Simplifying Emotion Classification Through Emotion Distillation

Emily Mower Provost* and Shrikanth Narayanan†
* University of Michigan, Ann Arbor, Michigan, USA
† University of Southern California, University Park, Los Angeles, California, USA

***Abstract*—Many state-of-the-art emotion classification systems are computationally complex. In this paper we present an emotion distillation framework that decreases the need for computational complex algorithms while maintaining rich, and interpretable, emotional descriptors. These representations are important for emotionally-aware interfaces, which we will increasingly see in technologies such as mobile devices with personalized interaction paradigms and in behavioral informatics. In both cases these technologies require the rapid distillation of vast amounts of data to identify emotionally salient portions. We demonstrate that emotion distillation can produce rich emotional descriptors that serve as an input to simple classification techniques. This system obtains results that match state-of-the-art classification results on the USC IEMOCAP data.**

I. Introduction

As computing power continues to grow, engineers and scientists have been able to obtain increasingly quantitative measures of human emotion and behavior. Emotion is integral to behavior comprehension as it underlies social communication [1], [2], [3]. Consequently, proper modeling of this complex signal will aid in the design of empathic assistive devices, devices designed to assess and describe emotional communication to their users. These systems must be able to describe emotion in an interpretable and robust manner. One method to meet these requirements is through emotion distillation. Emotion distillation is the process of generating a set of emotion-specific features from the original high-dimensional feature space that describes how emotion fluctuates over time. Simple classification of this distillation can then be performed (e.g., Figure 1). Simplified emotional computing can aid the development of emotionally aware mobile companions, mobile monitoring systems, and mobile assistive devices. There has been research into mobile emotion classification devices. For example, in [4], [5] the authors demonstrated that mobile affective devices could be developed using powerful handheld mobile computers. However, these systems are still too complex to implement natively on many hand-held mobile devices. In this work we demonstrate how emotion distillation can be used as a pre-processing stage in computationally simple emotion classification systems.

Distillation systems have important application in behavioral signal processing and natural language processing, in which vast amounts of data must be processed to make a few high-level judgments [6] or in which emotionally/behaviorally salient regions must be located [7]. Once located, these regions can be analyzed automatically or cued for follow-up by human experts in a closed-loop framework [8], [9]. Distillation features are strongly tied to the behavior of interest, naturally highlighting salient portions of the data. This simplifies the feature-space representation by reducing the quantity of spurious information. Emotion distillation can be viewed as a supervised emotion-specific approximation of factor analysis. The benefit of this approach is that classifiers operating on top of distilled emotional cues need not be computational complex. We will analyze a technique for emotion distillation and classification to assess the benefits this two-stage process.

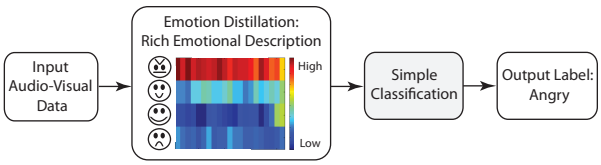

Fig. 1. A framework for emotionally rich classification.

The distillation approaches described in this paper are based on our Emotion Profile (EP) and emotogram representations. EPs characterize emotion in terms of affective blends (e.g., very angry and slightly sad) [10]. EPs are a multi-dimensional utterance-level characterization of emotion that describe the presence or absence of set basic emotions (e.g., angry, happy, neutral, sad). Emotograms are a temporal extension of EPs that capture the change in the presence and absence of these emotions in time. We have demonstrated that emotion can be classified by modeling the dynamics of the emotograms with Hidden Markov Models (HMM) [11]. However, it is not yet clear if computationally simple classification techniques can similarly model the temporal variation of these rich emotional descriptors.

In this work we distill emotion information using Emotion Profiles (EP) and emotograms. We classify the distilled emotogram representation using Hidden Markov Models (HMM) as demonstrated in [11] and compare these results to the computationally simple n-gram modeling, commonly employed in language modeling. We also assess the efficacy of computationally simpler static emotogram modeling using Linear Discriminant Analysis (LDA) and maximization over a time-condensed emotogram (simple summation, the simplest approach). Accurate classification results suggest that emotion distillation is a rich emotional representation with sufficient clarity to permit simple classification techniques.

