Ensemble of SVM Trees for Multimodal Emotion Recognition

Viktor Rozgić, Sankaranarayanan Ananthakrishnan, Shirin Saleem, Rohit Kumar, and Rohit Prasad
Speech Language and Multimedia Technologies, Raytheon BBN Technologies
vrozgic,santhana,ssaleem,rkumar,rprasad@bbn.com

Abstract— **In this paper we address the sentence-level multi-modal emotion recognition problem. We formulate the emotion recognition task as a multi-category classification problem and propose an innovative solution based on the automatically generated ensemble of trees with binary support vector machines (SVM) classifiers in the tree nodes.**

We demonstrate the efficacy of our approach by performing four-way (anger, happiness, sadness, neutral) and five-way (including excitement) emotion recognition on the University of Southern California's Interactive Emotional Motion Capture (USC-IEMOCAP) corpus using combinations of acoustic features, lexical features extracted from automatic speech recognition (ASR) output and visual features extracted from facial markers traced by a motion capture system. The experiments show that the proposed ensemble of trees of binary SVM classifiers outperforms classical multi-way SVM classification with one-vs-one voting scheme and achieves state-of-the-art results for all feature combinations.

I. INTRODUCTION

Automated emotion recognition has an important role in many applications related to modeling and analysis of human behavior. Choice of classification methods and informative features sets dominantly influence accuracy on the emotion recognition tasks and represent active research areas [1].

The main contribution of this paper is an emotion recognition method based on ensemble of trees with binary SVM classifiers in nodes. We define each tree node by: (a) *category set* - a set of emotion categories that can be discriminated at that node; and (b) a binary classifier that operates on some subset of the node's *category set* and is trained on some subset of the available features. For example, if both acoustic and lexical features can be used the binary classifier could be trained using only acoustic, only lexical or combination of these features. The choice of a binary classifier in a parent node defines *category sets* of the child nodes. Namely, the category set for the child node is a union of categories associated with the class "1" (or class "0") of the parent node and categories from the parent's *category set* that are not used in parent node classifier. We generate an ensemble of such trees of classifiers by randomizing classifier choices in each tree node. Namely, we randomly pick a classifier from the pool of best classifiers (with respect to the cross validation accuracy on the training set as the selection criterion) that operate on the node's *category set*.

Our approach is related to several other research works on tree structured classifiers for emotion recognition. While we automatically generate a tree ensemble, methods introduced in [2,3,4] use a single tree structure designed manually based on experts' knowledge. Another difference, compared to the methods proposed in [2,3] that exploit only classifiers that discriminate between two emotion categories, is that, as in [4], we allow classifiers to discriminate between sets of multiple emotion categories. From the algorithmic perspective, our method is motivated by work on tree structured SVM classifiers. While the most similar approaches presented in [5,6] emphasize minimization of the tree training complexity we train and optimize all possible binary classifiers for the given category set and generate trees using only the classification performance as a criterion. Additionally, the proposed method belongs to a broader class of approaches for combining multiple binary classifiers into a single multi-class classifier, e.g. one-vs-one and one-vs-all voting schemes and their generalizations via error correcting code schemes [7].

We tested our approach on the University of Southern California's Interactive Emotional Motion Capture (USC-IEMOCAP) corpus [8] using combinations of acoustic, lexical and visual features. The acoustic feature set contains the sentence level voicing statistics and the sentence-level statistical functionals (mean, variance, range, etc.) of the basic frame-level features (energy, pitch, formants) and voice quality features (jitter, shimmer). Instead of representing the non-stationary MFCC features using statistical functionals we use a set of model based features [12] obtained by scoring all MFCC vectors in a sentence by emotion-dependent Gaussian mixture models (GMM). We normalize these score vectors and fuse them by: (a) calculating mean of the normalized score vectors on the sentence-level, (b) generating histograms by voting for the highest scoring emotion model, and (c) estimating parameters of the Dirichlet distribution that generates normalized scores within a sentence. We augment acoustic features by two types of linguistic features derived from the ASR output: binary word-stem presence indicators and sentiment features based on bag-of-words sentiment categories. Finally, visual feature set uses frame level marker-distance features derived from the trajectories of the facial motion capture markers and fuses them on the sentence level by the same set of functionals used for the acoustic features.

