Active Learning by Label Uncertainty for Acoustic Emotion Recognition

Zixing Zhang[1], *Jun Deng*[1], *Erik Marchi*[1], *and Björn Schuller*[2,1]

[1] Machine Intelligence & Signal Processing Group, MMK,
Technische Universität München, Germany
[2]Institute for Sensor Systems, University of Passau, Germany
{zixing.zhang | jun.deng | erik.marchi}@tum.de, {Bjoern.Schuller}@uni-passau.de

10.21437/Interspeech.2013-250

Abstract

Speech data is in principle available in large amounts for the training of acoustic emotion recognisers. However, emotional labelling is usually not given and the distribution is heavily unbalanced, as most data is 'rather neutral' than truly 'emotional'. In the 'hay stack' of speech data, Active Learning automatically identifies the 'needles', i.e., the more informative instances to reduce human labelling effort when building a classifier, e.g., for acoustic emotion recognition. The critical issue thus is the determination and quantification of informativeness. To this end, we suggest to exploit the reliability of the usual ambiguity of emotional labels, i.e., we propose a novel approach based on label uncertainty. By building a certainty model and predicting the candidate instances, informativeness is thus based on labeller agreement. In addition, we consider class sparseness. The results of extensive test runs under well standardised conditions show the method's great potential in reducing labelling costs while boosting performance.

Index Terms: Active Learning, Label Uncertainty, Confidence Values, Class Sparseness

1. Introduction

For acoustic emotion recognition – as practically in any other pattern recognition task – we always consider ways to improve the robustness of a classifier: On the one hand, because of massive speech resources coloured by emotion existing in the real-world, we expect to annotate more instances and add more instances to build a better classifier, upon the idea of "there is no data like more data" in pattern recognition [1, 2, 3, 4]. On the other hand, we make an effort to control the total number of instances. A crucial problem is that this process is extremely time-consuming and costly. It is well-known that data collection, cleaning, and annotation consume about 80 percent of the effort in a typical data mining project [5]. In addition, it may avoid potential adversities of a larger amount of training instances [6], e.g., the training time will take long, or the training set will include much noisy data which are harmful to the classifier performance [7, 8]. Active learning (AL) seems to be an promising approach to minimize the amount of human supervision required and maximize the performance given transcribed and untranscribed data [9, 10].

Several AL approaches have been proposed and investigated in machine learning. A well known method is uncertainty-based AL in which the active learner determines the certainties of the predictions on the unlabelled data based on posterior probabilities. The samples with least certainty are generally presented to the labellers for annotation. This method is well established in automatic speech recognition [11] and information extraction [9], etc. Another common AL strategy is the committee-based method which utilises multiple classifiers and is investigated in [12, 13] for text categorisation. Predictions for unlabelled data are made by multiple classifiers. The samples considered as most informative are those with the lowest agreement. Other AL methods include the expected-error-reduction method [14], the expected-model-change-based method [15], the diversity-density-related method [16], etc.

However, those approaches mainly deal with objective pattern recognition tasks with certain 'ground truth', like automatic speech recognition [11], image retrieval [17], and vehicle recognition and tracking [18]. In tasks with subjective speech phenomena such as emotion, labels are determined by several labellers' personal judgement [19, 20, 21]. Because of the variety of personal perception, those labels have a large deviation. Several methods are recommended to alleviate the variation, e.g., labelling by multiple annotators, employing evaluator weighted estimator (EWE) [22], and filtering outliers [23]. Those labels ultimately form the 'gold standard' with an inherent label uncertainty.

To exploit this information which is reflected in the levels of human agreement, a novel AL approach is proposed by the usage of label uncertainty. By building an uncertainty model based on human agreement levels, we predict all the instances in the candidate pool. Then, we select the instances by either of two methods: 1) Based on class sparseness. That is, select and add the 'likely to be' sparse class instances to the training set. 2) Based on confidence value. The prediction of those instances among a range of labeller agreement levels will be chosen. Compared to our previous work [3], where methods of instance selection (sparse instance tracking and medium confidence score) are based on class predictions, both methods implemented in this paper are based upon predictions of human agreement levels.

