INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Retrieving Categorical Emotions using a Probabilistic Framework to Define Preference Learning Samples

Reza Lotfian and Carlos Busso

Multimodal Signal Processing (MSP) Laboratory, Department of Electrical Engineering
The University of Texas at Dallas, Richardson TX 75080, USA
reza.lotfian@utdallas.edu, busso@utdallas.edu

Abstract

Preference learning is an appealing approach for affective recognition. Instead of predicting the underlying emotional class of a sample, this framework relies on pairwise comparisons to rank-order the testing data according to an emotional dimension. This framework is relevant not only for continuous attributes such as arousal or valence, but also for categorical classes (e.g., is this sample happier than the other?). A preference learning system for categorical classes can have applications in several domains including retrieving emotional behaviors conveying a target emotion, and defining the emotional intensity associated with a given class. One important challenge to build such a system is to define relative labels defining the preference between training samples. Instead of building these labels from scratch, we propose a probabilistic framework that creates relative labels from existing categorical annotations. The approach considers individual assessments instead of consensus labels, creating a metrics that is sensitive to the underlying ambiguity of emotional classes. The proposed metric quantifies the likelihood that a sample belong to a target emotion. We build *happy*, *angry* and *sad* rank-classifiers using this metric. We evaluate the approach over cross-corpus experiments, showing improved performance over binary classifiers and rank-based classifiers trained with consensus labels.

Index Terms: emotion recognition, preference learning, information retrieval, basic emotions

1. Introduction

Emotions play an important role in human interaction. Recognizing basic emotions such as happiness, anger or sadness can have an important role in many practical applications. For example, an affect aware game can adjust the level of difficulty by tracking the emotion of the player, making the game more enjoyable [1,2]. Algorithms with the capability of detecting emotions can serve as instrumental tools for healthcare to facilitate the diagnosis and prognosis of many mental health conditions including schizophrenia [3], depression [4], and autism [5]. They can also serve as instrument to detect the cognitive learning state of a student such as frustration or uncertainty [6,7].

Current studies address emotion recognition of categorical emotions as a multi-class problem, where all the samples labeled with a given emotional class are assumed to be similar. However, emotional classes have different intensities showing clear inter-class variability (e.g., 'cold' versus 'hot' anger). We have consistently annotated our emotional corpora with continuous attributes such as valence (negative versus positive) and arousal (calm versus active), in addition to categorical labels [8,9]. This approach has allowed us to explore the relationship between categorical labels and attribute-based annotations [10], showing clear emotional differences within a given emotional class. Binary and multi-class classifiers cannot distinguish between samples associated with the same class having different intensity, which is a major barrier for the aforementioned applications. Instead, preference learning can offer a principled framework to evaluate emotional samples within a given emotional class (e.g., is sample one happier than sample two?). While preference learning algorithms designed to retrieve emotional content have been used in many retrieval applications in music [11, 12], text [13, 14] , image [15, 16], and video [17], their use in emotion recognition from speech has being limited to few studies [18–20].

A key challenge in preference learning is to derive reliable labels describing preference between pairwise samples. Direct annotation of these relative labels is expensive and time demanding as the number of pairwise comparison is $N(N-1)/2$ for N samples. When we use preference leaning for continuous attributes such as arousal or valence, these relative labels can be created by comparing the average scores associated with two samples [18,20]. We can control the reliability of these labels by imposing a minimum margin required to establish preferences (e.g., the difference of their scores needs to be higher than a margin to establish that one sample is preferred to another). While this approach is ideal for dimensional attributes, the same approach face many practical obstacles in dealing with basic categorical emotions. How can we derive similar framework for relative labels associated with categorical emotions?

