

Affective rating ranking based on face images in arousal-valence dimensional space*

Guo-peng XU, Hai-tang LU, Fei-fei ZHANG, Qi-rong MAO[‡]

School of Computer Science and Communication Engineering, Jiangsu University, Zhenjiang 212013, China

E-mail: gpxu@ujs.edu.cn; 1406404872@qq.com; susanzhang1231@sina.com; mao_qr@mail.ujs.edu.cn

Received Apr. 19, 2017; Revision accepted Aug. 31, 2017; Crosschecked June 12, 2018

Abstract: In dimensional affect recognition, the machine learning methods, which are used to model and predict affect, are mostly classification and regression. However, the annotation in the dimensional affect space usually takes the form of a continuous real value which has an ordinal property. The aforementioned methods do not focus on taking advantage of this important information. Therefore, we propose an affective rating ranking framework for affect recognition based on face images in the valence and arousal dimensional space. Our approach can appropriately use the ordinal information among affective ratings which are generated by discretizing continuous annotations. Specifically, we first train a series of basic cost-sensitive binary classifiers, each of which uses all samples relabeled according to the comparison results between corresponding ratings and a given rank of a binary classifier. We obtain the final affective ratings by aggregating the outputs of binary classifiers. By comparing the experimental results with the baseline and deep learning based classification and regression methods on the benchmarking database of the AVEC 2015 Challenge and the selected subset of SEMAINE database, we find that our ordinal ranking method is effective in both arousal and valence dimensions.

Key words: Ordinal ranking; Dimensional affect recognition; Valence; Arousal; Facial image processing

<https://doi.org/10.1631/FITEE.1700270>

CLC number: TP391

1 Introduction


Affect recognition which involves multiple scientific disciplines such as neuroscience, psychology, cognitive science, and computer science, has attracted great interest in the past two decades. The mainstream of research in this field has mostly focused on the recognition of facial and vocal affect in terms of basic emotions. However, a number of researchers have found that in everyday interac-

tions people usually exhibit non-basic, subtle, and rather complex mental or affective states such as thinking, depression, and embarrassment (Baron-Cohen, 2004). A small number of discrete emotion categories are not enough to reflect the complexity of affective states in those scenarios. Therefore, the use of the dimensional description of human affect is advocated, in which an affective state is characterized by a number of latent dimensions (Russell, 1980; Scherer, 2000; Scherer et al., 2001). The most widely used dimensional description of affect is the two-dimensional (2D) emotional space of arousal (active vs. passive) and valence (positive vs. negative).

In the research of affect recognition in the arousal and valence dimensional space, the machine learning methods, which are used to model and predict affect, are mostly classification and regression.

[‡] Corresponding author

* Project supported by the National Natural Science Foundation of China (Nos. 61272211 and 61672267), the Open Project Program of the National Laboratory of Pattern Recognition (No. 201700022), the China Postdoctoral Science Foundation (No. 2015M570413), and the Innovation Project of Undergraduate Students in Jiangsu University (No. 16A235)

 ORCID: Guo-peng XU, <http://orcid.org/0000-0002-2062-0763>

© Zhejiang University and Springer-Verlag GmbH Germany, part of Springer Nature 2018

The classification methods, which use audio signals or visual signals as feature representations, usually reduce the recognition problem to a two-class problem (active vs. passive or positive vs. negative classification (Schuller et al., 2009; Nicolaou et al., 2010)) or a four-class problem (classifying the quadrants of the 2D A-V space (Glowinski et al., 2008; Wöllmer et al., 2010b)). The shortcoming of these methods, on the one hand, is that the categories to be classified are obviously coarse and the affect states cannot be recognized at a fine level of granularity. On the other hand, the labels of affect are naively treated as independent tags and the ordinal property of the annotations is not used. The regression models typically learn a function which can best fit the mapping from the feature space to the affective annotation space with appropriate regularizations. Although they take the inter-relationship of labels into account, transforming implicitly the ordinal scale into a numerical scale introduces strong non-linear scale bias (Martinez et al., 2014). Specifically, the regression methods treat the annotations simply as proportional numerical scales and neglect the fact that the difference between affect annotations in a certain dimensionality is non-uniform (Martinez et al., 2014). Therefore, to improve the affect recognition performance in the A-V space, learning to rank (or ranking) methods have begun to attract the attention of researchers.

Learning to rank, or ranking for short, being the algorithm of machine learning, is widely used in the field of information retrieval (Joachims, 2002; Xu and Li, 2007). Ranking approaches can be generally divided into three categories: pointwise, pairwise, and listwise (Liu, 2011). The pointwise approach learns a function, which is trained on individual instances, to map the feature vectors of the given samples to corresponding real-value scores or ordinal labels. It is similar to classification or regression, but the ordinal property of labels is considered. In practice, it is widely used because of its simplicity and effectiveness. The pairwise approach learns a scoring function using relative ordering relationship of two input candidates in pairs. It exploits more information about the ground truth, but it needs to handle a larger number of training instances (usually quadratic in the size of training data), which potentially causes slower or less efficient training. The listwise approach views the list of input candidates

as a single instance to learn a scoring function that is employed to rank new instances. It uses the most ordering information of input candidates, but it has a high computing complexity and also easily leads to overfitting. What is more, the pairwise and listwise approaches always aim at predicting ordering relationship of the given samples, not the exact ordinal labels.

In this study, we explore the use of a pointwise approach, specifically, ordinal regression, to solve the affect recognition problem in the A-V space. Our algorithm uses the relative order of affect annotations to conduct an effective prediction for exact affective ratings. We first discretize the continuous and real-value annotations in the arousal and valence dimensions respectively to form finite affective ratings. Then a series of cost-sensitive binary classifiers are trained using all samples which are relabeled according to the comparison results between corresponding ratings and a given rank of a binary classifier. We finally obtain the affective ratings by aggregating the results of basic binary classifiers. Our method can be used in the real applications where the affect intensity in the arousal or valence dimension needs to be estimated. Specifically, it can be applied to film recommendation based on affect content, human depression detection, driver emotion analysis, and online education. In these scenarios, facial images are detected and cropped, and the feature vectors of the faces are extracted as the inputs of our affective rating ranking (ARR) framework. Then, the affect intensity or rating in arousal or valence can be estimated using the ARR framework. The prediction results can be ranked to be applied to the recommendation task or mapped to subtler emotions in the emotion analysis task. The main contributions of this paper are summarized as follows:

1. As far as we know, this is the first paper proposing the ordinal regression approach to predict affect ratings based on face images in the A-V space. The ranking method appropriately uses the ordinal property of dimensional affect annotations. The experimental results show that the ordinal property is available and important for improving the dimensional affect recognition performance.