We conducted all experiments in a leave-one-speaker-out configuration, on the 4-way (*anger*, *happiness*, *sadness*, *neutral*) classification problem, a standard approach for the USC-IEMOCAP dataset, and the 5-way classification problem (including the "excitement). Since the tree generation involves randomization we compared a set of tree ensembles with the baseline multi-way support vector machine (SVM) with 1-vs-1 voting scheme. The statistical tests proved that the proposed tree ensemble outperforms the baselines. The overview of the methodology is presented in Fig. 1.

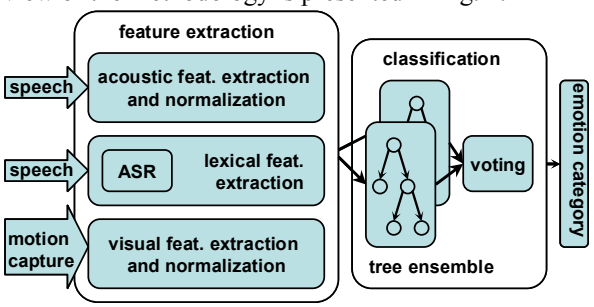

Fig. 1. Methodology overview

The remainder of the paper is organized in the following way. Section II provides a description of the USC-IEMOCAP

dataset and acoustic, lexical and visual features we used. In Section III we describe the algorithm we used to generate ensemble of SVM trees. Section IIV details our experimental methodology and summarizes key results. In Section V we conclude the paper and provide directions for future research.

II. DATASET AND FEATURES

We tested the proposed emotion recognition scheme on the USC-IEMOCAP database [8]. USC-IEMOCAP contains 12 hours of data (audio, video and motion capture trajectories of facial markers), split into 5min dyadic interactions between professional female-male actor pairs. Each dyadic interaction is manually split into sentences corresponding to speaker turns in the dialogue, manually transcribed and annotated by human annotators. Sentences are labeled by at least three annotators with categorical emotion labels (*anger*, *happiness*, etc.) and emotion attributes (*valence*, *activation* and *dominance* on the scale 1 to 5).

In order to match experimental conditions in previously reported categorical emotion recognition studies on USC-IEMOCAP [4,9], we consider only 5531 sentences with labelers' majority vote agreement on one of five categories: *anger*, *happiness*, *sadness*, *excitement* and *neutral*.

A. Acoustic Features

We use numerous sentence-level acoustic features derived as statistical functionals (mean, standard deviation, range) of frame-level features: F0, intensity, and first two formants. We extract 36-dimensional MFCCs with delta and acceleration values. In order to mitigate effects of variability in speaker characteristics and recording conditions, we normalize the energy mean of the speech signal, and perform cepstral mean normalization of MFCCs for each speaker in the dataset.

We extract jitter and shimmer, voice quality features correlated with negative emotions. We use Praat [10] to compute local and perturbation values for a total of five jitter features and six shimmer averaged across the utterance.

Additionally, we extract sentence-level voicing statistics (fraction of unvoiced frames in a sentence, number and average duration of voice break intervals) as intuitive indicators of excitement. We have successfully used this feature set for emotion recognition on the Berlin Emotional Speech Corpus [11].

Further, we transform the segmental MFCC features in a way that retains information relevant for emotion classification and avoid simplification through the direct computation of sentence level statistics. We train M C-component Gaussian mixture models, one for each emotion category using training frame-level feature vectors. We calculate likelihoods for all feature vectors within the sentence for each model component. For each frame we normalize the component likelihoods over all models and components creating a vector of normalized model likelihoods. Vectors of normalized likelihoods are multinomial distributions and we used three ways to fuse them into a single fixed dimensional sentence-level feature vector [12]. First, we calculate mean distribution over all normalized likelihood vectors corresponding to frames within a sentence. Second, we calculate histogram of votes for the most likely component on all frames in the sentence. Finally, we calculate parameters of the Dirichlet distribution that generate all normalized likelihood vectors within the sentence. We estimate these parameters in the maximum likelihood sense using the moment matching approximations [13].