This paper proposes a novel method to derive relative labels for preference learning using categorical emotions. We introduce a probabilistic framework to assign a relevance score to samples in our corpus for each categorical emotion. This relevance score quantifies the level of the target emotion conveyed in the samples. Assuming that each sample in the corpus was evaluated by R separate evaluators using categorical labels, we consider the R individual ratings instead of just the consensus labels. The relevance score considers the inherent confusion made by raters between emotional classes. Samples that are consistently evaluated with the target emotion receive higher scores than samples with conflicting evaluations. We build rank-based classifiers for categorical emotions based on pairs that are reliably separated according to this relevance score. We compare our proposed method to a rank-based approach which relies on consensus labels [19]. We also compare the proposed preference learning algorithm with binary classifiers adapted for retrieval problems. The cross-corpus evaluation is conducted by training the models on the IEMOCAP corpus [8] and testing the them on the MSP-IMPROV database [9], showing the benefits of the proposed approach.

This work was funded by NSF (IIS-1217104 and IIS-1453781).

http://dx.doi.org/10.21437/Interspeech.2016-1052

2. Related Work

Building relative pairwise labels for categorical emotions is not easy. Current emotional annotations are typically collected using questionnaires, where the evaluator is asked to select the most appropriate emotional class from a list of emotions. These labels are later aggregated to create consensus labels using rules such as majority vote. These labels are noisy, usually with low inter-evaluator agreement [21]. These labels do not directly indicate the intensity level within an emotional class, so creating relative preference labels is not straightforward. One solution is to collect new evaluations from scratch, where the raters use a Likert-like scales to rank the intensity for each of the target emotions. Another solution is to use continuous-time labels for discrete emotions. For example, the evaluations of the Semaine database includes discrete emotions such as happiness, sadness, and disgust as optional traces [22]. The evaluators assessed the intensity of the target emotion by moving the mouse cursor over a *user graphical interface* (GUI) using FEELTRACE. These annotations can be used to derive relative labels similar to the ones derived for dimensional attributes such as arousal and valence [18, 20].

Instead of collecting new evaluations from scratch, it is appealing to derive relative labels from existing categorical emotional annotations. To the best of our knowledge, there is only one study that have proposed a method for this purpose. Cao et al. [19] studied rank-based classifiers for emotions by using the consensus labels assigned to the samples. They created the relative labels by assigning preference for pairwise comparisons between a sample belonging to the target emotion, and another belonging to a different emotion. The sample from the target emotion was preferred over the sample that does not belong to the target emotions. Our proposed methods uses individual assessments instead of consensus labels, creating a continuous relevance score to reliably form the pairwise preference.

3. Databases and Features

The study uses cross-corpus evaluation relying on the *interactive emotional dyadic motion capture database* (IEMOCAP) database [8] and the MSP-IMPROV corpus [9]. This section also describes the acoustic features.

3.1. IEMOCAP

The IEMOCAP database [8] was collected to study expressive human behaviors. Five dyadic sessions were recorded using 10 trained actors. Two elicitation schemes were used based on scripts and spontaneous improvisations targeting the emotional categories sadness, happiness, anger and frustration. Other emotions were also elicited as dictated by the course of the conversation between the actors. Although the database is collected from actors, the elicitation techniques rooted in well-established theories and methods of theater provide emotional manifestations closer to natural iterations [23]. The corpus contains approximately twelve hours of recordings, which were manually segmented, transcribed and emotionally annotated with categorical (three evaluators) and attribute-based labels (two evaluators) at the turn level. This study considers turns in which three independent evaluators reached majority vote agreement. Turns with overlapped speech were excluded from the experiments. Table 1 shows the number of samples per emotion using majority vote rule. Further information about the database is provided in Busso et al. [8].

Table 1: Number of utterances from each emotion category based on majority voting: Anger(A), Happiness(H), Sadness(S), Neutral(N), Other(O) and No Agreement(NA).

Database	A	H	S	N	O	NA	Total
IEMOCAP	289	284	608	1099	1704	800	4784
MSP-IMPROV	792	2644	885	3477	85	555	8438

3.2. MSP-IMPROV

The MSP-IMPROV database [9] is an audiovisual corpus collected to explore emotional behaviors during dyadic improvisations. The scenarios were designed to promote the naturalness of the recordings, while keeping control over the lexical content. This goal was achieved by defining target sentences with emotional-dependent contexts, one for each of the target emotional classes: anger, sadness, and happiness, plus neutral state. The database includes the target sentences, all the improvisation turns that lead the actors to utter the target sentences, and all the interactions during breaks. This corpus consists of 8,438 turns (over 9 hours) of emotional sentences. The emotional labels were collected through perceptual evaluations using crowdsourcing [24], receiving at least five evaluations. Table 1 shows the number of samples per emotion, where the consensus labels are assigned using majority vote. Further description of the corpus is given in Busso et al. [9].