The novelty of this work is its description of how emotion distillation can be used to create simple light-weight classification frameworks and interpretable representations. Our

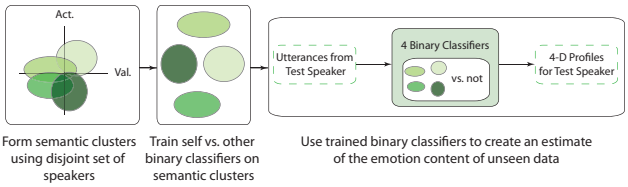

Fig. 2. Training paradigm used in EP creation.

results suggest that this method is effective for classifying the affective label of utterances across all classification methods employed. We demonstrate that given emotionally distilled information we can use simple classification techniques to achieve state-of-the-art performance (Metallinou et al., 2010, 62.42 $\pm$ 3.16 [12]), which used HMMs across multiple modalities and fused with decision-level Bayesian fusion.

II. Description of Data

In this paper we use the audio-visual+motion capture IEMOCAP database collected at the University of Southern California [13]. This study is restricted to the audio and motion-capture data. These data are dyadic emotional interactions between actors elicited using emotionally evocative scripts and improvisational scenarios and contain a wide range of emotions. Six evaluators labeled the categorical emotional labels of the data from the set of: angry, happy, neutral, sad, frustrated, excited (later merged with happy), surprised, disgusted, fear, and other. Approximately 17% of the data has no majority-agreed ground-truth label, highlighting the variability and non-prototypicality of the data. We use data from the emotion classes of: angry, happy, neutral, and sad. There are a total of 661, 1,155, 515, 572 angry, happy, neutral, and sad utterances, respectively (2,904 in total).

The audio features are extracted using Praat [14]. These features include pitch, intensity, and Mel Filterbank outputs (MFB), commonly used features in emotion recognition [15]. The motion-capture features are an adaptation of facial animation parameters (FAP) (discussed in [16]) to the IEMOCAP database. The features are grouped by facial region and include mouth, cheek, eyebrow, and forehead features.

The initial feature set includes statistical functionals calculated over the audio and motion capture signals. The functionals are calculated over windows of 0.25 seconds (with a half window length overlap) and include: mean, variance, upper and lower quantiles, and quantile range (685 features). The feature set size was reduced using Principal Feature Analysis (PFA) [17], also employed in [11]. PFA identifies a subset of features in the original feature space that explain a majority of the variance in the data. This results in a set of features that are less correlated than the initial set, explain the variance, and are interpretable. We retained 190 features as in [11].

III. Methods

A. Emotion Distillation via Emotion Profiles (EPs)

Emotion Profiles quantify the affective content of an utterance in terms of four components defined by the human-centered basic emotion labels: angry, happy, neutral, and sad. Each component describes the estimated level of the emotion class in the analyzed data, providing a quantitative description of the affective content. This description also provides human-interpretable information, important when this system is used in the context of a human-computer interaction.

EPs quantify the presence of emotion cues within an utterance using classifier-derived confidence, proven effective in our previous work [10], [11]. EPs are trained using four binary subject-independent self vs. other Support Vector Machine (SVM) classifiers over the classes of angry, happy, neutral, and sad (e.g., angry vs. not angry). The output of each classifier is a class membership ($\pm$ 1) and a distance from the hyperplane, which we use to approximate the confidence of the assertion of class membership (Figure 2). We use the same training paradigm as in [11], with Radial Basis Function kernels and sequential minimal optimization (SMO) [18], implemented using Matlab's svmclassify function. The testing data are windowed portions of the utterances. The training data are utterance-level statistics to decrease run-time. SVM training is between $O(n^2)$ and $O(n^3)$ depending on data noise. In our dataset, a 0.25 second window generates over 80,000 windowed EPs, making training computationally complex. However, this training complexity does not extend to testing. Consequently, assuming the availability of offline training, systems employing these techniques could potentially be employed on lightweight platforms.

EPs can be estimated over different window sizes to estimate the sub-utterance affective content. We previously extracted EPs over sliding windows of 0.25, 0.5, 1, 1.5, and 2 seconds to determine the effect of window size on classification performance [11]. In this paper, we focus only on a window size of 0.25 seconds. This permits the dynamic analysis over all utterances greater than 0.5 seconds (the HMM is a three-state model), eliminating the need to analyze results by utterance length and makes our work more directly comparable to published findings.