In the following, we firstly introduce the chosen database in Section 2; then, we describe the details of our novel AL (Section 3); further, we evaluate our approaches on acoustic emotion recognition in Section 4; finally, in Section 5 we draw conclusions and point out some future work.

2. Database

To evaluate the effectiveness of our approaches, we select the frequently used, spontaneous emotion database FAU Aibo Emotion Corpus (AEC) [24], which is the official corpus of the INTERSPEECH 2009 Emotion Challenge (EC) [25]. It deals with recordings of children interacting with Sony's pet robot Aibo via German speech. The Wizard-of-Oz controlled Aibo robot sometimes disobeyed children's commands thus provoking various emotional reactions. The recording was executed

at two schools – MONT and OHM –, and features 51 children with 21 males and 30 females, with ages ranging from 10 to 13.

Five annotators listened to the turns in sequential order and labelled each word independently from each other as neutral or as belonging to one of ten other classes. In our experiments – as in the Challenge – the unit of analysis is neither the word nor the turn, but some intermediate chunk being the best compromise between the length of the unit of analysis and the homogeneity of the different emotional/emotion-related states within one unit. The final labelling and labeller agreement levels for chunk are determined by the majority voting from labels of the five labellers on the word level onto one label for the whole chunk. Following this, chunks are classified into the 2-class labelling: **NEG**ative (subsuming $angry$, $touchy$, $reprimanding$, and $emphatic$) and **IDL**e (consisting of all other states). Fig. 1 displays the instance distribution of NEG and IDL with labeller agreement levels. For our experiments, we use the whole corpus consisting of 18 216 chunks, and guarantee speaker independence by using the data recorded at school of OHM as candidate pool, and the data recorded at another school of MONT as test set. Table 1 shows the details of the partition of the instances and the speakers for the pool and the testing set.

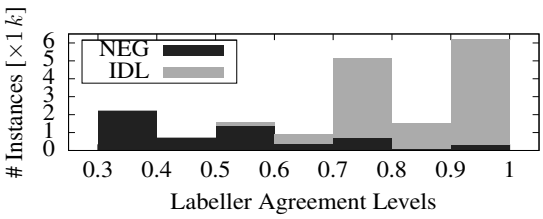

Figure 1: Number of instances with the according labeller agreement levels in the whole AEC.

Table 1: Number of speakers and instances per partition of FAU AIBO 2-class task. m/f: male/female; NEG/IDL: negative/idle.

	# speakers	# instances		
		NEG	**IDL**	Σ
Pool	13m/13f	3 358	6 601	9 959
Test	8m/17f	2 465	5 792	8 257
Σ	21m/30f	5 823	12 393	18 216

3. Methodology

Given a small amount of labelled data with its gold standard and corresponding labeller agreement levels, we build an uncertainty model for each class based on respective labeller agreement levels for all instances in the training set. Then, a regression is made for all instances in the candidate pool by each uncertainty model, resulting in multiple predictions of labeller agreement levels per instance. Upon these predictions of each instance, we assign a class according to a predefined decision-making mechanism, and calculate its corresponding confidence value. Later, for each class we select the instances by either of the following two methods: 1) By class sparseness (AL_{CS}). This means the instances predicted as the sparse classes will be chosen randomly. 2) By confidence values (AL_{CV}). Those instances will be selected whose confidence value is within a certain range, trying to avoid adding the annotation noise which the instances with lowest confidence value could contain. Afterwards, we deliver these selected instances for human labelling, producing human agreement levels and the corresponding 'gold standard'. Finally, we move the selected instances from the candidate pool to the training set. This process repeats until some criteria are met.

Algorithm 1: Active learning by label uncertainty.