3.3. Acoustic Features

This study uses the feature set provided for the speaker state challenge at INTERSPEECH 2013. This set includes 6,308 *high level descriptors* (HLDs) extracted using OpenSMILE [25]. The set includes prosodic, spectral and voice quality features, from which we estimate functionals at the turn level such as minimum, maximum, and range. We apply speaker dependent feature normalization, where we subtract the means of the functionals, dividing by their standard deviation (i.e., z-normalization). Further description of the features is given in Schuller et al. [26]

4. Methodology

The training data and the labels assigned to them play a key role in building reliable emotional classifiers. We are interested in building rank-based classifiers for categorical labels. The key challenge is to define relative labels with pairwise preference. For example, for a happy ranker, we would like to establish whether a sample is happier than another. These relative labels should be derived from existing annotations. We propose a probabilistic method to estimate the intensity of each target emotion for every sample.

The proposed method is to define a relevance score which is created by considering all the individual evaluations assigned to the data. In the IEMOCAP database, each turn received three primary labels from the list *Happiness* (H), *Excited* (E), *Surprised* (Su), *Fear* (F), *Anger* (A), *Frustration* (Fr), *Disgust* (D), *Sadness* (S), *Neutral* (N) and *Other* (O). The intuition behind our method is that samples that are consistently evaluated with the target emotion are more likely to convey the target emotion. For example, we can assume a sample with labels (H,H,H) is *happier* than a sample with labels (H,S,H) (i.e., (H,H,H) $\succ_H$ (H,S,H)). Furthermore, emotional categories are not orthogonal. For example, excitement and happiness are closely related. Therefore, a samples with labels (H,H,E) is expected to be *happier* than a sample with labels (H,H,S) (i.e., (H,H,E) $\succ_H$

(H,S,H)). We provide a probabilistic framework to generalize these intuitive ideas.

4.1. Probabilistic Framework to Create Relevance Score

To rank samples based on primary emotions, we can estimate the posteriori probability of a sample to belong to an emotional class, given the individual annotations assigned to the sample. For example $P(H|H,N,E)$ is the probability that the true emotion is *happiness*, given that the annotators chose the labels *happiness*, *neutral* and *excited* during the subjective evaluation. In practice, the true label for the categorical emotion is unknown. Therefore, we approximate the true emotions using the majority vote rule. We assume that the annotations are conditionally independent, given the true emotion. This is also an approximation since the true emotions derived from majority vote are clearly dependent on the individual evaluations. However, these assumptions greatly simplify the formulation of the problem, providing reasonable results, as discussed in the experimental evaluation. We denote the true emotion with w_k, and the individuals annotation with $x_1, x_2, \ldots, x_R$, where R is the number of annotators. We can estimate the posteriori probability with Equation 1.

$$P(w_k|x_1,x_2,\ldots,x_R) = \frac{P(w_k)\prod_{i=1}^{R} p(x_i|w_k)}{p(x_1,x_2,\ldots,x_R)} \qquad (1)$$

We use a simplification of this expression that is commonly used in the context of combining different classifies. When the goal is to find the class w_j with $j \in \{1,\ldots,E\}$ that maximizes Equation 1, we can use the product rule (Eq. 2) or the sum rule (Eq. 3) [27]:

$$\max_{k=1}^{E} \quad P^{-(R-1)}(w_k)\prod_{i=1}^{R} P(w_k|x_i) \qquad (2)$$

$$\max_{k=1}^{E} \quad (1-R)P(w_k) + \sum_{i=1}^{R} P(w_k|x_i) \qquad (3)$$

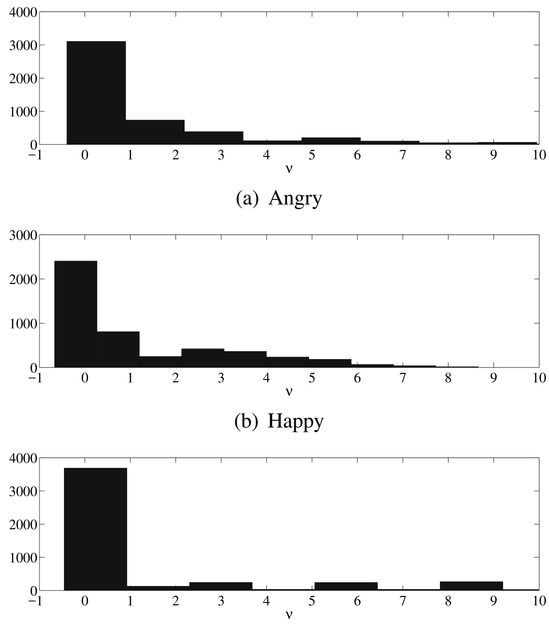
These expressions depend on the prior probability of the target emotions $P(w_k)$, and the posteriori probability $P(w_k|x_i)$, which can be easily estimated from the perceptual evaluations. In spite of imposing stronger assumptions, Kittler et al. [27] indicated that the sum rule generally performs better since it is robust against zero-value posteriori probabilities. While this expression is only related to the posteriori probability in Equation 1, we define our relevance score as

$$\nu_j = (1-R)P(w_j) + \sum_{i=1}^{R} P(w_j|x_i) \qquad (4)$$

We calculate this relevance score for sadness, happiness and anger. Figure 1 shows the distribution where high values indicate higher intensity of the target emotion. The figure shows a clear separation for samples with low and high relevance scores for happiness and anger. For sadness, the relevance scores are mostly determined by the number of evaluators perceiving the samples as sadness. This result is mainly due to the low correlation between sadness and other emotions considered in the IEMOCAP database.

4.2. Preference Learning Algorithms

We use the relevance score to train preference learning algorithms using the IEMOCAP database. We select pairs of samples where the difference of their relevance scores are larger



(a) Angry

(b) Happy

(c) Sad

Figure 1: Distributions of relevance scores, defined on Equation 4, for the target emotions.

than a minimum margin, reducing the ambiguity on the training set. This margin is set by maximizing precision in retrieving the number of samples for the target emotion using two-fold cross-validation over the training set. We consider two rank classifiers: preference learning with *Gaussian process* (GP) using the implementation provided by Chu and Ghahramani [28] and Rank-SVM [29] using the toolkit presented in Joachims [30]. Since the Gaussian process preference learning method has a better training performance on smaller training sets, we divide the training set into four subsets and create rank classifiers for each. The resulting rankings are then combined to minimize the conflict between the corresponding lists. We create rank-based classifiers for sadness, happiness and anger.

4.3. Feature selection

Given the high dimension of the original feature set (e.g., 6,308), we use a two-level feature selection approach to select the most relevant features for each emotion. First, we remove less-informative features according to information gain criterion [31], reducing the number of feature to 500. The information gain is implemented per emotion with binary labels: target emotion versus others. Second, we select 100 features from this set by maximizing the precision in retrieving the top 10% of the number of sentences from the target emotion using *floating forward feature selection* (FFFS) [32]. This framework is separately implemented for Rank SVM, and Gaussian process preference learning over the IEMOCAP corpus. We implement this approach using two fold cross-validation over the training set (IEMOCAP corpus), averaging the precision associated with the selected features.

5. Experimental Evaluation

The evaluation of the proposed approaches is conducted on the MSP-IMPROV database, where the ground truth labels are assigned based on majority vote rule. The goal of the emotion