We formally define two terms: an **EP slice** and an **emotogram**. An EP slice is a EP over a windowed portion of an utterance. An EP slice for utterance i at time t is designated as $EP_{i,t}$ (see vertical slices in Figure 3, left). Each four-dimensional EP slice defines the confidence, c, in the presence or absence of each emotion: $\mathbf{EP_{i,t}} = [\mathbf{c_{t,angry}}, \mathbf{c_{t,happy}}, \mathbf{c_{t,neutral}}, \mathbf{c_{t,sad}}]$. An emotogram is the set of EP slices extracted over utterance i, $\mathbf{emot_i} = \{[\mathbf{c_{t,ang}}, \mathbf{c_{t,hap}}, \mathbf{c_{t,neu}}, \mathbf{c_{t,sad}}]\}_{\mathbf{t=0}}^{\mathbf{t=T_N}}$, T_N is the number of EP slices. The testing EPs are generated using leave-one-subject-out cross-validation; the training emotograms were generated using leave-one-training-speaker-out cross-validation.

In the remainder of this paper we will describe four methods for classifying the affective content of utterances using the information described by the emotogram via HMMs and the simple techniques of N-Gram modeling, LDA, and simple summation. All experiments are performed using leave-one-speaker-out cross-validation. The results are presented in terms of unweighted accuracy (UW) averaged over the 10 speakers as in [12].

Fig. 3. Transformation from original emotogram (left) to quantized emotogram (center) and quantized EP slice (right).

B. Dynamic Methods: HMM and N-Gram Modeling

We first demonstrate that the emotogram labels can be classified by modeling the trajectory of the unquantized EP slices, a strategy introduced in [11]. We model the emotograms using a left-to-right three-state two-mixture description using HTK, the Hidden Markov Model Toolkit [19]. This defines an emotion in terms of a start, middle, and end state. The input features to the HMM modeling are the normalized continuous-valued EP slices of the emotograms. We will compare these results to the simplified classification results.

N-gram modeling uses discrete units obtained by quantizing the emotograms' EP slices using the 0.8 and 0.2 quantiles (determined empirically) for each component: angry, happy, neutral, sad over all training emotograms. This results in a four-dimensional discrete-valued representation (Figure 3). The binning values are defined over utterances with the ground truth of the component (e.g., the angry component is binned using angry utterances). These bins have designations of "high", "medium", and "low" for each component. Quantiles are used to account for the varied and skewed distribution of profile values over each component. We code the representation, resulting in 81 one-dimensional **EP codes** (39 of which occur in the data). The data are now described as D_t, one EP code from time slice, t. The coded emotogram is: $\{D_t\}_{t=0}^{T_N}$. We calculate unigram, bigram, and trigram probabilities over the EP codes for each emotogram.

We posit that the distillation captures emotional structure. We seek to identify the relationship between this structure and the label of each test utterance's emotogram by treating the distillation as an observable Markov chain. We focus on the evolution of emotion within the utterance (Equation 1). We assume conditional independence and a uniform prior to accommodate training-testing distribution mismatch. Thus, for all $m \in \{$angry, happy, neutral, sad$\}$, the prior, $p(m)$, is 0.25 and is not included any of the equations.

Our preliminary formulation assumes that emotion perception is cumulative and that humans attend to all emotional evidence when making an assessment (Equations 1 and 2). However, psychological research has suggested that people attend to salient information when making these assessments [20]. Consequently, we propose an alternate formulation that accumulates the probabilities as a sum, rather than a product (Equation (3)), emphasizing evidence that is a strong indication of a given class (an approximation of salience). The final label reflects the presence of emotionally salient data while mitigating both noise and emotionally inconsequential information (which will have little effect on the final sum).

$$m^* = \underset{m}{argmax}\ P(D|m) = \underset{m}{argmax}\ P(D_{t_n}, ..., D_1|m) \quad (1)$$

$$= \underset{m}{argmax} \sum_{i=2}^{n} log(P(D_t|D_{t-1}, m)) + log(P(D_1|m)) \quad (2)$$

$$m^* = \underset{m}{argmax} \sum_{i=2}^{n} P(D_t|D_{t-1}, m) + P(D_1|m) \quad (3)$$

We model the emotion flow using unigram, bigram, and trigram models. The bigram and trigram models use back-off interpolation using parameters from largest n-gram to smallest n-gram: [0.8, 0.2] and [0.5, 0.3, 0.2], respectively, determined empirically. We also used used add-less-than-one smoothing to account for unseen data. We use a voting mechanism to arrive at a final estimate of emotion content. We sum the probabilities derived from the estimates of each n-gram model. The maximal probability is chosen as the final class estimate.