B. Lexical Features

We capture content-based cues for emotion recognition using two types of lexical features extracted from ASR output, sentiment and word stem features. We removed stop-words and applied Porter stemming to remove common morphological and inflectional endings. We augmented the set of stemmed words with non-speech tokens, such as laughter and sighs, present in the ASR output. We defined word stem features as binary indicators of the presence or absence of all unigram tokens from (approximately 2000) the training dataset in the processed sentence. Additionally, we extracted 125 features representing numbers of words in a sentence that belong to different sentiment categories. Each category, based on Linguistic Inquiry and Word Count (LIWC) [14] and General Inquirer (GI) system [15] lexicons, is represented by a bag of word stems. The LIWC-derived categories include: positive emotions, optimism, negative emotions, anxiety and fear, anger, sadness, swear words; and the GI categories include arousal, activation and valence.

C. Visual Features

We extracted a set of visual features inspired by Facial Animation Parameters (FAPs) from the motion capture marker trajectories. These features represent vectors between different points on the face and include: (1) vectors between the top of the cheek and the eyebrow markers (approximating the squeeze present in a smile); (2) vectors from the cheek to the nose, mouth, and chin markers; (3) mouth opening measures (length and width); (4) vectors between lip corners and the nose marker (related with smiles); (5) vectors between forehead and eyebrow markers; and (6) eyebrow angle and curvature measures. Further, we extracted the same set of the statistical functionals as for the acoustic features.

III. ENSEMBLE OF SVM CLASSIFIER TREES

A recent study [4] showed that better accuracies on multi-category emotion recognition task can be achieved if the one-vs-one and one-vs-all multi-class voting schemes are replaced by a handcrafted tree with binary classifiers in its nodes.

We propose to generalize this approach and automatically generate an ensemble of trees with binary SVM classifiers in tree nodes. As the first step we train all possible binary SVM classifiers for the given number of categories (e.g. for 5 categories there are 90 different binary classifiers) and combinations of the available features (e.g. for acoustic and lexical features we would train binary classifiers using only acoustic, only lexical and combination of these features). Second, we generate ensemble of trees of binary SVM classifiers. Each node in the tree is defined by a *category set* and a binary classifier. The *category set* describes data available in the node, and the binary classifier can operate on any subset of the *category set*. We generate the classifier tree in the following way. First, we randomly pick a classifier from a set of binary classifiers that have accuracy (estimated by cross validation on the training set) that is at least ε percent of the best classifier's accuracy. Second, we augment all categories that belong to class "1" (class "0") by categories that belong to the *category set* but do not belong to class "1" or to class "0" of the chosen binary classifier. The obtained set represents *category set* for left (right) child of the current node. Finally, we repeat the first two steps while the *category sets* of leaf nodes have more than one element. Randomization in the first step allows us to generate different

classifier trees in the ensemble, at the same time biasing choices towards more confident classifiers.

The described method is formalized by algorithm in Fig. 2.

GenerateNode (S, ε)

S : Set of available categories.

ε : Classifier performance threshold parameter.

- Collect a set $C(S)$ of all binary classifiers C_i $(i = 1,..., N)$ that operate on some subset of *category set* $S(C_i,1) \cup S(C_i,0) \in S$, where $S(C_i,1)$ and $S(C_i,0)$ denote categories that belong to class "1" and class "0" of the classifier C_i .
- Select a subset $C_\varepsilon(S)$ that classifier with accuracy at least ε percent of the best accuracy in $C(S)$ (based on the cross validation on the training set).
- Randomly select classifier C^* from the set $C_\varepsilon(S)$. This is the selected classifier in the current tree node.
- Create sets of categories for the right (class "1") and left (class"0") child of the current node: $S_1 = S \setminus S(C^*,0)$ and $S_0 = S \setminus S(C^*,1)$.
- If S_1 is nonempty, ***GenerateNode*** (S_1, ε) . If S_0 is nonempty, ***GenerateNode*** (S_0, ε) .

Fig. 2. Randomized tree generation algorithm

A sample tree for the 4-category problem is presented in Fig. 3.