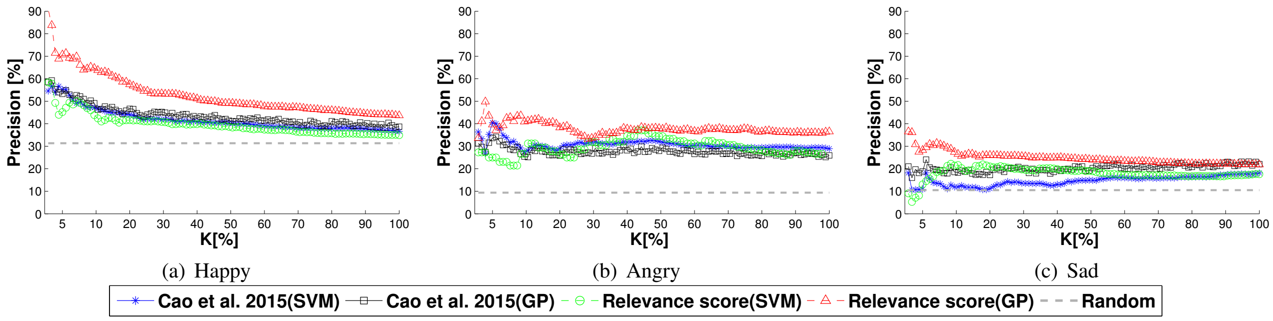


Figure 2: Precision at K for different rankers of categorical emotions. *P@K* is the precision after retrieving K% of the sample labeled with the target emotion. The continuous dashed lines represent performance at random.

Table 2: Comparison of Kendall rank correlation coefficient for Gaussian process preference learning.

Emotion category	Cao et al. [2015]	Relevance scores
Angry	0.118	0.126
Happy	0.194	0.242
Sad	0.158	0.194

Table 3: Comparison of proposed approach at *P@30* and *P@100*, with binary SVM and GP rankers using labels estimated with the approach in Cao et al. [2015].

Emotion	Cao et al. [2015]		Relevance scores		Binary SVM	
	P@30	P@100	P@30	P@100	P@30	P@100
Angry	31.8	28.9	36.5	33.6*	30.4	29.1
Happy	42.4	37.1	53.4*	43.7	45.0	42.8
Sad	20.1	22.0	25.6	21.9	16.7	15.2

* approach outperforms alternatives asserting significance at p = 0.05.

rankers is to retrieve the target emotions from the entire corpus. The performance of preference learning methods are assessed with the precision at K (*P@K*). In our implementation, this metric provides the precision after retrieving K% of the sample labeled with the target emotion. For example, there are 2644 happy samples in the MSP-IMPROV corpus. Therefore, *P@10* is the precision achieved by the happy ranker when retrieving 264 samples. We compare the preference learning algorithms with the approach proposed by Cao et al. [19]. This method defines relative pairwise preference between samples, using the majority vote labels (Sec. 2).

Figure 2 shows the results for sadness, happiness and anger. The continuous dashed lines represent the performance achieved by randomly selecting samples, reflecting the proportion of the target emotion in the corpus. We implement Rank-SVM and GP method using labels defined by the relevance score and the approach proposed by Cao et al. [19]. The best performance is achieved by the GP framework using the labels derived from the relevance score (red line). Those labels tend to produce more accurate rankers. While not complete consistent across experiments, the GP framework usually achieves better performance than Rank-SVM. We use GP rankers for the rest of the evaluation.

Kendall rank correlation coefficient is a common approach to compare preference learning algorithms. This measure compared the overall ranking estimated by the retrieval system and the ground truth ranking. This ground truth ranking is obtained by estimating the relevance scores over the test samples, sorting them in descendent order. Given the noisy nature of the these labels, we may have negligible mistakes by erroneously ranking adjacent samples. Therefore, we downsample the testing set to 20 samples, ensuring that the emotional differences between adjacent samples are meaningful. Table 2 shows the Kendall rank correlation for GP rankers using the relevance scores and the approach proposed by Cao et al. [19]. Notice that the values reported in this study are higher than the ones reported in related studies using this metric [18]. We observe higher Kendall rank correlation for the proposed method for all the emotional rankers, demonstrating the benefits of the relevance scores.

Finally, we compare the performance with binary SVMs trained for each emotion (one emotion versus other emotions). To rank the samples using binary SVM, we sort the results according to the distance to the hyperplane. The preference learning algorithms are trained with GP rankers. We consider two conditions: *P@30* and *P@100*. Table 3 shows the results. The preference learning ranker trained with the relevance scores outperforms other methods. The only exception is *P@100* for sadness where the results are almost identical to the results achieved by the approach proposed by Cao et al. [19]. The asterisk in the table indicates that the performance for our method is significantly better than any other method for that condition (difference between two population proportion test asserting significance at p=0.05).