Input:
$\mathcal{L}$: small amount of labelled data (with labeller agreement levels and gold standard);
$\mathcal{U}$: large amount of unlabelled candidate data pool;
m: number of labellers;
k: number of classes;
n: number of selected instances for each repetition;
Output:
$\mathcal{H}$: enhanced emotion classifier;

Process
Obtain the priors of each class P_i $(i = 1, \ldots, k)$ in $\mathcal{L}$;
repeat
  (Option) Upsample the training set $\mathcal{L}$ to even class distribution $\mathcal{L}_{\mathcal{D}}$;
  Given the labeller agreement levels, build uncertainty models by usage of $\mathcal{L}/\mathcal{L}_{\mathcal{D}}$ for each class, M_i, $i = 1, \ldots, k$;
  for $i = 1, \ldots, k$ **do**
    Regress $\mathcal{U}$ by uncertainty models M_i, then, assign every instance in $\mathcal{U}$ with a predicted labeller agreement level S_i;
    Normalise S_i into [0,1];
  end
  Given k normalised predicted labeller agreement levels S_i $(i = 1, \ldots, k)$ for each instance in $\mathcal{U}$, assign it with the class determined by a decision-making mechanism. Then, calculate corresponding confidence values $V \in [0, 1]$;
  Select subset $\mathcal{T}$ from $\mathcal{U}$ with predefined instances number n, and label them (*cf. Method 1 or 2*);
  Add the selected subset $\mathcal{T}$ into the training set $\mathcal{L}$, $\mathcal{L} = \mathcal{L} \cup \mathcal{T}$;
  Remove the selected subset $\mathcal{T}$ out of the unlabelled set $\mathcal{U}$, $\mathcal{U} = \mathcal{U} \setminus \mathcal{T}$;
until
i) Targeted performance is achieved; or
ii) No more instances remain in the pool; or
iii) The number of predetermined iteration times or instances numbers has been reached ;

Method 1: By class sparseness

1. Randomly select n instances from $\mathcal{U}$ that are predicted as the sparse class ('NEG' in our case);
2. Deliver the selected subset $\mathcal{T}$ to m experts for labelling, respectively;

Method 2: By confidence values

1. **for** $i = 1, \ldots, k$ **do**
2. Sort the instances that are predicted as class C_i by confidence value S, producing queue Q_i;
3. Select n_i ($n_i = n \times P_i$) instances which are in the middle of queue Q_i;
4. **end**
5. Deliver the selected subset $\mathcal{T}$ to m experts for labelling, respectively;

The issue of unbalanced class distribution and labeller agreement levels might impact the affect recognition performance by producing several problems. For example, instances are generally inclined to be classified as the dominating classes, making the selection process weak. This means that the dominant class could be recognised incrementally better with respect to the minority classes. In order to avoid this, several techniques are considered in our algorithm. First, we upsample the training set to balance the class distribution if necessary. Second, we normalise the predicted levels for each class. This partly ensures a better class assignment to an instance according to its multiple prediction levels. Third, we choose the instances in proportion to the prior of each class in the initial training set. Fourth, if upsampling is not applied, we employ a class sparseness method, which aims to enhance the weight of sparse classes.

In the case of our experiments on acoustic emotion recognition, only a 2-class task (NEG vs. IDL) is considered. Thus, the AL algorithm can be simplified. Only one uncertainty model needs to be built, since the labeller agreement level for IDL is complementary with that for NEG. Further, the decision-making mechanism can be simplified to a threshold based decision to distinguish between NEG and IDL. In the same vein, the confidence value can be applied as:

$$V(x) = |x - S_T|/(S_{\max} - S_{\min}), \tag{1}$$

where x is the normalised predicted labeller agreement levels, S_T is a predefined threshold between NEG and IDL, and $S_{\max}$, $S_{\min}$ are the maximal and minimal normalised predicted levels of NEG or IDL. For example, if we build one uncertainty model for IDL, $S_{\max}$ and $S_{\min}$ will be 1 and S_T for IDL, and S_T and 0 for NEG, respectively.