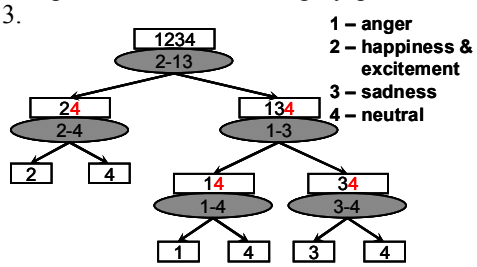

Fig. 3. Example – discriminative classifier tree (rectangles denote set of node classes and ellipses node classifiers)

IV. EXPERIMENTAL RESULTS

In order to match evaluation setups for emotion recognition results reported for the USC-IEMOCAP dataset [4,9] we compared the proposed method with two baselines on the 4-category (ANG-anger, HAP-happiness, SAD-sadness and NEU-neutral) classification task. The first baseline is a 4-way SVM classifier based on 1-vs-1 voting scheme. The second baseline is a 5-way SVM classifier based on 1-vs-1 voting scheme followed by merging the "excitement" and "happiness" categories to a single "happiness" category to produce 4 category output. Additionally, in order to separately analyze effects caused by merging "happiness" and "excitement" categories and benefits stemming from use of the tree ensemble, we compared the proposed tree-ensemble method with the multi-way SVM based on 1-vs-1 voting scheme on the 5-category emotion recognition task.

We performed all evaluations using leave-one-speaker-out 10-fold cross-validation. For each cross-validation split we extracted acoustic, lexical and visual features for test and train sentence sets. The acoustic features, extracted as described in Section II-A using 25ms processing windows with 10ms shift, formed 85–dimensional vectors (12 statistical functionals of F0,F1,F2 and energy, 11 voice quality, 3 voicing statistics and 59 model based MFCC features). Model based MFCCs for each emotion category were based on 5-mixture GMMs obtained using the training set feature vectors. Lexical features, described in Section II-B, were extracted from the output of the ASR system with models trained on 1700 hours of transcribed broadcast news using 65K word dictionary. Due to a mismatch between the acoustic models and spontaneous speech style of USC-IEMOCAP (including speaker overlaps) the ASR word error rate we got was high (~40%). We extracted ~2000 word-stems on the training sets and created binary word-stem indicator vectors for each train and test sample. We augmented these vectors with 125-dimensional vectors with coordinates that represent counts of words that belong to 125 sentiment bags-of-words categories. Finally, visual features were represented by 342-dimensional vectors of FAP inspired features described in Section II-C.

We formed test and train samples by concatenating feature vectors for all features used in a particular experiment. We further split each training set into 9 single speaker folds and used cross-validation multi-way classification accuracy on the training set as a cost function to optimize parameters of the SVM classifier with radial basis function (RBF) kernel. We optimized the parameter $C \in \{2^{-1},...,2^{4}\}$ that controls trade-off between training errors and the SVM margin size and the kernel width $\gamma \in \{2^{-7},...,2^{-2}\}$. We used the optimal parameters to train SVM classifiers on the full training set.

In Tables 1 and 2 we present unweighted and weighted classification accuracies as well as separate emotion category classification accuracies for different feature sets for two baseline classifiers. The unweighted accuracy is defined as a mean of accuracies for different emotion categories, while the weighted accuracy is a weighted mean with weights proportional to the emotion category sample size.

The baselines exhibited performance improvements with introduction of additional features, proving complementarity of the used feature sets. The second baseline outperformed the first one achieving the best reported performances [4, 16] on the USC-IEMOCAP dataset for all feature combinations proving adequacy of the chosen feature set.

