Context-Sensitive Multimodal Emotion Recognition from Speech and Facial Expression using Bidirectional LSTM Modeling

Martin Wöllmer[1], *Angeliki Metallinou*[2], *Florian Eyben*[1], *Björn Schuller*[1], *Shrikanth Narayanan*[2]

[1]Institute for Human-Machine Communication, Technische Universität München, Germany
[2]Signal Analysis and Interpretation Lab (SAIL), University of Southern California, Los Angeles, CA

[woellmer,eyben,schuller]@tum.de, metallin@usc.edu, shri@sipi.usc.edu

10.21437/Interspeech.2010-646

Abstract

In this paper, we apply a context-sensitive technique for multimodal emotion recognition based on feature-level fusion of acoustic and visual cues. We use bidirectional Long Short-Term Memory (BLSTM) networks which, unlike most other emotion recognition approaches, exploit long-range contextual information for modeling the evolution of emotion within a conversation. We focus on recognizing dimensional emotional labels, which enables us to classify both prototypical and non-prototypical emotional expressions contained in a large audio-visual database. Subject-independent experiments on various classification tasks reveal that the BLSTM network approach generally prevails over standard classification techniques such as Hidden Markov Models or Support Vector Machines, and achieves F1-measures of the order of 72 %, 65 %, and 55 % for the discrimination of three clusters in emotional space and the distinction between three levels of valence and activation, respectively.

Index Terms: emotion recognition, multimodality, long short-term memory, hidden markov models, context modeling

1. Introduction

Due to the increasing interest in designing multimodal human-machine interfaces, information retrieval systems and conversational agents that take into account the affective state of the user, automatic emotion recognition (AER) from voice and face has become a core discipline in machine learning and pattern recognition [1]. Humans express and perceive emotion through the complex interplay of multiple modalities [2]. Thus, considering multiple modalities may give a more complete description of the expressed emotion and generally tends to lead to more accurate results than unimodal techniques [3]. Audio-visual modalities are already supported by today's laptops or by other widely used portable devices and are thereby of special interest for AER applications.

As emotion recognition presumes the modeling of the dynamics of acoustic or visual features, some classification strategies in the field of AER make use of dynamic classifiers like Hidden Markov Models (HMM) and Dynamic Bayesian Networks (DBN) [3] [4]. Alternative strategies apply static techniques such as Support Vector Machines (SVM) that process statistical functionals of low-level features which are computed over longer data segments [5]. However, these techniques only model a limited amount of contextual information which does not take advantage of the fact that human emotion usually is slowly varying and highly context-dependent. A unimodal framework for short-term context modeling in dyadic interactions was proposed in [6].

In this paper we propose a multimodal emotion recognition framework that merges audio-visual information at the feature level and uses a classification technique that allows for the modeling of long-range temporal dependencies. We account for contextual information by applying Long Short-Term Memory (LSTM) networks [7] which have shown to prevail over standard recurrent neural networks (RNN) whenever a high amount of context has to be considered (e. g. [8]). This concept is able to model *emotional history* and overcomes the so-called *vanishing gradient problem* in conventional recurrent neural nets. A first attempt to use discriminatively trained LSTM networks for unimodal emotion prediction has been made in [5]. In contrast to [5], this paper also investigates the modeling of *bidirectional* context, which can be used to refine the emotion prediction of past observation once more context is available. Moreover, the system presented in this paper is the first *multimodal* technique using Long Short-Term Memory.

We focus on the recognition of dimensional emotional labels, valence and activation, instead of categorical emotional tags, such as 'anger' or 'happiness'. Therefore, our system is trained and evaluated also on non-prototypical data; meaning utterances that are labeled differently by different annotators and may not have a categorical label. We classify a variety of emotional manifestations, which may include ambiguous emotions, subtle emotions, or mixtures of emotions. This allows for a more realistic AER performance assessment, since a real-life system has to classify *all* data that is recorded. The acoustic and facial feature extraction applied herein is based on the technique introduced in [3]. Yet, in contrast to [3], our approach does not use phoneme-dependent models or viseme information and thus does not rely on the correct phoneme transcription. Finally, the feature extraction and the recognition experiments are performed in a subject-independent way, which is a challenging but realistic experimental setup. All the above design decisions can be seen as further steps towards real-life applicability.