In addition, a crucial point in AL_{CV} is the query function, which can be defined as:

$$Q(x) = \begin{cases} 1, & \text{if } D(x) = \arg\min_x |V(x) - V_m|, \\ 0, & otherwise, \end{cases} \tag{2}$$

where $V(x)$ evaluates the confidence value of instance x after normalisation, and V_m is the confidence value of that instance in the centre of the ranking queue. Ideally, for equally distributed predictions, V_m will be 0.5. Yet, practically, V_m is determined by the scenario of the actual distribution, and it is not fixed but varies within the change of the candidate pool through the learning iterations.

4. Experiments and Results

According to whether instance upsampling is undertaken, we evaluate AL based on label uncertainty by the class sparseness method in Subsection 4.2 and confidence value method in Subsection 4.3.

4.1. Experimental Setups

According to the INTERSPEECH 2009 Emotion Challenge (EC), we exactly follow the experimental setup as in [25]. The feature set contains 384 features resulting from a systematic combination of 16 low-level-descriptors (LLDs) and corresponding first order delta coefficients with 12 functionals. Thus, the total feature vector per chunk contains $16 \times 2 \times 12 = 384$ attributes. The features are extracted with openSMILE [26] and details can be found in [25]. For our experiments, we considered two classifiers: One is Support Vector Machines (SVMs) for evaluating on the test set. Here, we applied Sequential Minimal Optimisation (SMO) algorithm with polynomial kernel and a complexity constant of 0.05, as used in [25]. The other one is a DecisionStump Tree which was employed to build the uncertainty model and exploit instances from the candidate pool by label uncertainty. Here, we applied AdditiveRegression meta-learning with an iteration number of 100 and subspace size of 0.2. Both classifiers are implemented in the Weka toolkit [27]. As primary evaluation measure, we retain the choice of unweighted average recall as was used in the Challenge held in 2009 [28].

AL comprises a random selection of 500 instances from candidate pool as initial small training set, which resembles approximately 3 % of the whole corpus. Thus, the other 9 459 instances are maintained in the pool waiting to be exploited. Then, we choose n = 200 instances as step size in each learning iteration. Finally, to reduce the influence of 'lucky' or 'unlucky' selection for the initial training set, we repeat 20 times with different random initialisations ('seeds'), leading to 20 runs of the whole iteration process executed.

Moreover, the number of labellers m is equal to five and the number of classes k is set to two. According to the statistic distribution of IDL and NEG (cf. Fig. 1), we define S_T as 0.58. This means that if the normalised prediction levels of instances are greater than 0.58, they will be classified as IDL; otherwise, they will be classified as NEG.

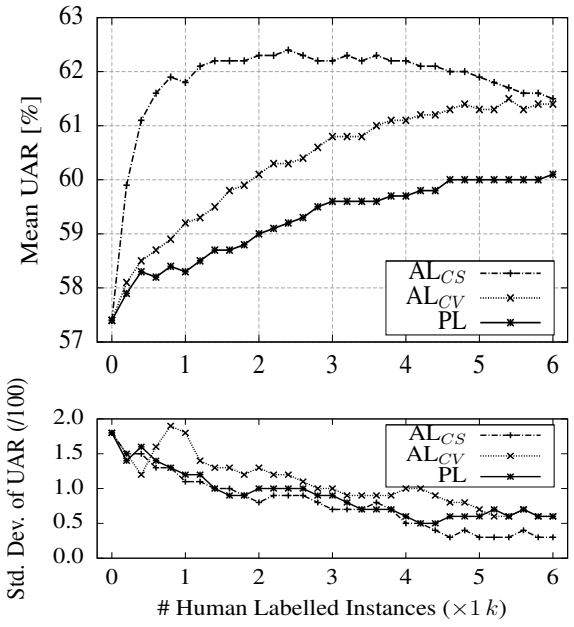

Figure 2: Mean UAR (top) and its standard derivation (std. dev.) (bottom) vs. number of human labelled instances. Comparison of active learning (AL) by label uncertainty with the method based on confidence values (AL_{CV}), or class sparseness (AL_{CS}), and passive learning (*PL*) in 20 independent runs of the whole process *without* instance upsampling.