2. In our ranking framework, a cost-sensitive setting is adopted for each basic binary classifier. We conduct exhaustive experiments to compare the performance of our ranking method with different

cost-sensitive settings being employed and empirically find the most suitable cost-sensitive settings for affect rating ranking in the arousal and valence dimensions, respectively.

2 Related works

In this section, we first review the typical studies where the traditional classification or regression methods are employed to model the affect recognition problem in the A-V space. Then we focus on the ordinal ranking algorithms related to our work and the existing attempts using ranking-based approaches in affect recognition.

2.1 Classification methods

In terms of affect recognition using classification methods in the A-V dimensional space, the widely adopted strategy is to simplify the problem to a three-class valence-related classification problem: positive, neutral, and negative affect classification (Yu et al., 2004; McDuff et al., 2010). A similar and simple method is to reduce the dimensional affect classification problem to a two-class problem (active vs. passive or positive vs. negative (Schuller et al., 2009; Nicolaou et al., 2010)) or a four-class problem (classifying the quadrants of the 2D A-V space (Ioannou et al., 2005; Caridakis et al., 2006; Glowinski et al., 2008; Wöllmer et al., 2010b)). Systems that aim at dimensional affect recognition, considering that the affective states are represented along a continuum, commonly tend to quantize the continuous range into several levels. Wöllmer et al. (2008) used the sensitive artificial listener (SAL) database, quantized the annotations of valence and arousal into four and seven levels respectively, and adopted the conditional random fields (CRFs) and support vector machine (SVM) to predict the quantized affective labels. Wöllmer et al. (2010a) used a context-sensitive technique and multimodal data containing facial and audio information to recognize three to five levels of the A-V values. Obviously, the aforementioned classification methods recognize only the affect states at a coarse level. In addition, naively treating annotations as independent category tags does not take the inter-relationship (such as the ordering relationship of labels) into account.

2.2 Regression methods

Some models based on regression methods have been proposed to conduct continuous dimensional

affect prediction. Nicolaou et al. (2011) proposed a multimodal system to continuously predict valence and arousal states of a speaker using support vector regression (SVR) and long-short term memory (LSTM) regression. He et al. (2015) used multimodal feature selection and feature fusion, and used a deep bidirectional long-short term memory recurrent neural network framework to obtain the best prediction results in AVEC 2015 Challenge. Even though these methods are popular in the research of continuous affect prediction in the A-V space, the aforementioned non-linear scale bias still exists in continuous annotations.

2.3 Ranking methods

In ordinal ranking research, Li and Lin (2006) proposed a reduction framework from ordinal regression to binary classification using extended samples and formally proved that a weighted 0/1 loss of the binary classifier could bound the mislabeling cost of the ranking rule constructed from the binary classifier. Chang et al. (2010) treated the age estimation problem as an ordinal regression problem and adopted the reduction framework above to obtain better age prediction performance than traditional classification and regression methods. Recently, more and more ranking algorithms have been developed to solve the human age estimation problem. Chang and Chen (2015) developed their ranking-based age estimation approach and presented a cost-sensitive ordinal hyperplane ranking algorithm. In their approach, the age ranks were inferred by aggregating a series of basic binary classification results, in which cost sensitivities among the ranks were introduced to improve the final aggregating performance. Lim et al. (2015) proposed a VRank framework for facial age estimation, using a deep learning architecture to achieve efficient features and ranking each age with the ranking SVM algorithm. In the last stage, the proposed voting system algorithm was used to infer the age by weighted relational information. Abousaleh et al. (2016) presented a deep learning framework called the ‘comparative region convolutional neural network (CRCNN)’, in which a set of hints (comparative relations) were generated by comparing the input face with reference faces first and then all the hints obtained were aggregated to infer the age of a person. Feng et al. (2017) combined the strength of cost-sensitive label ranking methods

with the power of low-rank matrix recovery theories, in which the correlations among different age labels were captured and the model complexity was also controlled. In the test stage, the decision values for different age labels of the test image were ranked in descending order and the age label ranked at the top was selected as the estimated age value.

Considering the defect of classification and regression methods mentioned above, ranking-based approaches have attracted the attention of researchers though the attempts in affect recognition are far fewer than those using classification and regression. Yang and Chen (2011) introduced a novel learning to rank algorithm called RBF-ListNet to rank the affect of a set of music pieces. This list-wise approach determining the ordering relationship of given samples is applicable to the task of music retrieval according to affect content, but it is not quite suitable for the task of human affect recognition, which prefers to obtain exact affect labels, not just their ordering relationship. Martinez et al. (2014) compared the performance of pairwise preference learning and binary classification on the dimensional affect dataset SAL and other datasets. Although the results suggested that the preference learning method leads to more reliable, generic, and robust models that capture more information about the ground truth, the pairwise approach here could not be directly used to handle the problem of learning and predicting exact affective ratings. Our proposed method based on the pointwise approach was inspired by Chang and Chen (2015), and it focuses on affective rating estimation by ordinal ranking based on face images in the A-V space. Unlike classification methods that transform continuous real-value annotations into independent nominal categories or regression approaches that treat the annotations simply as proportional numerical scales, we try to use the ordinal regression approach which appropriately employs the ordering relationship among annotations to improve the affect recognition performance in the A-V dimensional space.

3 Affective rating ranking framework

In our ARR framework, we first discretize the continuous, real-value affect annotations in arousal and valence dimensions respectively to form finite affective ratings. Then a series of cost-sensitive bi-

nary classifiers are trained using relabeled samples and their weights. Finally, the results of binary classifiers are aggregated to obtain the affective ratings. Fig. 1 shows the illustration of our ARR framework for arousal rating estimation of a given sample, and the valence rating inference is just similar. Next, we will describe these three parts in detail.

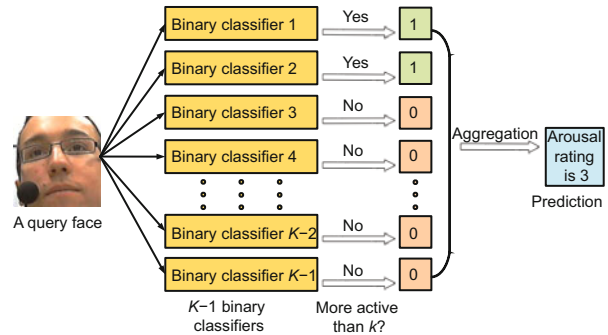


Fig. 1 Illustration of our affective rating ranking framework for arousal rating estimation of a given sample

3.1 From continuous annotations to finite ratings

In the field of dimensional affect recognition, the annotations of the affect states in a certain dimension are usually continuous real values limited in a range such as $[-1, 1]$. Although the interval in which the annotation lies is small, the number of possible annotation values is infinite. Directly using the ordinal ranking algorithms on the annotations in a continuous interval is not feasible, because the ordinal ranking algorithms need finite labels or ranks. Therefore, we have to transform the continuously labeled annotations into discrete and finite ratings.

