ± 5.37
3 clusters			
HMM	v+f	67.33 ± 5.15	66.18 ± 6.69
cHMM	v+f	68.45 ± 3.38	67.95 ± 3.18
cHMM+HMM(w=4)	v+f	70.36 ± 3.48	69.76 ± 3.09
cHMM+BLSTM	v+f	68.09 ± 4.16	68.02 ± 4.72
BLSTM	v+f	72.35 ± 5.10	71.83 ± 5.46
4 clusters			
HMM	v+f	56.54 ± 4.29	56.64 ± 5.90
cHMM	v+f	57.28 ± 3.65	57.87 ± 4.33
cHMM+HMM(w=4)	v+f	58.65 ± 3.80	58.89 ± 4.59
cHMM+BLSTM	v+f	58.21 ± 5.24	57.94 ± 5.89
BLSTM	v+f	62.80 ± 6.69	61.96 ± 7.02

zero mean Gaussian noise with standard deviation 0.6 to the input statistical features during training. The BLSTM networks that are trained to process the HMM outputs (HMM+BLSTM) consist of 32 memory blocks per input direction.

B. Context-Free vs Context-Sensitive Classifiers

In this section, we compare the classification performance between emotion-specific HMMs or c-HMMs, that do not make use of context information, and our proposed context-sensitive frameworks: hierarchical HMMs (or c-HMMs) using sequential Viterbi Decoding with bidirectional window of $w + 1$ utterances (HMM+HMM, c-HMM+HMM), BLSTM trained with utterance-level feature functionals (BLSTM) and hybrid HMM/BLSTM classifiers (HMM+BLSTM, c-HMM+BLSTM). Table I shows the performances for discriminating three levels of valence and activation, as well as classification into three and four clusters in the valence-activation space. To test the statistical significance of the differences in average F1 performance between (c-)HMM, (c-)HMM+HMM, (c-)HMM+BLSTM and BLSTM classifiers, we conducted repeated measures ANOVA at the subject level with Bonferonni adjustment for post-hoc tests, using SPSS [34].

Concerning the valence results, we observe that facial features are much more effective in classifiying valence than voice features, which agrees with our previous findings [8]. Regarding the audio-visual classifiers, BLSTM achieves the highest average F1 measure, while the emotion-specific HMMs benefit from the use of long-range context, either through higher-level Viterbi Decoding (HMM+HMM) or through the use of a higher-level BLSTM (HMM+BLSTM). Statistical significance tests reveal that the average HMM+BLSTM F1 measure is significantly higher than that of the HMM at the 0.05 level. HMM+HMM performance was not found significantly higher than HMM performance, but it has a p value very close to the threshold (p=0.055). Similarly the comparison of BLSTM and HMM gives a p value of 0.06.

For the activation task, incorporating visual cues does not improve performance significantly, indicating that audio cues are more informative that visual cues. This agrees with previous results in the emotion literature [35], [36]. Overall, we notice that taking temporal context into account does not benefit activation classification performance for HMMs (HMM+HMM), and the BLSTM and HMM/BLSTM classifiers on average perform worse than the context-free HMMs. This could be attributed to the isolated nature of the extreme activation instances.

For the three cluster task, we notice that context-sensitive classifiers, such as c-HMM+HMM and BLSTM, on average perform higher than the simple HMMs and c-HMMs, and that the BLSTM classifier achieves the highest average F1 measure. The average F1 measure of cHMM+HMM was found significantly higher that of c-HMM at the 0.05 level. Similarly, for the four cluster task context-sensitive classifiers tend to outperform simple HMMs and cHMMs in terms of average F1, although these differences are not statistically significant at the 0.05 level.

Overall, our results suggest that incorporating context is beneficial for the valence and the three and four cluster classification. The BLSTM classifier generally achieves the highest classification performance, although performance across folds has a relatively high variance. The hierarchical HMM and hybrid HMM/BLSTM classifiers perform similarly in general, and lower than the BLSTM in terms of average F1 measure, although they tend to have more consistent performance across subjects (smaller variance). Regarding the hierarchical HMM approach we notice that a small amount of bidirectional context (e.g., w=4) can give a performance increase. We have omitted the results of the HMM+HMM architecture where Viterbi Decoding is used over the total observation sequence. For all our classification tasks, the results are very similar to the ones obtained through sequential VD with small window sizes, which suggests that it is possible to increase recognition performance even when a small amount of bidirectional context is used. These observations are encouraging and suggest that this algorithm could be applied in practical scenarios where an emotion recognition system might not afford to wait for the conversation to end in order to perform recognition, while it might be acceptable to wait a few utterances before making a decision.