In our experiments we use a large multimodal and multi-subject acted database [9], which was collected so as to contain emotional manifestations that are non-prototypical and resemble as much as possible real-life emotional expression. In addition to classifying the degree of valence and activation separately, we also investigate the modeling of clusters in the emotional space (as in [5]). We compare the recognition performance of our bidirectional LSTM network to a conventional SVM approach and to fully-connected HMMs. Short-term context is incorporated into the HMM framework using a first-order 'language model', based on emotional state transition probabilities as observed in the training set. According to our results, bidirectional LSTM networks generally outperform HMM and SVM classifiers, a finding which suggests that long-term context modeling is important for emotion recognition tasks.

2. Database and Dimensional Labeling

In this study, we use the Interactive Emotional Dyadic Motion Capture (IEMOCAP) database [9]. This database contains approximately 12 hours of audio-visual data from five mixed gender pairs of actors, male and female (ten subjects in total). IEMOCAP contains detailed face information obtained from motion capture as well as video and audio of each session. Two techniques of actor training were used; scripts and improvisation of hypothetical scenarios. The goal was to elicit emotional displays that resemble natural emotional expression. Dyadic sessions of approximately five minute length were recorded and were later manually segmented into utterances. Each utterance was annotated into nine categorical (such as anger, happiness, or neutrality) as well as dimensional tags (valence, activation, dominance), by multiple human annotators. Dimensional tags take integer values that range from one to five. The dimensional tag of an utterance is the average of the tags given by at least two annotators. In this study, we focus on the classification of valence and activation, which enables us to make use of all the available data, even utterances for which there was no inter-annotator agreement, and thus no categorical label exists. Such data are a relatively large portion of the database (approximately 17 % of the total utterances).

3. Methodology

3.1. Feature Extraction and Feature Selection

The IEMOCAP database contains detailed facial marker information, as illustrated in figure 1. Face markers are normalized for head rotation and translation and the tip nose marker is defined as the local coordinate center of each frame. We use the (x,y,z) coordinates from 46 facial markers. In order to obtain a low-dimensional representation of the facial marker information, we use Principal Feature Analysis (PFA) [10]. This method performs Principal Component Analysis (PCA) as a first step and selects features (here marker coordinates) so as to minimize the correlations between them. We select 30 features, because the PCA transformation explains more than 95% of the total variability, and we append the first derivatives, resulting in a 60-dim representation. In addition, the facial features are normalized per speaker in order to smooth out individual facial characteristics that are unrelated to emotion. 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 detail in our previous work [11].

In addition, we extract from the waveform a variety of speech features; 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 and are normalized using z-standardization (the statistics are computed from the corresponding trainset). The audio and visual features are extracted at the same framerate of 25 ms, with a window size of 50 ms. Since our experiments are organized in a cyclic leave-one-speaker-out (LOSO) cross validation, all the normalization constants for the face and voice features, as well as the PCA transforms, are computed in a subject-independent way from the training set of each fold.

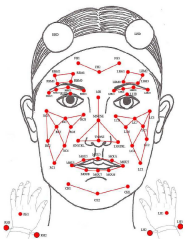

Figure 1: *Facial marker positions*

For LSTM and SVM classification we use a set of utterance level statistical functionals that are computed from the low-level acoustic and visual features. These functionals include means, standard deviations, linear and quadratic regression parameters (slope, offset, linear/quadratic approximation error), maximum and minimum positions, skewness, kurtosis, quartiles, interquartile ranges, and percentiles. All functionals were calculated using our openEAR toolkit [12]. 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. This results in an automatic selection of between 66 and 224 features, depending on the classification task and the fold. For the valence classification task, on average 84 % of the selected features are facial features, whereas for classification of the degree of activation, only 44 % of the features selected via CFS are facial features. This underlines the fact that visual features tend to be well-suited for determining valence while acoustic features rather reveal the degree of activation and agrees with the unimodal classification results that are presented in the results section. For a detailed analysis of the selected features see table 1.