However, little attention has been paid to whether there are definite boundaries along the continuous annotations to distinguish among different intensities or levels (Gunes and Pantic, 2010). The most common way to study this issue is to quantize the valence and arousal dimensions into an arbitrary number of intensities and levels (Wöllmer et al., 2008, 2010a). We adopt a similar approach, which divides the value range of annotations of a certain dimension into finite intervals with the same length. Each interval stands for an affective intensity or rating. For example, if we divide the continuous annotations into K intervals, we annotate these K intervals with $1, 2, \dots, K$ ratings respectively, where

the smaller real value lies in the interval, and the lower rating corresponds to the interval. We describe a transforming process on a dimensional affect dataset in detail in Section 4. After this processing, we can use the aforementioned ordinal ranking algorithm to solve the affect recognition problem in the A-V space.

3.2 Cost-sensitive binary classifiers for affective rating ranking

Assume a set of training face images $I_i, i = 1, 2, \dots, m$. We use $\mathbf{x}_i \in \mathbb{R}^d$ to represent the feature vector extracted from I_i and let y_i be the transformed affective rating of I_i . According to the rating transformation method above, $y_i \in \{1, 2, \dots, K\}$, where K is the number of affective ratings and y_i is treated as a rank order. Denote $\mathcal{S} = \{\mathbf{x}_i | i = 1, 2, \dots, m\}$. For a given affective rating k in a certain dimensionality (i.e., arousal or valence), the whole dataset is split into two subsets as \mathcal{S}_k^+ and \mathcal{S}_k^- :

$$\begin{cases} \mathcal{S}_k^+ = \{(\mathbf{x}_i, +1) | y_i > k\}, \\ \mathcal{S}_k^- = \{(\mathbf{x}_i, -1) | y_i \leq k\}. \end{cases} \quad (1)$$

Then the two subsets are used to train a basic binary classifier. This binary classifier will answer such a query: “Is the face more active than arousal rating k ?” or “Is the face more positive than valence rating k ?”. With k from 1 to $K - 1$, $K - 1$ binary classifiers are learned and their binary decision results for a given sample will indicate the ordering relationships between the affective rating of the given sample and affective ratings 1 to $K - 1$. We can use these ordering relationships to infer the exact affective rating of the given sample. Thus, the affective rating ranking problem is reduced to a series of binary classification subproblems. It is naturally thought that if each binary classifier is trained well, the correct ranking result is more likely to be obtained.

Before concentrating on each binary classification subproblem, we introduce the performance evaluation measurements first. In our ARR framework, we select the mean absolute error (MAE) (Geng et al., 2007) and cumulative score (CS) (Geng et al., 2007) which are widely adopted in human age estimation evaluation as performance indices. They are easy to calculate and can be used for performance comparison among classification, regression, and ranking methods from different angles. MAE

measures the mean difference between the labels and the predictions on test samples. It can be defined as

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^N |y_i^* - y_i|, \quad (2)$$

where y_i^* is the predicted affective rating, y_i is the ground truth rating, and N is the number of test samples. CS calculates the percentage of test samples whose prediction errors are not larger than an error tolerance level L , which is defined as

$$\text{CS}(L) = \frac{1}{N} \sum_{i=1}^N [\![|y_i^* - y_i| \leq L]\!] \times 100\%, \quad (3)$$

where $[\![\cdot]\!]$ is the truth-test operator, which is 1 if the inner condition is true, and 0 otherwise. The adoption of CS measurement is meaningful because when we divide the continuous annotations into several fine intervals, the affect states represented by adjacent ratings have little difference, and the predicted error at a low tolerance level is acceptable.

Let us focus on each binary classification subproblem. Chang and Chen (2015) applied a cost-sensitive setting to each binary classification subproblem and obtained better age estimation results than those generated by using common 0/1 misclassification cost. Therefore, to obtain better affective rating prediction results by ranking, we also introduce a cost-sensitive setting to each binary classifier. Assume that the cost of misclassifying a sample for affective rating label t in subproblem k is $c_k(t)$, where $t = 1, 2, \dots, K$ and $k = 1, 2, \dots, K - 1$. For the i^{th} sample \mathbf{x}_i , the cost of subproblem k can be represented as $c_k(\mathbf{x}_i)$. We consider mainly three kinds of cost settings and select the best setting from them. The first one is the absolute cost, which is defined as

$$c_k(t) = |t - k|. \quad (4)$$

The second cost is a special type of 0/1 cost corresponding to CS measure, not common accuracy measure, which is defined as

$$c_k(t) = \begin{cases} 0, & \text{if } (t - L) \leq k \leq (t + L), \\ 1, & \text{otherwise.} \end{cases} \quad (5)$$

For the last cost setting, it combines the characteristics of the first two costs and can be defined as

$$c_k(t) = \begin{cases} 0, & \text{if } (t - L) \leq k \leq (t + L), \\ |t - k| - L, & \text{otherwise.} \end{cases} \quad (6)$$

The error within L is discarded and only the absolute error outside L is counted.

Then we just focus on training each better cost-sensitive binary classifier to improve the final ranking performance. Fig. 2 shows the training process for the k^{th} cost-sensitive binary classifier.

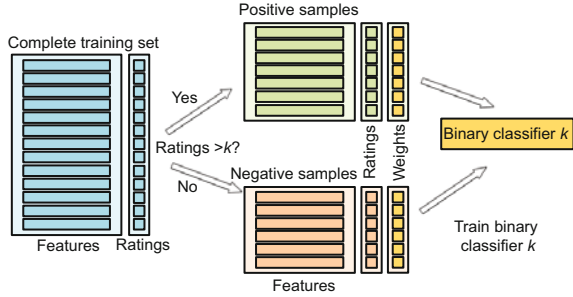


Fig. 2 Training process of the k^{th} cost-sensitive binary classifier

In our approach, to implement a cost-sensitive setting, the data reweighted SVM is employed to solve the k^{th} binary classification subproblem as follows:

$$\min_{\mathbf{w}_k, b_k, \xi} \frac{1}{2} \|\mathbf{w}_k\|^2 + C \left(\sum_i c_k(\mathbf{x}_i) \xi_i \right)$$