Table 1. First baseline: Accuracies for the multi-way SVM with one-vs-one voting scheme for 4 categories (A-acoustic, L–lexical, V-visual features; UW-unweighted accuracy and W- weighted accuracy)

	UW	W	ANG	HAP	SAD	NEU
A	60.2	59.3	65.2	55.7	64.5	55.5
L	43.8	45.6	43.8	41.9	28.3	61.4
V	50.4	51.3	40.5	64.8	50.7	45.7
A+L	65.7	64.9	70.0	63.8	69.2	59.8
A+L+V	66.4	66.1	73.1	72.4	61.9	58.1

Table 2. Second baseline: Accuracies for the 4-way classifier based on 5-way SVM one-vs-one voting scheme and late merge of "happiness" and "excitement" categories (A-acoustic, L–lexical and V-visual features; UW-unweighted accuracy and W- weighted accuracy)

	UW	W	ANG	HAP	SAD	NEU
A	60.5	59.4	66.4	55.8	64.9	55.0
L	47.7	45.8	64.9	47.8	40.0	38.1
V	50.9	50.8	43.9	65.6	50.3	43.8
A+L	67.1	65.1	77.9	64.5	68.6	57.2
A+L+V	66.5	65.5	76.1	74.0	61.3	54.6

In Table 3 we present average accuracies for the proposed tree ensemble method. As described in Section III we generated 10 ensembles, each with 20 trees. Accuracies proved not to be sensitive to choice of the epsilon parameter value in the range [0.5, 0.95] and exhibited decrease for

epsilon values in interval [0.95, 1]. We present results for epsilon value 0.75. We emphasize that for "A+L" tree ensemble we selected classifiers from the pool of binary classifiers trained using only acoustic, only lexical and combination of acoustic and lexical features. Similarly, for the "A+L+V" ensemble we used classifiers based on only acoustic, lexical and visual features and on combinations of acoustic and lexical and acoustic, lexical and visual features.

Table 3. Average accuracies over 10 ensembles of 20 SVM trees (A-acoustic, L–lexical, and V-visual features; UW-unweighted accuracy, W- weighted accuracy)

	UW	W	ANG	HAP	SAD	NEU
A	60.9	60.8	66.1	53.9	65.5	58.1
L	48.6	48.5	63.1	49.6	42.0	39.5
V	51.3	51.5	41.8	63.6	52.6	47.0
A+L	67.4	67.4	77.8	63.2	68.3	60.4
A+L+V	69.4	69.5	78.1	69.2	67.1	63.0

The averaged tree-ensembles outperformed both baselines. The greedy tree generation algorithm solves easier classification tasks first and avoids error propagation by postponing more difficult classification tasks for the leaf nodes. The biggest improvement was obtained for the "A+L+V" classification task. Reason for this is that all binary classifiers in the multi-way SVM have the same kernel parameters and these parameters were optimized based on the joint performance after 1-vs-1 voting and not for each binary classifier separately. For this reason some binary classifiers trained on a feature subset perform better then the classifiers trained using all available features.

Table 4. Comparison of the SVM tree ensembles and the second baseline on 4 category classification (mean, standard deviation, minimum and maximum of the unweighted accuracy over ten 20-tree ensembles; A-acoustic, L–lexical, and V-visual features)

	μ	σ	min	max	UW_2	α
A	60.9	0.03	60.8	61.0	60.5	10^{-6}
L	48.6	0.05	48.5	48.7	47.7	10^{-8}
V	51.3	0.16	51.0	51.5	50.9	0.02
A+L	67.4	0.12	67.2	67.6	67.1	0.02
A+L+V	69.4	0.32	68.9	69.7	66.5	10^{-5}

In Table 4 we present mean, standard deviation, minimum and maximum of the unweighted accuracy calculated over 10 tree ensembles for the 4-category classification task and compare it with the second baseline using one-sided t-test. The last column presents critical significance levels at which the tree ensembles outperform the baseline classifier.

Table 5. Comparison of the SVM tree ensembles and the second baseline on 5 category classification (mean, standard deviation, minimum and maximum of the unweighted accuracy over ten 20-tree ensembles; A-acoustic, L–lexical, and V-visual features)

	μ	σ	min	max	UW_2	α
A	50.8	0.02	50.7	50.8	50.5	10^{-7}
L	41.1	0.02	41.0	41.1	40.3	10^{-9}
V	44.5	0.05	44.4	44.6	44.5	0.5
A+L	57.9	0.15	57.7	58.2	57.4	0.005
A+L+V	59.9	0.20	59.6	60.1	58.8	0.001

We compare the proposed tree ensemble classifier with the second (better) baseline on 5-category recognition task to provide insight into ensemble performance decoupled from the category merging effects (Table 5). The proposed method outperforms the baseline with 0.5% significance level for all classification tasks except for 5-category classification using only the visual features where the performances are equal.