C. Static Methods: Linear Discriminant Analysis and Simple Summation

In this section we will look at two simple static modeling techniques. We must alter our treatment of the emotogram. Instead of modeling the dynamics of the profile, we treat the emotogram as a set of emotional evidence that can be accumulated over the course of that utterance. We start with the original emotogram and define $sumEP$, the accumulation, and $normEP$, the accumulation normalized by emotogram length, T_N.

We define $sumEP_i$ as the summation of the emotogram ($emot_i$, defined in Section III-A) over all time steps, $sumEP_i = \sum_{t=0}^{t=T_N} [c_{t,ang}, c_{t,hap}, c_{t,neu}, c_{t,sad}]$. We define sums of confidences over each emotion component as $C_m, m \in angry, happy, neutral, sad$ and redefine $sumEP_i$ as: $sumEP_i = [C_{angry},\ C_{happy},\ C_{neutral},\ C_{sad}]$. We normalize $sumEP$ by the length of the utterance, T_N, to create $normEP$. The input to both the LDA and simple summation classification is $normEP$.

Linear Discriminant Analysis (LDA) identifies Bayes optimal decision boundaries to separate points assigned to w_i from those not assigned to w_i. The LDA decision boundary separates based on differences between the weight vectors, rather than the value of the weight vectors [21]. This formulation is ideal for the aggregated emotogram representation. The $normEP$ representation describes the accumulated evidence of the relative flow of each emotion component. Thus, classification that focuses on the differences in the component definitions of $normEP$ will be well positioned to classify unlabeled test utterances. The summation classification assigns an utterance to the class with the highest value of $normEP$.

IV. Results and Discussion

The results demonstrate that there are no statistically significant differences between any of the overall accuracies of the presented methods (Table I). The major differences in the accuracies across techniques presented in this paper

Type	Angry	Happy	Neutral	Sad	Unweighted
HMM	74.61	72.57	45.90	65.59	64.67 $\pm$ 5.85
1-gram	76.82	76.82	28.70	76.27	64.67 $\pm$ 5.59
2-gram	70.50	79.72	32.18	75.86	64.57 $\pm$ 5.91
3-gram	71.92	78.12	30.78	77.51	64.58 $\pm$ 6.27
vote	73.55	78.56	29.31	77.52	64.74 $\pm$ 6.12
LDA	76.78	72.28	41.93	70.61	65.40 $\pm$ 5.61
sum	76.91	71.85	39.55	69.15	64.36 $\pm$ 6.56

TABLE I
ACCURACIES OF HMM AND N-GRAM MODELING TECHNIQUES (NOTE: THE N-GRAMS RESULTS ARE OBTAINED WITH INTERPOLATION).

are the per-class accuracies. HMM modeling results in the highest neutral class accuracies while the n-gram techniques had the highest sadness and happiness accuracies. The emotion distillation classification results are comparable to those of the more computationally complex state-of-the-art, 62.42 $\pm$ 3.16 [12]. The two systems rely on slightly different subsets of the IEMOCAP dataset, which makes it difficult to assess the specific performance differences.

The n-gram results demonstrate that the four techniques, unigram, bigram, trigram, and voting, perform comparably. This first suggests that the information contained in individual EP slices (modeled via unigrams) is expressed differently across the four emotion classes in an utterance. The parity between the unigram and the bigram/trigram results suggests that the structure predicted by the emotogram representation are not noise, otherwise the bigram/trigram results would be lower than those of the unigram results. However, the lack of improvement when considering the dynamics of the emotogram changes suggests that there exist additional opportunities to more effectively model the emotional structure of emotograms. Future work will explore methods to better extract this intra-utterance emotional structure.

V. CONCLUSIONS

In this paper we present results demonstrating that emotion distillation augmented with simple classification performs comparably to distillation when augmented with Hidden Markov Models. We show that all techniques perform comparably to the state-of-the art results on this database, 62.42 $\pm$ 3.16 [12]. We highlight that the maximal accuracies of the four methods (HMM, n-gram, LDA, and simple summation) do not differ significantly. This suggests that emotion fluctuation can be modeled and interpreted in multiple contexts. Further, the n-gram experiments suggest that emotion can be modeled by implicitly focusing only on regions of the utterance that are highly representative of a certain emotion class.

The presented success of emotion distillation suggests that this process can be used to simplify the design of classification algorithms for emotion recognition. In our future work we will explore methods to assess the nature of intra-utterance patterns. We will investigate simplified intra-utterance versions of the inter-utterance context sensitive emotion computing proposed in [22]. We also seek to employ distillation frameworks in the context of mobile interface design and behavioral informatics computation. We see emotion distillation as both a method to simplify computation and as a technique to understand more about the complex nature of emotion expression.