$$\text{s.t. } \forall i, l_k(\mathbf{x}_i) (\mathbf{w}_k^T \phi_k(\mathbf{x}_i) + b_k) \geq 1 - \xi_i, \xi_i \geq 0, \quad (7)$$

where $l_k(\mathbf{x}_i) = +1$ if $\mathbf{x}_i \in \mathcal{S}_k^+$ and $l_k(\mathbf{x}_i) = -1$ if $\mathbf{x}_i \in \mathcal{S}_k^-$, ϕ_k is a function mapping the feature vector \mathbf{x}_i into a high-dimensional space, and \mathbf{w}_k and b_k are the hyperplane parameters in the high-dimensional space. Note that all subproblems do not have to share a single kernel and each of them can project its own feature space via ϕ_k . The discriminating function $f_k(\mathbf{x})$ is then employed as

$$f_k(\mathbf{x}) = \mathbf{w}_k^T \phi_k(\mathbf{x}) + b_k. \quad (8)$$

3.3 Aggregating binary decisions for affective rating inference

After training a series of cost-sensitive binary classifiers (SVMs), we can aggregate these binary decision results of binary classifiers for a given sample, such as \mathbf{x}_i , to obtain its affective rating. It can be represented as

$$r(\mathbf{x}_i) = 1 + \sum_{k=1}^{K-1} \llbracket f_k(\mathbf{x}_i) > 0 \rrbracket, \quad (9)$$

where $f_k(\mathbf{x}_i)$ is the binary decision result of the k^{th} binary classifier for a test sample \mathbf{x}_i . Fig. 1 shows a specific aggregation of arousal rating inference of a given sample.

4 Experiments

4.1 Datasets

We perform experiments on two datasets. The first is the benchmarking database of the AVEC 2015 Challenge (Ringeval et al., 2015). It is a subset of the RECOLA multimodal corpus of remote and collaborative affective interactions (Ringeval et al., 2013). There are 27 videos of different subjects in total in this dataset (9 for training, 9 for development, and 9 for testing). The gold standard ratings included in the dataset give the annotations of arousal and valence dimensions for training and development sets. The annotations of the testing set are not provided. Therefore, we use only the training and development videos, i.e., 18 in total, for our experiments. We divide the data into 3 subsets, all 9 training videos for the training set, the 1st, 3rd, 7th, and 8th development videos for the validation set and the rest for the testing set, which is a subject-independent setting.

The second dataset we use in our experiments is a subset selected from the SEMAINE database. The SEMAINE database is a large audiovisual database, which is recorded to study natural social signals occurring during conversations between humans and artificially intelligent agents. The scenario used in the recordings follows the sensitive artificial listener (SAL) paradigm. A user in a recording or a session interacts with an emotionally stereotyped ‘character’ who can be one of four personalities (Prudence, Poppy, Spike, and Obadiah). Recordings in this database involve 150 participants, in total 959 conversations with individual SAL characters, lasting approximately 5 min each. The dimensional affect annotations like arousal, valence, and expectation are included. In our experiments, we select video data of 12 subjects, 6 (subjects 16, 5, 2, 10, 14, and 17) for training, 3 (subjects 3, 12, and 15) for validation, and 3 (subjects 7, 8, and 11) for testing, which is also a subject-independent setting.

4.2 Data processing

We extract all frames from the videos first. Then we crop the face from each frame and resize each face image to the same size 224×224 .

The annotations of arousal and valence dimensions are continuous real values in the datasets. As discussed above, we should transform the continuous and infinite values into discrete and finite ratings. For arousal annotation in the AVEC 2015 benchmarking database, the value range in the dataset is $(-0.6, 0.6)$ and we first divide the whole range into several intervals with the same length 0.1, thus creating 12 derived ratings in total. However, actually the face images corresponding to the ratings which are close to the two ends are far fewer than those belonging to ratings near the center. To mitigate the data imbalance problem, we combine some ratings and do oversampling and undersampling. Finally, the discretized ratings and some example face images are shown in Fig. 3a. For valence annotation, the value range in the dataset is $(-0.2, 0.6)$, and we process the annotation in the same way as with the arousal one, forming seven ratings shown in Fig. 3b. The specific image numbers of different ratings on different data subsets in arousal and valence dimensions are shown in Tables 1 and 2, respectively. For the dataset from SEMAINE, we process data in a similar way and transform the annotations into seven ratings in arousal and eight ratings in valence.

4.3 Feature extraction

We extract two kinds of facial features, local Gabor binary patterns (LGBP) (Senechal et al., 2012) and scattering transform (ST) (Bruna and Mallat, 2013). In LGBP, the Gabor and LBP filtering operations follow one after the other. Different from the normal LBP whose filter operates only on the original images, the LBP filter in LGBP operates on a number of images which have been filtered by a bank of different Gabor filters. The final LGBP feature histogram of an image is formed by concatenating the histograms composed for each Gabor picture, with histogram blocks in the same manner as for LBP. In our experiments, we use 18 Gabor filters (3 wavelet scales and 6 filter orientations) and apply Uniform LBP (59 patterns) to an image split into 16 (4×4

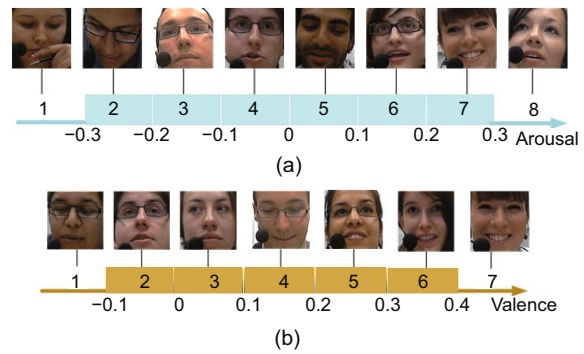


Fig. 3 Presentation of discretized affective ratings and some corresponding face images of the AVEC 2015 benchmarking database in arousal (a) and valence (b)

Table 1 Image numbers of different ratings on different data subsets of the AVEC 2015 benchmarking database in arousal

Data subset	Image number							
	1	2	3	4	5	6	7	8
Training set	5004	4999	6539	8299	8419	8451	8243	4838
Validation set	2576	2868	3763	4042	4040	4492	2660	2580
Testing set	3329	4288	4187	4121	4509	4262	2399	2352

Table 2 Image numbers of different ratings on different data subsets of the AVEC 2015 benchmarking database in valence

Data subset	Image number						
	1	2	3	4	5	6	7
Training set	4465	8197	8492	8111	7923	3979	4680
Validation set	3096	4077	3907	4161	2293	883	2184
Testing set	1572	4443	4803	4346	4930	1927	3004

blocks) local regions. Then we obtain a feature vector of dimensionality $18 \times 16 \times 59 = 16\,992$. For the AVEC 2015 benchmarking database, after performing PCA, which preserves the 97% energy, the feature vectors are reduced to 137-D for valence and 131-D for arousal. For another dataset selected from SEMAINE, the final feature vectors are 102-D for valence and 99-D for arousal. We discretize ratings, preprocess images, and extract features for arousal and valence dimensions independently, so the numbers of samples of these two dimensions are different, causing different dimensionality of feature representation.