Table 1: *Distribution of the features selected via CFS for the classification of valence (VAL) and activation (ACT) as well as for the discrimination of 3, 4, and 5 clusters in emotional space (see section 3.4).*

feature group	VAL	ACT	3 clusters	4 clusters	5 clusters
pitch	5 %	4 %	3 %	4 %	3 %
energy	0 %	1 %	1 %	1 %	1 %
MFCC	4 %	21 %	11 %	11 %	10 %
MFB	7 %	30 %	18 %	19 %	21 %
lower face	63 %	32 %	50 %	49 %	48 %
upper face	21 %	12 %	17 %	16 %	17 %

3.2. Bidirectional LSTM Networks

In order to model contextual information between successive utterances, we apply bidirectional Long Short-Term Memory (BLSTM) recurrent neural networks. Bidirectional RNNs are composed of two recurrent network layers, whereas the first one processes the sequence forwards and the second one processes it backwards. Since both networks are connected to the same output layer, the bidirectional net has access to the entire information about past and future data points in the sequence. During training, the amount of contextual information that the network uses is learned and does not have to be specified manually. For emotion recognition, bidirectional networks can be used e. g. to refine the emotion prediction of past observations.

A drawback of conventional bidirectional and unidirectional RNN architectures is that the range of context that can actually be accessed is limited as the influence of a given input on the hidden layer either decays or blows up exponentially over time (*vanishing gradient problem*). This led to the introduction of Long Short-Term Memory RNNs [7]. An LSTM layer is composed of recurrently connected memory blocks, each of which contains one or more recurrently connected memory cells, along with three multiplicative 'gate' units: the input,

output, and forget gates. The gates perform functions analogous to read, write, and reset operations. More specifically, the cell input is multiplied by the activation of the input gate, the cell output by that of the output gate, and the previous cell values by the forget gate. Their effect is to allow the network to store and retrieve information over long periods of time. This principle solves the vanishing gradient problem and gives access to long range context information. The combination of bidirectional networks and LSTM is called bidirectional LSTM. A detailed explanation of BLSTM networks can be found e. g. in [13].

The LSTM networks applied for our experiments consist of 128 memory blocks with one memory cell per block. The number of input nodes corresponds to the number of different features per utterance whereas the number of output nodes corresponds to the number of target classes. To improve generalization, we add zero mean Gaussian noise with standard deviation 0.6 to the inputs during training. All networks are trained using a learning rate of 10^{-5}. The bidirectional networks consist of two LSTM layers (one for forward and one for backward processing) with 128 memory blocks per input direction.

3.3. HMM and SVM Classification

As an alternative classification approach, we also examine a dynamic, generative classification framework using Hidden Markov Models. Our motivation is to model the underlying dynamics of audio-visual emotional expression. We train fully-connected 3-state HMMs for the facial and vocal modality, as well as audio-visual HMMs. For each classification task, we train one HMM for each class using the training utterances and during the test stage we recognize the most probable class. We use frame-level features, as opposed to the BLSTM experiments where we process statistical functionals over features. For the facial HMMs we use a 60-dim feature vector containing 30 normalized PFA features and their first derivatives. For the vocal HMMs, we use a 58-dim feature vector containing 27 normalized MFBs, normalized pitch and energy values and their first derivatives. For the audio-visual HMMs, we combine the synchronous face and speech features at the feature level (118 dimensions). All HMMs are trained using the HTK Toolbox[14]. In order to have a rough, local description of the past emotional context, we incorporate a first-order 'language model' (LM) in our classification framework. Specifically, from the training set of each fold, we count the number of transitions for each pair of the classes of our problem. In that way we obtain an estimate of the transition probabilities from one class to the other. During the test stage, we select the class that maximizes the product of the class probability for the current utterance, and the transition probability from the previous class to the current class.

Furthermore, we compare the performance of the BLSTM networks to static classification of utterance level feature functionals via Support Vector Machines. The SVMs have a polynomial kernel (degree 1) and are trained using the sequential minimal optimization (SMO) algorithm.