4.2. By Class Sparseness

Fig. 2 displays the performance (upper: UAR, bottom: standard derivation) of AL_{CS} and AL_{CV}. Obviously, in the case

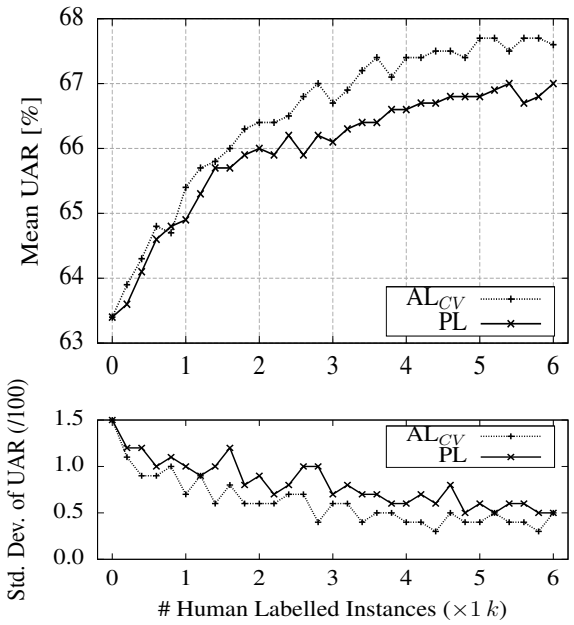

Figure 3: Mean UAR (top) and its standard derivation (std. dev.) (bottom) vs. number of human labelled instances. Comparison of active learning (AL) by label uncertainty with the method based on confidence values (*AL$_{CV}$*), and passive learning (*PL*) in 20 independent runs of the whole process *with* instance upsampling.

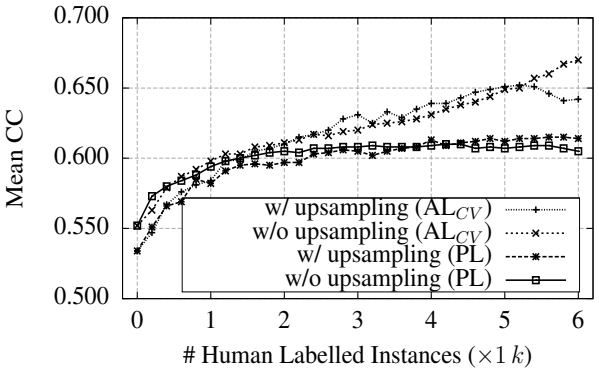

Figure 4: Mean Correlation Coefficient (CC) of the uncertainty model evaluated on the remaining pool vs. the number of human labelled instances. Comparison of active learning (AL) by label uncertainty based upon confidence values (AL$_{CV}$) and passive learning (PL) with/-out upsampling in five independent runs.

of no upsampling technique executed in the process, the two methods outperform the passive learning (PL) which considers random instance selection. However, the best performance can be observed by applying AL$_{CS}$. The most impressive performance trend can be found in the first few iterations. After the 4th iteration (800 instances added), the accuracy increases from 57.4 % to 61.9 % of UAR, with 4.5 % absolute gain. The best performance is achieved at roughly 62.5 % after 2.4 k instances are added. Compared to the performance (60.7 %) of the classifier trained on the whole candidate pool (about 10 k instances), UAR is increased by about 2 % and the training set is reduced by 70 %. This can be expected: It accelerates the learning rate for selecting the 'right' instances within the first few iterations, which helps to improve the acoustic model rapidly for the sparse class 'NEG'. Apart from that, it is worth noting that its performance becomes somewhat worse when adding more instances after the best performance is achieved. This could be due to the limited number of sparse instances (NEG) in the candidate pool.