Note that there are significant variations in the performance and rankings across different folds for all classification approaches, as indicated by the variances of Table I and the results of the statistical significance tests. These suggest that no approach is clearly superior for all speakers. Our insight is that these variations could result from speaker dependent characteristics of emotional expression, i.e., some speakers may be more overtly expressive than others or may make different expressive use of the audio and visual modalities.

C. Context-Sensitive Neural Network Classifiers

In this section, we compare the recognition performances of various Neural Network classifiers which take into account different amount of unidirectional and bidirectional context. The results are presented in Table II. (B)LSTM architectures achieve a higher average F1 measure compared to (B)RNN architectures which indicates the merit of learning a longer range of temporal context for emotion recognition tasks. Also, bidirectional neural networks, such as BLSTMs and BRNNs, outperform their respective unidirectional counterparts, such as LSTM and RNN, which suggests the importance of bidirectional context for these architectures. The performance differences between these context-sensitive NNs, although not statistically significant, indicate a consistent trend in performance across all classification tasks, with BLSTM being the highest performing classifier.

TABLE II
Comparing context-sensitive Neural Network classifiers for discriminating three levels of valence and activation, and three and four clusters in valence-activation space using face (f) and voice (v) features: mean and standard deviation of F1-measure and unweighted Accuracy across the 10 folds (10 speakers).

classifier	features	F1	Acc.(uw)
		valence	
RNN	v+f	63.34 ± 4.58	62.92 ± 6.00
BRNN	v+f	64.10 ± 5.05	63.68 ± 6.64
LSTM	v+f	63.71 ± 4.86	63.76 ± 5.95
BLSTM	v+f	65.12 ± 5.13	64.67 ± 6.48
		activation	
RNN	v+f	52.78 ± 5.21	48.54 ± 5.59
BRNN	v+f	53.93 ± 4.12	49.98 ± 4.62
LSTM	v+f	53.65 ± 4.97	50.35 ± 5.83
BLSTM	v+f	54.90 ± 5.02	52.28 ± 5.37
		3 clusters	
RNN	v+f	69.59 ± 5.75	69.34 ± 5.95
BRNN	v+f	69.94 ± 5.65	69.76 ± 6.00
LSTM	v+f	70.34 ± 5.85	69.53 ± 6.61
BLSTM	v+f	72.35 ± 5.10	71.83 ± 5.46
		4 clusters	
RNN	v+f	58.30 ± 6.63	57.29 ± 7.28
BRNN	v+f	60.10 ± 5.96	59.14 ± 6.72
LSTM	v+f	61.93 ± 5.96	61.02 ± 6.15
BLSTM	v+f	62.80 ± 6.69	61.96 ± 7.02

VII. Conclusion

In this work we have described and analyzed context-sensitive frameworks for emotion recognition, i.e., frameworks that take into account temporal emotional context when making a decision about the emotion of an utterance. These methods, which utilize powerful and popular classifiers, such as HMMs and BLSTMs, could be viewed under the common framework of a hierarchical, multimodal approach, which models the observation flow both at the utterance level (within an emotion) and at the conversation level (between emotions). The different classifiers that can be chosen for each level reflect different modeling assumptions on the underlying sequences and account for different system requirements. Our emotion classification experiments indicate that taking into account temporal context tends to improve emotion classification performance. Overall, context-sensitive approaches outperform methods that do not consider context for the recognition of valence states and emotional clusters in the valence-activation space, in terms of average F1 measure. However, the relatively large performance variability between subjects suggests that no method is clearly superior for all subjects. Additionally, the use of context from both past and future seems beneficial, as suggested by the slightly higher performance of bidirectional neural networks (BLSTM, BRNN) compared to their unidirectional counterparts. Even the use of a small amount of context around the current observation, e.g., from the use of the sequential VD algorithm with a small window of $w+1$ utterances, leads to performance improvement, which is an encouraging result for designing context sensitive frameworks with performance close to real-time. The only emotion classification task that does not benefit significantly from context is activation, possibly because of the isolated nature of the extreme activation events, which makes this structure difficult to model.