For ST, it is implemented by cascading wavelet modulus operators in a deep convolution network, where the signal information is scattered along multiple paths. The coefficients formed in different layers react to details of patterns in different degrees. Thus, the concatenation of coefficients can be used as a good feature representation, which guarantees the translational invariance to a certain extent and is stable to deformations overall. In our experiments, we use wavelets of four scales and eight orientations, and the number of layers of the scattering network is two (i.e., the maximum scattering order is 1). After the scattering operation, we concatenate scattering coefficients to obtain a 25 872-D feature vector for each face image. We also perform PCA which preserves the 99% energy to reduce the dimension of the feature vector on both datasets we use. This results in a 73-D feature vector for valence and a 69-D one for arousal in the AVEC 2015 benchmarking database, and a 71-D feature vector for valence and a 68-D one for arousal in the other one. Finally, we normalize the feature value in each dimension of a feature vector to the range 0 to 1.

4.4 Performance evaluation

To evaluate the effectiveness of our approach called ARR-SVM here, we compare the results with the traditional baseline methods of multi-class classification SVM (C-SVC) and SVM for regression (epsilon-SVR), and deep learning based methods multi-class classification CNN (C-CNN) and deep BLSTM for regression (R-DBLSTM). Note that epsilon-SVR and R-DBLSTM are conducted on the discretized ratings, instead of the original continuous annotations. For C-SVC, we directly use the C-SVC setting in LIBSVM (Chang and Lin, 2011) to train

classifiers on the training set and select the model that provides the highest accuracy on the validation set to test and record the performance on the test set. For epsilon-SVR, the model that obtains the lowest mean squared error (MSE) on the validation set is used to conduct a testing on the test set. In terms of C-CNN, we finetune the age net trained in Levi and Hassner (2015) using the training data and select the model that obtains the highest accuracy on the validation set to perform testing on the test set. For R-DBLSTM, we train models on the training set and select the one that gives the lowest sum of square errors on the validation set to test. For ARR-SVM, we train and obtain $K - 1$ binary classifiers, each of which uses all training samples that are relabeled with binary labels according to the compared results between rating labels and the specific rank in $\{1, 2, \dots, K - 1\}$ during training and performs best on the corresponding validation set. Then we use Eq. (9) to obtain the predicted ratings of test samples.

In terms of parameter selection, for each SVM, no matter whether in C-SVC, epsilon-SVR, or our ARR-SVM, the kernel is always selected as the RBF kernel. We employ the grid search method to find the best combination of parameters C and γ . Parameter C is searched in the range $\{2^{-5}, 2^{-4}, 2^{-3}, 2^{-2}, 2^{-1}, 2^0, 2^1, 2^2, 2^3, 2^4, 2^5, 2^6, 2^7\}$, and parameter γ is selected from a slightly smaller range $\{2^{-5}, 2^{-4}, 2^{-3}, 2^{-2}, 2^{-1}, 2^0, 2^1\}$. Parameter ϵ in epsilon-SVR is selected as 0.1 for our experiments. For C-CNN and R-DBLSTM, the parameters that obtain the best performance on the validation set are selected as optimal parameters.

4.4.1 Performance comparison of CS-1 ($L=1$) and MAE

Table 3 reports the CS-1 ($L=1$) and MAE results of different approaches in arousal. For ARR-SVM, the costs applied to reweight the training samples of the k^{th} cost-sensitive binary classifier are derived from Eq. (4). This selection is motivated by the experimental results of ARR-SVM when using different costs, which will be discussed later. Clearly, our ARR-SVM method with the ST feature provides the best performance in both CS-1 and MAE measures. For CS-1 evaluation, the regression based methods epsilon-SVR and R-DBLSTM obtain poor results at only around

30%. However, for the MAE measure, epsilon-SVR achieves better performance than C-SVC, and R-DBLSTM outperforms C-CNN. This is mainly because the optimization targets of epsilon-SVR and R-DBLSTM are more consistent with the target of reducing the MAE error than those of C-SVC and C-CNN.

Table 3 CS-1 and MAE comparison in arousal in the AVEC 2015 benchmarking database

Learning method	Feature type	CS-1 (%)	MAE
Epsilon-SVR	LGBP	29.42	1.9011
C-SVC	LGBP	38.44	2.2211
ARR-SVM	LGBP	42.11	1.8559
Epsilon-SVR	ST	31.07	1.7922
C-SVC	ST	39.29	2.3596
ARR-SVM	ST	50.83	1.7507
C-CNN	–	44.58	1.9469
R-DBLSTM	ST	32.27	1.7653

Bold numbers denote the best results. CS-1: cumulative score calculated by Eq. (3) when $L=1$; MAE: mean absolute error

Table 4 shows the experimental results of two evaluation indices in valence. Each cost-sensitive binary classifier of ARR-SVM uses the data reweighting costs generated by Eq. (6) where $L=3$. The selection of this cost function is also supported by the experimental results shown in what follows. Note that C-SVC using the ST feature provides the best performance in CS-1 evaluation. This could be due to the fact that the ST feature extracted for the valence affect recognition task is more suitable for classification. Our ARR-SVM approach using the ST feature still obtains the best result in MAE measure. Similar to the results in arousal, regression-based methods tend to obtain worse results than classification-based methods in the CS-1 measure, but provide better performance in MAE evaluation. It is also not difficult to find that the ST feature performs better than the LGBP feature in the affective rating estimation task, and the prediction results obtained in valence are better than those in arousal in the AVEC 2015 benchmarking database.