4.3. By Confidence Values

Fig. 3 compares the performance of AL$_{CV}$ and PL when the instance upsampling strategy is used. The subfigure (upper) in Fig. 3 shows that AL$_{CV}$ outperforms PL. Especially after 10 iterations (2 k instances added), UAR speeds up, rising to an average improvement of 0.8 % UAR absolute. For AL$_{CV}$, we obtain an UAR of 67.7 % which is equal to the baseline in [24] after 5 k instances are added in the training set, reducing the training set to 4 % of its size. Moreover, the subfigure (bottom) in Fig. 3 further indicates that the AL$_{CV}$ shows more stable performances than PL. This would greatly save a lot of time and money for human labelling, and reduce the training complexity and training time.

We further investigate the impact of human agreement levels on instance selection of AL. Fig. 4 gives a comparison of correlation coefficient (CC) between AL$_{CV}$, and PL with/-out instance upsampling. Note that these results are evaluated on the remaining candidate pool instead of the testing set. It shows clearly that after 2 k human labelled instances are added, both uncertainty models built by AL$_{CV}$ with/-out upsampling outperform the models built by PL.

5. Conclusions and Future Work

In this paper, we proposed novel active learning approaches based on label uncertainty for the subjective speech phenomenon of emotion, aiming to exploit the reliability of labelling. Two instance selection methods based on class sparseness and confidence value were investigated. The experimental results show the efficiency of our proposed AL. With AL$_{CS}$, the best performance was achieved by 62.5 % UAR by adding only 2.4 k instance for training. This overtakes the performance of the classifier trained on the whole pool set with approximately 2 % UAR absolute gain, and reduces the training instances by roughly 70 %. By AL$_{CV}$, the amount of training instances is also reduced by 45 % when balancing the training instances to achieve the baseline of the INTERSPEECH 2009 Emotion Challenge. This is quite interesting when we put acoustic emotion recognition into practise, where large unlabelled instances can be easily collected in automated ways, but labelling is expensive and time consuming.

Future work will continue on other subjective speech tasks to investigate robustness and universality. In addition, the performance of semi-supervised learning on label uncertainty is of interest.

6. Acknowledgements

The research leading to these results has received funding from the Chinese Scholarship Council (CSC) and the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement No. 289021 (ASC-Inclusion).

7. References

[1] B. Schuller, Z. Zhang, F. Weninger, and G. Rigoll, "Using Multiple Databases for Training in Emotion Recognition: To Unite or to Vote?" in *Proc. INTERSPEECH 2011*, Florence, Italy, 2011, pp. 1553–1556.

[2] Z. Zhang, F. Weninger, M. Wöllmer, and B. Schuller, "Unsupervised Learning in Cross-Corpus Acoustic Emotion Recognition," in *IEEE workshop on Automatic Speech Recognition and Understanding (ASRU)*, Big Island, HY, 2011, pp. 523–528.

[3] Z. Zhang and B. Schuller, "Active Learning by Sparse Instance Tracking and Classifier Confidence in Acoustic Emotion Recognition," in *Proc. INTERSPEECH 2012*, Portland, OR, 2012, 4 pages.

[4] Z. Zhang, J. Deng, and B. Schuller, "Co-Training Succeeds in Computational Paralinguistics," in *Proc. 2013 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Vancouver, Canada, 2013, to appear.

[5] D. Braha, Ed., *Data Mining for Design and Manufacturing: Methods and Applications*. Kluwer Academic, 2001.

[6] B. Schuller, Z. Zhang, F. Weninger, and G. Rigoll, "Selecting Training Data for Cross-Corpus Speech Emotion Recognition: Prototypicality vs. Generalization," in *Proc. 2011 Speech Processing Conference*, Tel Aviv, Israel, 2011, 4 pages.

[7] Y. Freund, H. S. Seung, E. Shamir, and N. Tishby, "Selective sampling using the query by committee algorithm," *Machine learning*, vol. 28, no. 2, pp. 133–168, 1997.