According to our results, neural network architectures, and specifically (B)LSTM networks trained with utterance level feature functionals achieve a higher average performance than HMM classification schemes. This could be attributed to their discriminative training, fewer modeling assumptions and their ability to capture long-range, bidirectional temporal patterns of the input feature streams and output activations. BLSTM networks can learn an adequate amount of relevant emotional context around the current observation, during the training stage. When such context is not present, for example when we randomly shuffle the utterances of a conversation, the performance of the BLSTM classifiers significantly decreases. However (B)LSTM and (B)RNN classifiers seem to have difficulties handling emotional expression variability between subjects, therefore their performance may vary significantly across people. HMM classification frameworks and hybrid HMM/BLSTM frameworks, on average perform lower than neural networks, but generally achieve more consistent classification results across subjects. They provide a structured approach for modeling and classifying sequences at multiple levels, they have more trasparency as to what amount of context is used and they are generally flexible.

REFERENCES

[1] R. El Kaliouby, P. Robinson, and S. Keates, "Temporal context and the recognition of emotion from facial expression," in *HCI International, Crete*, June 2003.

[2] J. M. Carroll and J. A. Russell, "Do facial expressions signal specific emotions? judging emotion from the face in context," *Journal of Personality and Social Psychology*, vol. 70, pp. 205–218, 1996.

[3] H. R. Knudsen and L. H. Muzekari, "The effects of verbal statements of context on facial expressions of emotion," *Journal of Nonverbal Behavior*, vol. 7, pp. 202–212, 1983.

[4] T. Masuda, P.C. Ellsworth, B. Mesquita, J. Leu, S. Tanida, and E. Van de Veerdonk, "Placing the face in context: Cultural differences in the perception of facial emotion," *Journal of Personality and Social Psychology*, vol. 94, pp. 365–381, 2008.

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

[6] B. de Gelder and J. Vroomen, "The perception of emotions by ear and by eye," *Cognition and Emotion*, pp. 289–311, May 2000.

[7] K. Oatley and J. M. Jenkins, *Understanding Emotions*, Blackwell Publishers Ltd, 1996.

[8] 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, Japan*, 2010.

[9] S. Fine, Y. Singer, and N. Tishby, "The Hierarchical Hidden Markov Model: Analysis and applications," *Machine Learning*, vol. 32, pp. 41–62, 1998.

[10] A. Graves, S. Fernandez, M. Liwicki, H. Bunke, and J. Schmidhuber, "Unconstrained online handwriting recognition with recurrent neural networks," *Advances in Neural Information Processing Systems*, vol. 20, pp. 1–8, 2008.

[11] A. McCabe and J. Trevathan, "Handwritten signature verifcation using complementary statistical models," *Journal Of Computers*, vol. 4, pp. 670–680, 2009.

[12] F. Weninger, J. Geiger, M. Wöllmer, B. Schuller, and G. Rigoll, "Feature Enhancement by Deep LSTM Networks for ASR in Reverberant Multisource Environments," *Computer Speech and Language*, vol. 28, no. 4, pp. 888–902, July 2014.

[13] E. Marchi, F. Vesperini, F. Weninger, F. Eyben, S. Squartini, and B. Schuller, "Non-Linear Prediction with LSTM Recurrent Neural Networks for Acoustic Novelty Detection," in *Proceedings 2015 International Joint Conference on Neural Networks (IJCNN)*, Killarney, Ireland, July 2015, IEEE, to appear.

[14] R. Brückner and B. Schuller, "Social Signal Classification Using Deep BLSTM Recurrent Neural Networks," in *Proceedings 39th IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2014*, Florence, Italy, May 2014, pp. 4856–4860, IEEE.

[15] J. Geiger, E. Marchi, F. Weninger, B. Schuller, and G. Rigoll, "The TUM system for the REVERB Challenge: Recognition of Reverberated Speech using Multi-Channel Correlation Shaping Dereverberation and BLSTM Recurrent Neural Networks," in *Proceedings REVERB Workshop, held in conjunction with ICASSP 2014 and HSCMA 2014*, Florence, Italy, May 2014, IEEE, pp. 1–8, IEEE.