V. CONCLUSIONS

In this paper we presented state-of-the-art results for the 4-category emotion recognition on the USC-IEMOCAP database for all feature set combinations: acoustic – 60.9%, acoustic and lexical – 67.4%, and acoustic, lexical and visual – 69.4%. We propose a novel classifier based on the automatically generated ensemble of SVM tree classifiers. The proposed classifier outperformed two state-of-the-art baselines both on 4-category and 5-category recognition tasks.

Future work will focus on refinement of the feature set identifying frame subsets which are the most relevant for sentence-level emotion recognition. Additionally, we plan to develop alternative ways to generate classifier trees without need to train all possible binary classifiers.

ACKNOWLEDGMENT

This paper is based upon work supported by the DARPA DCAPS Program. The views expressed are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

We would like to thank Professor Shrikanth Narayanan for providing us access to the USC-IEMOCAP database.

REFERENCES

[1] B. Schuller, A. Batliner, S. Steidl, D. Seppi, "Recognising Realistic Emotions and Affect in Speech: State of the Art and Lessons Learnt from the First Challenge", Speech Communication, 53(10), 2011.

[2] Z. Xiao, E. Dellandrea, W. Dou, L. Chen, "Automatic hierarchical classification of emotional speech", In ISMW, 2007

[3] A. Hassan, R. I. Damper, "Multi-class and hierarchical SVMs for emotion recognition", In Proc. Interspeech, 2010

[4] C.-C. Lee, E. Mower, C. Busso, S. Lee and S. S. Narayanan, "Emotion recognition using a hierarchical binary decision tree approach",In Speech Communication, 53(10), pp. 1162-1171, 2011

[5] B. Fei, J. Liu, "Binary Tree of SVM: A New Fast Multiclass Training and Classification Algorithm", In IEEE Trans. Neural Networks, 17(3), 2006

[6] G. Madzarov, D. Gjorgjevikj and I. Chorbev, "A Multi-class SVM Classifier Utilizing Binary Decision Tree", Informatica, 33:233-241, 2009

[7] A. Passerini, M. Pontil, M., and P. Frasconi, "New results on error correcting output codes of kernel machines", In IEEE Trans. On *Neural Networks,* 15(1), 2004

[8] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. Chang, S. Lee and S. S. Narayanan,"IEMOCAP: Interactive emotional dyadic motion capture database", In Journal of Language Resources and Evaluation, 42(4):.335-359, 2008

[9] A. Metallinou, S. Lee and S. S. Narayanan, "Decision level combination of multiple modalities for recognition and analysis of emotional expression", In Proc. of ICASSP, 2010

[10] P. Boersma and D. Weenink, "Praat, a system for doing phonetics by computer," Glot International, 5(9): 341–345, 2001.

[11] S. Ananthakrishnan,A. Vembu, R. Prasad,"Model-based parametric features for emotion recognition from speech", In Proc. ASRU, 2011

[12] V. Rozgic, S. Ananthakrishnan, S. Saleem, R. Prasad, "Emotion Recognition using Acoustic and Lexical Features", In Proc. Interspeech 2012

[13] N. Wicker, J. Muller,R. K. R. Kalathur O. Poch, "A maximum likelihood approximation method for Dirichlet's parameter estimation", Journal of Comput. Stat. and Data Anal., 52(3), 2008

[14] J. W. Pennebaker, R. J. Booth, M. E. Francis, "Linguistic Inquiry and Word Count: A text analysis program", LIWC 2007

[15] P. J. Stone, "The General Inquirer: A Computer Approach to Content Analysis", The MIT Press, 1966

[16] E. Mower, M. J. Mataric, S. Narayanan, "A Framework for Automatic Human Emotion Classification Using Emotion Profiles", In IEEE Trans. On Audio, Speech and Language Processing, 19(5), 2011