[8] W. Chen, G. Liu, J. Guo, and Y.-J. Guo, "A new method for sample selection in active learning," in *2009 International Conference on Machine Learning and Cybernetics*, Baoding, China, 2009, pp. 2270–2274.

[9] M. Li and I. Sethi, "Confidence-based active learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 8, pp. 1251–1261, 2006.

[10] B. Settles, "Active learning literature survey," Department of Computer Sciences, University of Wisconsin–Madison, Wisconsin, WI, Tech. Rep., 2009.

[11] G. Riccardi and D. Hakkani-Tur, "Active learning: theory and applications to automatic speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 4, pp. 504–511, 2005.

[12] R. Liere, "Active learning with committees: An approach to efficient learning in text categorization using linear threshold algorithms," Ph.D. dissertation, Oregon State Univ., Portland, 2000.

[13] A. McCallum and K. Nigam, "Employing EM in pool-based active learning for text classification," in *Proc. 15th International Conference on Machine Learning (ICML)*, Madison, WI, 1998, pp. 359–367.

[14] N. Roy and A. McCallum, "Toward optimal active learning through sampling estimation of error reduction," in *Prof. Int'l. Conf. on Machine Learning (ICML)*, Williamstown, MA, 2001, pp. 441–448.

[15] B. Settles and M. Craven, "An analysis of active learning strategies for sequence labeling tasks," in *Proc. Empirical Methods in Natural Language Processing(EMNLP)*, Honolulu, HY, October 2008, pp. 1070–1079.

[16] Z. Xu, R. Akella, and Y. Zhang, "Incorporating diversity and density in active learning for relevance feedback," in *Proc. European Conference on Information Retrieval (ECIR)*, Rome, Italy, 2007, pp. 246–257.

[17] P.-H. Gosselin and M. Cord, "Active learning methods for interactive image retrieval," *IEEE Transactions on Image Processing*, vol. 17, no. 7, pp. 1200–1211, 2008.

[18] S. Sivaraman and M. Trivedi, "A general active-learning framework for on-road vehicle recognition and tracking," *IEEE Transactions on Intelligent Transportation Systems*, vol. 11, no. 2, pp. 267–276, 2010.

[19] I. Sneddon, M. McRorie, G. McKeown, and J. Hanratty, "The belfast induced natural emotion database," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 32–41, 2012.

[20] B. Schuller, "Multimodal Affect Databases — Collection, Challenges & Chances," in *Handbook of Affective Computing*, R. A. Calvo, S. DMello, J. Gratch, and A. Kappas, Eds. Oxford, UK: Oxford University Press, 2013.

[21] D. Seppi, A. Batliner, B. Schuller, S. Steidl, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, N. Amir, and V. Aharonson, "Patterns, Prototypes, Performance: Classifying Emotional User States," in *Proc. INTERSPEECH 2008*, Brisbane, Australia, 2008, pp. 601–604.

[22] M. Grimm and K. Kroschel, "Evaluation of natural emotions using self assessment manikins," in *2005 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Cancun, Mexico, 2005, pp. 381–385.

[23] S. G. Karadogan and J. Larsen, "Combining semantics and acoustic features for valence and arousal recognition of speech," in *Proc. 3rd International Workshop on Cognitive Information Processing*, Baiona, Spain, 2012, no pagination.

[24] S. Steidl, *Automatic Classification of Emotion-Related User States in Spontaneous Children's Speech*. Berlin: Logos Verlag, 2009.

[25] B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 Emotion Challenge," in *Proc. INTERSPEECH 2009*, Brighton, UK, 2009, pp. 312–315.

[26] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE – The Munich Versatile and Fast Open-Source Audio Feature Extractor," in *Proc. ACM Multimedia (MM)*, Florence, Italy, 2010, pp. 1459–1462.

[27] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[28] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," *Speech Communication*, vol. 53, no. 9, pp. 1062–1087, 2011.