[16] E. Coutinho, F. Weninger, K. Scherer, and B. Schuller, "The Munich LSTM-RNN Approach to the MediaEval 2014 "Emotion in Music" Task," in *Proceedings of the MediaEval 2014 Multimedia Benchmark Workshop*, Martha Larson, Bogdan Ionescu, Xavier Anguera, Maria Eskevich, Pavel Korshunov, Markus Schedl, Mohammad Soleymani, Georgios Petkos, Richard Sutcliffe, Jaeyoung Choi, and Gareth J.F. Jones, Eds., Barcelona, Spain, October 2014, CEUR.

[17] M.J.F. Gales and S.J. Young, "The application of Hidden Markov Models in speech recognition," *Foundations and Trends in Signal Processing*, vol. 1, pp. 195–304, 2008.

[18] 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.

[19] A. Graves, S. Fernandez, and J. Schmidhuber, "Bidirectional LSTM networks for improved phoneme classification and recognition," in *Proc. of ICANN*, Poland, 2005, vol. 18, pp. 602–610.

[20] 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," *Language Resources and Evaluation*, vol. 42, pp. 335–359, 2008.

[21] M. Wöllmer, F. Eyben, B. Schuller, E. Douglas-Cowie, and R. Cowie, "Data-driven clustering in emotional space for affect recognition using discriminatively trained LSTM networks," in *Proc. of Interspeech*, UK, 2009, pp. 1595–1598.

[22] I. Cohen, A. Garg, and T. S. Huang, "Emotion recognition from facial expressions using multilevel HMM," *Neural Information Processing Systems*, 2000.

[23] A. Metallinou, M. Wöllmer, A. Katsamanis, F. Eyben, B. Suller, and S. Narayanan, "Context-sensitive learning for enhanced audiovisual emotion classification," *IEEE Transactions on Affective Computing*, vol. 3, no. 2, pp. 184–198, 2012.

[24] 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, UK*, 2009.

[25] I. Cohen, Q. T. Xiang, S. Zhou, X. Sean, Z. Thomas, and T. S. Huang, "Feature selection using principal feature analysis," 2002.

[26] A. Metallinou, C.Busso, S. Lee, and S. Narayanan, "Visual emotion recognition using compact facial representations and viseme information," in *Proc. of ICASSP*, Dallas, Texas, 2010, pp. 2474–2477.

[27] P. Boersma, "Praat, a system for doing phonetics by computer.," *Glot International*, vol. 5, no. 9/10, pp. 341–345, 2001.

[28] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE - the Munich versatile and fast open-source audio feature extractor," in *Proc. of ACM Multimedia*, Firenze, Italy, 2010.

[29] M. A. Hall, *Correlation-based Feature Selection for Machine Learning*, Ph.D. thesis, University of Waikato, 1999.

[30] Ian H. Witten and Eibe Frank, *Data Mining: Practical machine learning tools and techniques*, Morgan Kaufmann, San Francisco, 2nd edition edition, 2005.

[31] A. V. Nefian, L. Liang, X. Pi, L. Xiaoxiang, C. Mao, and K. Murphy, "A coupled HMM for audio-visual speech recognition," in *Proc. of ICASSP*, 2002, pp. 2013–2016.

[32] M. Brand, N. Oliver, and A. Pentland, "Coupled Hidden Markov Models for complex action recognition," Proc. of CVPR, 1997.

[33] G. Gravier, G. Potamianos, and C. Neti, "Asynchrony modeling for audio-visual speech recognition," in *Proc. of the 2nd Intl. Conf. on Human Language Technology Research*, San Francisco, CA, USA, 2002, HLT '02, pp. 1–6, Morgan Kaufmann Publishers Inc.

[34] SPSS Inc., "SPSS base 10.0 for windows user's guide," *SPSS Inc., Chicago IL.*, 1999.

[35] J. A. Russell, J.-A. Bachorowski, and J.-M. Fernndez-Dols, "Facial and vocal expressions of emotion," *Annual Review of Psychology*, vol. 54, pp. 329–349, February 2003.

[36] A. Metallinou, S. Lee, and S. Narayanan, "Decision level combination of multiple modalities for recognition and analysis of emotional expression," in *Proc. of ICASSP*, Dallas, Texas, 2010, pp. 2462–2465.