Table 5 reports the experimental results in arousal on the selected subset of SEMAINE database. C-CNN achieves the best results in both CS-1 and MAE measures. It is probably because C-CNN has learned better feature representation directly from the facial images. Our ARR-SVM

Table 4 CS-1 and MAE comparison in valence in the AVEC 2015 benchmarking database

Learning method	Feature type	CS-1 (%)	MAE
Epsilon-SVR	LGBP	39.91	1.5362
C-SVC	LGBP	58.74	1.5048
ARR-SVM	LGBP	65.89	1.3387
Epsilon-SVR	ST	43.47	1.3840
C-SVC	ST	66.45	1.3854
ARR-SVM	ST	63.84	1.3286
C-CNN	–	54.88	1.6687
R-DBLSTM	ST	47.60	1.3483

Bold numbers denote the best results. CS-1: cumulative score calculated by Eq. (3) when $L=1$; MAE: mean absolute error

method still outperforms baseline methods epsilon-SVR and C-SVC. In addition, R-DBLSTM obtains good results in MAE measure but poor performance in CS-1 evaluation.

Table 5 CS-1 and MAE comparison in arousal in the subset of the SEMAINE database

Learning method	Feature type	CS-1 (%)	MAE
Epsilon-SVR	LGBP	28.93	1.8084
C-SVC	LGBP	29.11	2.7851
ARR-SVM	LGBP	45.79	1.8660
Epsilon-SVR	ST	26.21	1.8576
C-SVC	ST	44.60	2.0549
ARR-SVM	ST	45.23	1.7197
C-CNN	–	53.30	1.5828
R-DBLSTM	ST	29.45	1.6413

Bold numbers denote the best results. CS-1: cumulative score calculated by Eq. (3) when $L=1$; MAE: mean absolute error

Table 6 shows the experimental results in valence. ARR-SVM using the LGBP feature gives the best result in CS-1 measure. Regression-based methods epsilon-SVR and R-DBLSTM provide lower MAE, and C-CNN still obtains a good CS-1 result.

4.4.2 CS comparison at different error tolerance levels

Fig. 4 shows the CS results in arousal and valence dimensions for different approaches, using the ST feature on the AVEC 2015 benchmarking database when the error tolerance level is in the range of $\{0, 1, 2, 3\}$. ARR-SVM outperforms all the other methods consistently at different error tolerance levels in arousal. It suggests that the ordinal information is available and important to improve affect recognition performance in arousal. Note that

Table 6 CS-1 and MAE comparison in valence in the subset of the SEMAINE database

Learning method	Feature type	CS-1 (%)	MAE
Epsilon-SVR	LGBP	35.38	1.7214
C-SVC	LGBP	42.65	2.3794
ARR-SVM	LGBP	46.25	2.0434
Epsilon-SVR	ST	20.98	2.1378
C-SVC	ST	42.88	2.1558
ARR-SVM	ST	44.18	1.9989
C-CNN	-	45.16	2.2341
R-DBLSTM	ST	34.77	1.8238

Bold numbers denote the best results. CS-1: cumulative score calculated by Eq. (3) when $L=1$; MAE: mean absolute error

CS-0 ($L=0$) is equivalent to the common accuracy and for epsilon-SVR and R-DBLSTM, we calculate CS-0.5 for CS-0. For valence, C-SVC obtains competitive results with the ARR-SVM approach and even performs better than ARR-SVM in CS-0 and CS-1 measures. However, ARR-SVM consistently provides higher accuracy than other methods at different error tolerance levels and performs better than C-SVC at higher error tolerance levels.

Fig. 5 presents arousal and valence results on the subset of SEMAINE database. For arousal, C-CNN obtains the highest accuracy at all error tolerance levels. It suggests that C-CNN has learned good feature representation from facial images for arousal rating recognition. Our ranking method ARR-SVM gives comparable results with C-CNN at high error tolerance levels and outperforms other methods at different error tolerance levels. For valence, classification-based methods C-SVC and C-CNN achieve competitive results with ARR-SVM and regression-based methods epsilon-SVR and R-DBLSTM give the worse performance.

4.4.3 Performance comparison when using different cost-sensitive settings

We conduct exhaustive experiments in the AVEC 2015 benchmarking database to compare results of our ARR-SVM method when different cost-sensitive settings are applied. Tables 7 and 8 show the experimental results for arousal and valence, respectively. For arousal results, it can be seen that the absolute cost generated by Eq. (4) produces the best performance on all three kinds of measures by using the ST feature. In terms of the LGBP feature, although we cannot find any kind of

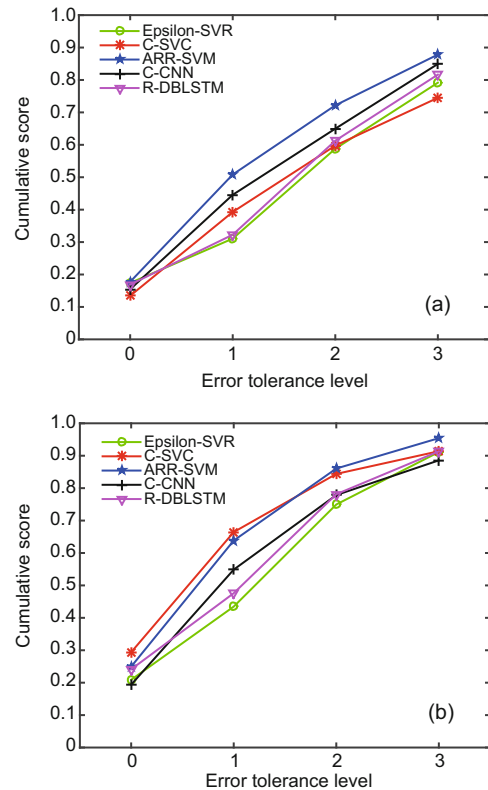


Fig. 4 Cumulative score (CS) comparison at error tolerance levels {0, 1, 2, 3} for different approaches in arousal (a) and valence (b) using the scattering transform (ST) feature on the AVEC 2015 benchmarking database

cost that consistently attains the best results on all measures, we can select the absolute cost that produces the best accuracy and the second best CS-1 and MAE measures as the best cost. What is more, the no-cost setting does not give even one best result on three kinds of measures when using ST and LGBP features. It can validate the importance of using a cost-sensitive setting. Therefore, we can conclude that the absolute cost is more suitable for reweighting instances to train each basic binary classifier in arousal rating ranking.