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*, pp. 32–80, Jan. 2001.

[2] Klaus R. Scherer, "Vocal communication of emotion: A review of research paradigms," *Speech Communication*, vol. 40, no. 1-2, pp. 227 – 256, 2003.

[3] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social signal processing: Survey of an emerging domain," *Image and Vision Computing*, vol. 27, no. 12, pp. 1743–1759, 2009.

[4] R. El Kaliouby and P. Robinson, "The emotional hearing aid: an assistive tool for children with asperger syndrome," *Universal Access in the Information Society*, vol. 4, no. 2, pp. 121–134, 2005.

[5] M. Madsen, R. El Kaliouby, M. Goodwin, and R. Picard, "Technology for just-in-time in-situ learning of facial affect for persons diagnosed with an autism spectrum disorder," in *ACM SIGACCESS conference on Computers and accessibility*. ACM, 2008, pp. 19–26.

[6] M.P. Black, P. Georgiou, A. Katsamanis, B. Baucom, and S. Narayanan, ""you made me do it": Classification of blame in married couples' interaction by fusing automatically derived speech and language information," in *Interspeech*, Florence, Italy, Aug. 2011.

[7] J. Gibson, A. Katsamanis, M.P. Black, and S. Narayanan, "Automatic identification of salient acoustic instances in couples' behavioral interactions using diverse density support vector machines," in *Interspeech*, Florence, Italy, Aug. 2011.

[8] B. Settles, "Closing the loop: Fast, interactive semi-supervised annotation with queries on features and instances," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2011, pp. 1467–1478.

[9] G. Druck and A. McCallum, "Toward interactive training and evaluation," in *ACM: Information and Knowledge Management*, 2011, pp. 947–956.

[10] E. Mower, M. Matarić, and S. Narayanan, "A framework for automatic human emotion classification using emotion profiles," *IEEE Trans. on Audio, Speech, and Language Processing*, pp. 1057–1070, 2011.

[11] E. Mower and S. S. Narayanan, "A hierarchical static-dynamic framework for emotion classification," in *In Proceedings of IEEE International Conference on Audio, Speech and Signal Processing (ICASSP)*, May 2011.

[12] A. Metallinou, S. Lee, and S. Narayanan, "Decision level combination of multiple modalities for recognition and analysis of emotional expression," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*. IEEE, 2010, pp. 2462–2465.

[13] C. Busso, M. Bulut, C.C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J.N. Chang, S. Lee, and S.S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Journal of Language Resources and Evaluation*, pp. 335–359, Nov. 5 2008.

[14] P. P. G. Boersma, "Praat, a system for doing phonetics by computer," *Glot international*, vol. 5, no. 9/10, pp. 341–345, 2002.

[15] F. Eyben, M. Wöllmer, and B. Schuller, "openEAR - introducing the munich open-source emotion and affect recognition toolkit," in *ACII*, Amsterdam, The Netherlands, Sept. 2009.

[16] N. Tsapatsoulis, A. Raouzaiou, S. Kollias, R. Cowie, and E. Douglas-Cowie, "Emotion Recognition and Synthesis Based on MPEG-4 FAP's," in *MPEG-4 Facial Animation: The Standard, Implementation, and Applications*, I. S. Pandzic and R. Forchheimer, Eds., chapter 9, pp. 141–167. John Wiley & Sons, Ltd., 2002.

[17] Y. Lu, I. Cohen, X. S. Zhou, and Q. Tian, "Feature selection using principal feature analysis," in *Int. Conf. on Multimedia*, New York, NY, USA, 2007, pp. 301–304.

[18] J. C. Platt, "Fast training of support vector machines using sequential minimal optimization," in *Advances in Kernel Methods*. MIT press, 1999, pp. 185–208.

[19] S. Young, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK book*, Entropic Cambridge Research Laboratory Cambridge, Eng., 1997.

[20] M. L. Phillips, W. C. Drevets, S. L. Rauch, and R. Lane, "Neurobiology of emotion perception i: the neural basis of normal emotion perception," *Biological Psychiatry*, vol. 54, no. 5, pp. 504 – 514, 2003.

[21] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, Wiley-Interscience Publication, 2000.

[22] A. Metallinou, M. Wollmer, A. Katsamanis, F. Eyben, B. Schuller, and S. Narayanan, "Context-sensitive learning for enhanced audiovisual emotion classification," *IEEE Trans. on Affective Comp.*, , no. 99, 2011.