6. Conclusion and future wok

This study proposed a probabilistic framework to derive relative labels, establishing preference between pairwise samples for a target emotion. These labels are created from existing annotations of categorical emotional classes, by defining a relevance score that considers individual annotations assigned to samples. We used these labels to train preference learning algorithms, creating emotional rankers for sadness, happiness, and anger. The experimental evaluation demonstrated that the precision rate in retrieving target categorical emotions are higher than the ones achieved with an alternative method using consensus labels, or with standard binary classifiers. The proposed approach also outperforms other methods in terms of Kendall rank correlation coefficient.

Future work includes modeling the evaluators' reliability, which is a relevant variable when the emotional annotations are collected with crowdsourcing services. We are also considering more sophisticated approaches to define relative margins between pairs of samples in the training set to ensure a minimum confidence level. We are also going to implement our preference learning classifiers using suitable *deep neural networks* (DNNs) for this task.

7. References

[1] M. Szwoch and W. Szwoch, "Emotion recognition for affect aware video games," in *Image Processing & Communications Challenges 6*, R. Choraś, Ed. Bydgoszcz, Poland: Springer International Publishing, September 2014, vol. 313, pp. 227–236.

[2] M. Obaid, C. Han, and M. Billinghurst, ""feed the fish": an affect-aware game," in *Australasian Conference on Interactive Entertainment*, Brisbane, Australia, December 2008.

[3] C. G. Kohler, W. Bilker, M. Hagendoorn, R. E. Gur, and R. C. Gur, "Emotion recognition deficit in schizophrenia: association with symptomatology and cognition," *Biological Psychiatry*, vol. 48, no. 2, pp. 127–136, July 2000.

[4] S. Langenecker, L. Bieliauskas, L. Rapport, J. Zubieta, E. Wilde, and S. S. Berent, "Face emotion perception and executive functioning deficits in depression," *Journal of Clinical and Experimental Neuropsychology*, vol. 27, no. 3, pp. 320–333, April 2005.

[5] C. A. Mazefsky and D. P. Oswald, "Emotion perception in Asperger's syndrome and high-functioning autism: the importance of diagnostic criteria and cue intensity," *Journal of Autism and Developmental Disorders*, vol. 37, no. 6, pp. 1086–1095, July 2007.

[6] W. Burleson and R. Picard, "Affective agents: sustaining motivation to learn through failure and a state of "stuck"," in *Social and Emotional Intelligence in Learning Environments Workshop In conjunction with the 7th International Conference on Intelligent Tutoring Systems (ITS 2004)*, Maceiò, Brazil, August - September 2004.

[7] S. D'Mello, S. Craig, B. Gholson, S. Franklin, R. Picard, and A. Graesser, "Integrating affect sensors in an intelligent tutoring system," in *Affective Interactions: The Computer in the Affective Loop Workshop at 2005 International Conference on Intelligent User Interfaces*, San Diego, CA, USA, January 2005, pp. 7–13.

[8] C. Busso, M. Bulut, C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. Chang, S. Lee, and S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Journal of Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, December 2008.

[9] C. Busso, S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, and E. Mower Provost, "MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception," *IEEE Transactions on Affective Computing*, vol. To appear, 2015.

[10] C. Busso and S. Narayanan, "Scripted dialogs versus improvisation: Lessons learned about emotional elicitation techniques from the IEMOCAP database," in *Interspeech 2008 - Eurospeech*, Brisbane, Australia, September 2008, pp. 1670–1673.

[11] K. Trohidis, G. Tsoumakas, G. Kalliris, and I. P. Vlahavas, "Multi-label classification of music into emotions," in *International Conference on Music Information Retrieval (ISMIR 2008)*, Philadelphia, PA, USA, September 2008, pp. 325–330.

[12] Y.-H. Yang, Y.-C. Lin, Y.-F. Su, and H. Chen, "A regression approach to music emotion recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 2, pp. 448–457, February 2008.