For valence results in Table 8, we can see that when the ST feature is used, the CSMAE3 cost generated by Eq. (6) where $L=3$ can be viewed as the cost to produce the best performance. It attains the best accuracy and MAE and produces good CS-1 result. The no-cost setting gives the worst MAE and relatively low accuracy and CS-1 measures. It is not difficult to find that the CS style costs tend to produce better CS-1 results. In terms of the LGBP feature, the CSMAE3 cost gives the best MAE, the

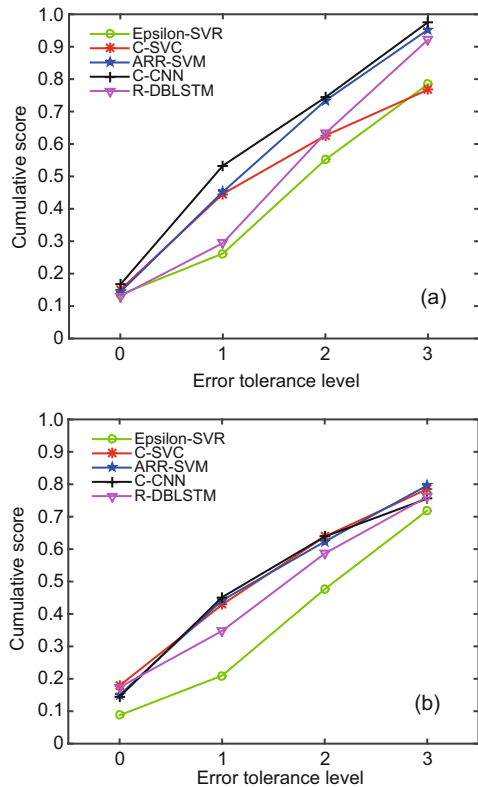


Fig. 5 Cumulative score (CS) comparison at error tolerance levels {0, 1, 2, 3} for different approaches in arousal (a) and valence (b) using the scattering transform (ST) feature on the subset of SEMAINE database

second best accuracy, and good CS-1 result which is very close to the best CS-1 performance. Therefore, similar to the ST feature, the CSMAE3 cost is also the best cost-sensitive setting in valence rating ranking when considering all three kinds of measures. Note that the absolute cost produces the poorest performance, which is even far worse than that given by the no-cost setting on each measure. The discussions above indicate mainly that the appropriate cost-sensitive settings, such as absolute cost for arousal and CSMAE3 for valence, which are induced by experiments, are important for improving the performance of our ranking method.

5 Conclusions and future work

In this paper, we have proposed an ordinal ranking framework, ARR, to solve the affect recognition problem in the A-V space. We first discretize the continuous, real-value annotations in arousal and valence dimensions, respectively, to form finite affect ratings. Then a series of cost-sensitive binary classi-

Table 7 Accuracy, CS-1, and MAE comparison of ARR-SVM when using different cost-sensitive settings and features in arousal

Cost	Feature type	Accuracy (%)	CS-1 (%)	MAE
No cost	LGBP	14.64	41.63	1.9023
CS-1	LGBP	14.73	41.55	1.9052
CS-2	LGBP	14.76	42.12	1.8835
CS-3	LGBP	14.50	41.40	1.9063
CSMAE1	LGBP	14.51	41.84	1.8870
CSMAE2	LGBP	14.45	41.65	1.8876
CSMAE3	LGBP	15.30	41.97	1.8434
Absolute	LGBP	15.33	42.11	1.8559
No cost	ST	17.57	49.10	1.7806
CS-1	ST	16.43	45.92	1.8160
CS-2	ST	16.75	48.24	1.7821
CS-3	ST	16.62	48.59	1.7938
CSMAE1	ST	16.83	49.03	1.8133
CSMAE2	ST	16.61	47.77	1.8158
CSMAE3	ST	16.98	48.48	1.7513
Absolute	ST	17.75	50.83	1.7507

No cost: no cost used for data reweighting; MAE: mean absolute error; CS- n : cost generated by Eq. (5) when $L=n$ ($n=1, 2, 3$); CSMAE n : cost generated by Eq. (6) when $L=n$ ($n=1, 2, 3$); Absolute: cost generated by Eq. (4). Bold numbers denote the best results

Table 8 Accuracy, CS-1, and MAE comparison of ARR-SVM when using different cost-sensitive settings and features in valence

Cost	Feature type	Accuracy (%)	CS-1 (%)	MAE
No cost	LGBP	22.93	65.91	1.3426
CS-1	LGBP	24.22	64.34	1.3969
CS-2	LGBP	23.88	64.56	1.3776
CS-3	LGBP	23.44	65.98	1.3402
CSMAE1	LGBP	23.66	63.38	1.4077
CSMAE2	LGBP	23.88	64.31	1.3785
CSMAE3	LGBP	23.91	65.89	1.3387
Absolute	LGBP	22.13	60.05	1.4607
No cost	ST	23.01	62.69	1.4035
CS-1	ST	23.72	64.08	1.3426
CS-2	ST	23.58	62.64	1.3528
CS-3	ST	23.60	64.26	1.3320
CSMAE1	ST	22.73	62.11	1.3642
CSMAE2	ST	23.88	62.49	1.3510
CSMAE3	ST	24.96	63.84	1.3286
Absolute	ST	23.96	63.24	1.3473

No cost: no cost used for data reweighting; MAE: mean absolute error; CS- n : cost generated by Eq. (5) when $L=n$ ($n=1, 2, 3$); CSMAE n : cost generated by Eq. (6) when $L=n$ ($n=1, 2, 3$); Absolute: cost generated by Eq. (4). Bold numbers denote the best results

fiers are trained using all samples relabeled according to the comparison results between corresponding ratings and a given rank of a binary classifier. We finally obtain the affect rating by aggregating the

results of basic binary classifiers. We compare CS and MAE evaluations of our ranking method with baseline and deep learning based classification and regression methods. Experimental results show that our proposed approach can produce effective results in affect recognition in the A-V space. The ordinal property of the annotations in the dimensional affect space should be appropriately used to enhance recognition performance.

In the future, we will explore the pairwise ranking framework and search for the feature representation that is more suitable for ranking methods in affect recognition in the A-V space.