3.4. Ground Truth Labels

The valence and activation values range from one to five and could be non-integer, since the decisions of two evaluators are averaged for each utterance label. We perform classification of three levels of valence (negative, neutral, and positive, corresponding to ratings {1,1.5,2}, {2.5,3,3.5}, and {4,4.5,5}, respectively) and activation (low, medium, and high, again corresponding to {1,1.5,2}, {2.5,3,3.5}, and {4,4.5,5}). The class sizes are not balanced since medium values of labels are more common than extreme values. We also examine the joint classification of the emotional dimensions by building three, four, and five clusters in the valence-activation space. The cluster midpoints in the emotional space are determined by applying the K-means algorithm on the annotations of the respective training sets. 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. The intuition for clustering the valence-activation space is to build classifiers that provide richer and more complete emotional information, that could correspond to generic emotional tags. For example, as can be seen in figure 2, the coordinates of the cluster midpoints are interpretable: when building three clusters, the midpoints roughly correspond to the affective states 'angry', 'neutral/sad', and 'happy'. The average standard deviation of the cluster centroid coordinates across the ten folds is as low as 0.05.

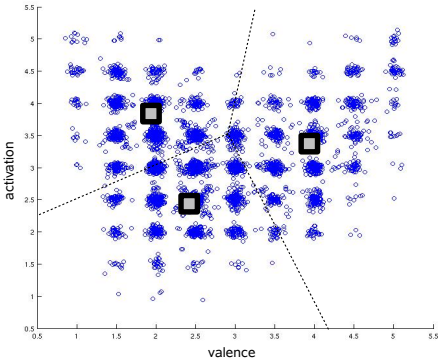

Figure 2: *Annotations of the IEMOCAP training set for fold 1 with cluster midpoints (gray squares) and resulting class borders (dotted lines) for the 3-class task; a small amount of random noise is added to the annotations for visualization purposes*

4. Experimental Results and Discussion

Our experiments are organized in a cyclic leave-one-speaker-out cross validation. The mean and standard deviation of the number of test and training utterances across the folds is 498 ± 60 and 4475 ± 61, respectively. Each utterance may range from a few tenths of seconds to a minute. For each fold, we compute the accuracy and the (unweighted) precision, recall, and F1 measure. The presented recognition results are the subject-independent averages over the ten folds.

Table 2 shows the recognition performances for discriminating three levels of valence and activation, respectively. The unimodal HMM results confirm that facial features tend to be more important for valence classification while acoustic features are well-suited for activation classification. Generally, multimodal classification outperforms unimodal AER. The best F1-measure for valence can be obtained using a BLSTM network (65.18 %), whereas the performance for unidirectional LSTM networks is only slightly lower (F1-measure of 63.66 %). This indicates that modeling the long-range context between successive utterances is very important. Incorporating a bigram language model into the HMM recognition framework, also leads to a performance gain, which again underlines the importance of context modeling. For activation, we observe a lower performance of LSTM modeling. A major reason for this is the imbalance of the class distribution: the majority of utterances are labeled as 'medium activation' so that the amount of training data for the remaining two activation classes is insufficient (also see figure 2). For the activation class, the HMM+LM framework handles this class imbalance better and achieves the

Table 2: *Recognition performances for discriminating three levels of valence and activation using face (f) and voice (v) features: accuracy (ACC), unweighted recall (REC), precision (PREC), and F1-measure (F1).*

classifier	features	ACC	REC	PREC	F1
		valence			
HMM	v	47.08	47.11	48.20	47.62
HMM	f	55.53	60.07	56.77	58.29
HMM	v+f	59.27	58.81	61.68	60.17
HMM+LM	v+f	61.07	62.85	61.11	61.91
SVM	v+f	61.49	61.50	63.59	61.45
LSTM	v+f	62.35	63.77	63.80	63.66
BLSTM	v+f	63.92	64.71	65.87	**65.18**
		activation			
HMM	v	55.06	61.68	50.93	55.77
HMM	f	43.87	51.86	47.48	49.30
HMM	v+f	51.33	52.56	60.16	55.90
HMM+LM	v+f	57.65	57.62	57.75	**56.89**
SVM	v+f	70.53	50.39	60.30	51.30
LSTM	v+f	68.84	50.58	58.45	53.89
BLSTM	v+f	67.31	52.53	58.46	55.18

highest performance (F1-measure of 56.89 %). This may also suggest that such dynamic classification frameworks may capture important dynamic information about the evolution of emotional expression.