[13] T. Danisman and A. Alpkocak, "Feeler: Emotion classification of text using vector space model," in *AISB 2008 Convention Communication, Interaction and Social Intelligence*, vol. 2, Aberdeen, Scotland, April 2008, pp. 53–59.

[14] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and trends in information retrieval*, vol. 2, no. 1-2, pp. 1–135, January 2008.

[15] W. Wang and Q. He, "A survey on emotional semantic image retrieval," in *IEEE International Conference on Image Processing (ICIP 2008)*, San Diego, CA, USA, October 2008, pp. 117–120.

[16] S. Schmidt and W. G. Stock, "Collective indexing of emotions in images. A study in emotional information retrieval," *Journal of the American Society for Information Science and Technology*, vol. 60, no. 5, pp. 863–876, May 2009.

[17] J. Kierkels, M. Soleymani, and T. Pun, "Queries and tags in affect-based multimedia retrieval," in *IEEE International Conference on Multimedia and Expo (ICME 2009)*, Amsterdam, The Netherlands, June-July 2009, pp. 1436–1439.

[18] H. Martinez, G. Yannakakis, and J. Hallam, "Don't classify ratings of affect; rank them!" *IEEE Transactions on Affective Computing*, vol. 5, no. 2, pp. 314–326, July-September 2014.

[19] H. Cao, R. Verma, and A. Nenkova, "Speaker-sensitive emotion recognition via ranking: Studies on acted and spontaneous speech," *Computer Speech & Language*, vol. 29, no. 1, pp. 186–202, January 2014.

[20] R. Lotfian and C. Busso, "Practical considerations on the use of preference learning for ranking emotional speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016)*, Shanghai, China, March 2016, pp. 5205–5209.

[21] C. Busso, M. Bulut, and S. Narayanan, "Toward effective automatic recognition systems of emotion in speech," in *Social emotions in nature and artifact: emotions in human and human-computer interaction*, J. Gratch and S. Marsella, Eds. New York, NY, USA: Oxford University Press, November 2013, pp. 110–127.

[22] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schröder, "The SEMAINE database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 5–17, January-March 2012.

[23] C. Busso and S. Narayanan, "Recording audio-visual emotional databases from actors: a closer look," in *Second International Workshop on Emotion: Corpora for Research on Emotion and Affect, International conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco, May 2008, pp. 17–22.

[24] A. Burmania, S. Parthasarathy, and C. Busso, "Increasing the reliability of crowdsourcing evaluations using online quality assessment," *IEEE Transactions on Affective Computing*, vol. To Appear, 2015.

[25] F. Eyben, M. Wöllmer, and B. Schuller, "OpenSMILE: the Munich versatile and fast open-source audio feature extractor," in *ACM International conference on Multimedia (MM 2010)*, Florence, Italy, October 2010, pp. 1459–1462.

[26] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, and S. Kim, "The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Interspeech 2013*, Lyon, France, August 2013, pp. 148–152.

[27] J. Kittler, "Combining classifiers: A theoretical framework," *Pattern Analysis & Applications*, vol. 1, no. 1, pp. 18–27, March 1998.

[28] W. Chu and Z. Ghahramani, "Preference learning with Gaussian processes," in *International conference on machine learning (ICML 2005)*, Bonn, Germany, August 2005, pp. 137–144.

[29] R. Herbrich, T. Graepel, and K. Obermayer, "Support vector learning for ordinal regression," in *International Conference on Artificial Neural Networks (ICANN 1999)*, Edinburgh, UK, September 1999, pp. 97–102.

[30] T. Joachims, "Training linear SVMs in linear time," in *ACM SIGKDD international conference on Knowledge discovery and data mining*, Philadelphia, USA, August 2006, pp. 217–226.

[31] Y. Yang and J. Pedersen, "A comparative study on feature selection in text categorization," in *International Conference on Machine Learning (ICML 1997)*, Nashville, TN, USA, July 1997, pp. 412–420.

[32] P. Pudil, J. Novovičová, and J. Kittler, "Floating search methods in feature selection," *Pattern Recognition Letters*, vol. 15, no. 11, pp. 1119–1125, November 1999.