References

- Abousaleh F, Lim T, Cheng W, et al., 2016. A novel comparative deep learning framework for facial age estimation. *EURASIP J Image Video Process*, 2016(1):47. <https://doi.org/10.1186/s13640-016-0151-4>
- Baron-Cohen S, 2004. *Mind Reading: the Interactive Guide to Emotions*. Jessica Kingsley Publishers.
- Bruna J, Mallat S, 2013. Invariant scattering convolution networks. *IEEE Trans Patt Anal Mach Intell*, 35(8):1872-1886. <https://doi.org/10.1109/TPAMI.2012.230>
- Caridakis G, Malatesta L, Kessous L, et al., 2006. Modeling naturalistic affective states via facial and vocal expressions recognition. 8th Int Conf on Multim Interfaces, p.146-154. <https://doi.org/10.1145/1180995.1181029>
- Chang C, Lin C, 2011. LIBSVM: a library for support vector machines. *ACM Trans Intell Syst Technol*, 2(3):27. <https://doi.org/10.1145/1961189.1961199>
- Chang K, Chen C, 2015. A learning framework for age rank estimation based on face images with scattering transform. *IEEE Trans Image Process*, 24(3):785-798. <https://doi.org/10.1109/TIP.2014.2387379>
- Chang K, Chen C, Hung Y, 2010. A ranking approach for human ages estimation based on face images. 20th Int Conf on Pattern Recognition, p.3396-3399. <https://doi.org/10.1109/ICPR.2010.829>
- Feng S, Lang C, Feng J, et al., 2017. Human facial age estimation by cost-sensitive label ranking and trace norm regularization. *IEEE Trans Multim*, 19(1):136-148. <https://doi.org/10.1109/TMM.2016.2608786>
- Geng X, Zhou Z, Smith-Miles K, 2007. Automatic age estimation based on facial aging patterns. *IEEE Trans Patt Anal Mach Intell*, 29(12):2234-2240. <https://doi.org/10.1109/TPAMI.2007.70733>
- Glowinski D, Camurri A, Volpe G, et al., 2008. Technique for automatic emotion recognition by body gesture analysis. Int Conf on Computer Vision and Pattern Recognition Workshops, p.1-6. <https://doi.org/10.1109/CVPRW.2008.4563173>
- Gunes H, Pantic M, 2010. Automatic, dimensional and continuous emotion recognition. *Int J Synth Emot*, 1(1):68-99. <https://doi.org/10.4018/jse.2010101605>
- He L, Jiang D, Yang L, et al., 2015. Multimodal affective dimension prediction using deep bidirectional long short-term memory recurrent neural networks. 5th Int Workshop on Audio/Visual Emotion Challenge, p.73-80. <https://doi.org/10.1145/2808196.2811641>
- Ioannou S, Raouzaoui A, Tzouvaras V, et al., 2005. Emotion recognition through facial expression analysis based on a neurofuzzy network. *Neur Netw*, 18(4):423-435. <https://doi.org/10.1016/j.neunet.2005.03.004>
- Joachims T, 2002. Optimizing search engines using click-through data. 8th ACM Int Conf on Knowledge Discovery and Data Mining, p.133-142. <https://doi.org/10.1145/775047.775067>
- Levi G, Hassnecr T, 2015. Age and gender classification using convolutional neural networks. Int Conf on Computer Vision and Pattern Recognition Workshops, p.34-42. <https://doi.org/10.1109/CVPRW.2015.7301352>
- Li L, Lin H, 2006. Ordinal regression by extended binary classification. *Advances in Neural Information Processing Systems*, p.865-872.
- Lim T, Hua K, Wang H, et al., 2015. VRank: voting system on ranking model for human age estimation. 17th IEEE Int Workshop on Multimedia Signal Processing, p.1-6. <https://doi.org/10.1109/MMSP.2015.7340789>
- Liu T, 2011. *Learning to Rank for Information Retrieval*. Springer-Verlag Berlin Heidelberg.
- Martinez H, Yannakakis G, Hallam J, 2014. Don't classify ratings of affect; rank them! *IEEE Trans Affect Comput*, 5(3):314-326. <https://doi.org/10.1109/TAFFC.2014.2352268>
- McDuff D, El Kaliouby R, Kassam K, et al., 2010. Affect valence inference from facial action unit spectrograms. Int Conf on Computer Vision and Pattern Recognition Workshops, p.17-24. <https://doi.org/10.1109/CVPRW.2010.5543833>
- Nicolaou M, Gunes H, Pantic M, 2010. Audio-visual classification and fusion of spontaneous affective data in likelihood space. 20th Int Conf on Pattern Recognition, p.3695-3699. <https://doi.org/10.1109/ICPR.2010.900>
- Nicolaou M, Gunes H, Pantic M, 2011. Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. *IEEE Trans Affect Comput*, 2(2):92-105. <https://doi.org/10.1109/T-AFFC.2011.9>
- Ringeval F, Sonderegger A, Sauer J, et al., 2013. Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. 10th IEEE Int Conf on Automatic Face and Gesture Recognition Workshops, p.1-8. <https://doi.org/10.1109/FG.2013.6553805>
- Ringeval F, Schuller B, Valstar M, et al., 2015. AVEC 2015: the 5th International Audio/Visual Emotion Challenge and Workshop. 23rd ACM Int Conf on Multimedia, p.1335-1336. <https://doi.org/10.1145/2733373.2806408>
- Russell J, 1980. A circumplex model of affect. *J Pers Soc Psychol*, 39(6):1161-1178. <https://doi.org/10.1037/h0077714>
- Scherer K, 2000. Psychological models of emotion. In: Borod J (Ed.), *The Neuropsychology of Emotion*. Oxford University Press, New York, USA.
- Scherer K, Schorr A, Johnstone T, 2001. *Appraisal Processes in Emotion: Theory, Methods, Research*. Oxford University Press, New York, USA.
- Schuller B, Vlasenko B, Eyben F, et al., 2009. Acoustic emotion recognition: a benchmark comparison of performances. IEEE Workshop on Automatic Speech Recognition and Understanding, p.552-557. <https://doi.org/10.1109/ASRU.2009.5372886>

- Senechal T, Rapp V, Salam H, et al., 2012. Facial action recognition combining heterogeneous features via multikernel learning. *IEEE Trans Syst Man Cybern Part B (Cybern)*, 42(4):993-1005.
<https://doi.org/10.1109/TSMCB.2012.2193567>
- Wöllmer M, Eyben F, Reiter S, et al., 2008. Abandoning emotion classes—towards continuous emotion recognition with modelling of long-range dependencies. *Interspeech*, p.597-600.
- Wöllmer M, Metallinou A, Eyben F, et al., 2010a. Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional LSTM modeling. *Interspeech*, p.2362-2365.
- Wöllmer M, Schuller B, Eyben F, et al., 2010b. Combining long short-term memory and dynamic Bayesian networks for incremental emotion-sensitive artificial listening. *IEEE J Sel Top Signal Process*, 4(5):867-881.
<https://doi.org/10.1109/JSTSP.2010.2057200>
- Xu J, Li H, 2007. AdaRank: a boosting algorithm for information retrieval. 30th ACM Int Conf on Research and Development in Information Retrieval, p.391-398.
<https://doi.org/10.1145/1277741.1277809>
- Yang Y, Chen H, 2011. Ranking-based emotion recognition for music organization and retrieval. *IEEE Trans Audio Speech Lang Process*, 19(4):762-774.
<https://doi.org/10.1109/TASL.2010.2064164>
- Yu C, Aoki P, Woodruff A, 2004. Detecting user engagement in everyday conversations. 8th Int Conf on Spoken Language Processing, p.1329-1332.