Table 3: *Recognition performances for discriminating three, four, and five clusters in emotional space using face (f) and voice (v) features: accuracy (ACC), unweighted recall (REC), precision (PREC), and F1-measure (F1).*

classifier	features	ACC	REC	PREC	F1
		3 clusters			
HMM	v+f	67.03	66.87	67.99	67.37
HMM+LM	v+f	67.03	66.89	68.04	67.41
SVM	v+f	68.91	68.58	69.20	67.95
LSTM	v+f	70.17	69.54	71.20	70.33
BLSTM	v+f	72.31	71.88	72.84	**72.34**
		4 clusters			
HMM	v+f	55.70	55.93	55.69	55.73
HMM+LM	v+f	56.87	56.33	56.44	56.31
SVM	v+f	60.77	58.36	59.12	57.10
LSTM	v+f	63.69	61.00	62.86	61.87
BLSTM	v+f	64.30	61.92	63.85	**62.78**
		5 clusters			
HMM	v+f	49.94	50.94	48.87	49.76
HMM+LM	v+f	50.81	50.99	50.17	50.41
SVM	v+f	51.49	49.52	50.99	48.55
LSTM	v+f	56.19	53.89	56.25	**55.00**
BLSTM	v+f	56.31	53.76	56.13	54.84

A more balanced class distribution and a better class separability can be obtained when jointly classifying valence and activation by assigning the utterances to clusters that are learned in a data-driven way as explained in section 3.4: for the distinction between three clusters BLSTM networks achieve an F1-measure of 72.34 % (see table 3). For four and five clusters they achieve F1-measures of 62.78 % and 55.00 % respectively. For all cluster prediction tasks, we observe similar trends: LSTM modeling prevails over HMM and SVM classification and bidirectional context outperforms unidirectional context (except for the five-cluster task, where there is no significant difference between LSTM and BLSTM). The HMM+LM and SVM classification frameworks achieve comparable, and lower, results.

In general, the BLSTM framework which is able to incorporate long-range bidirectional context information, prevails over other classification frameworks which use no or limited contextual emotional information, such as the SVM and the HMM+LM respectively.

5. Conclusion and Future Work

We applied a context-sensitive multimodal framework for affect recognition from acoustic and facial features, exploiting long-range contextual information by using bidirectional Long Short-Term Memory Networks. Various challenging subject-independent classification tasks revealed that BLSTM modeling prevails over conventional dynamic or static classification strategies. Thereby we focused on a realistic experimental setup in the sense of including ambiguous and non-prototypical emotions.

In the future, it would be interesting to investigate dynamic modeling of *low-level* features using a multimodal Long Short-Term Memory framework. This might enable a more accurate description of emotional expression dynamics *within* an utterance and could increase multimodal affect recognition performance. Additionally, the benefit of including linguistic information into the presented multimodal AER system should be examined.

6. References

[1] 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*, vol. 18, no. 1, pp. 32–80, 2001.

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

[3] 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.

[4] Y. Zhang and Q. Ji, "Active and dynamic information fusion for facial expression understanding from image sequences," *IEEE Transactions On Pattern Analysis And Machine Intelligence*, vol. 27, no. 5, pp. 1–16, May 2005.

[5] 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*, Brighton, UK, 2009, pp. 1595–1598.

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

[7] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9(8), pp. 1735–1780, 1997.

[8] M. Wöllmer, B. Schuller, F. Eyben, and G. Rigoll, "Combining long short-term memory and dynamic bayesian networks for incremental emotion-sensitive artificial listening," *IEEE Journal of Selected Topics in Signal Processing, Special Issue on Speech Processing for Natural Interaction with Intelligent Environments*, 2010.

[9] 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.

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

[11] 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.

[12] F. Eyben, M. Wöllmer, and B. Schuller, "openEAR - introducing the Munich open-source emotion and affect recognition toolkit," in *Proc. of ACII*, Amsterdam, The Netherlands, 2009, pp. 576–581.

[13] A. Graves, "Supervised sequence labelling with recurrent neural networks," Ph.D. dissertation, Technische Universität München, 2008.

[14] 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.