Audio-Visual Emotion Recognition with Preference Learning based on Intended and Multi-modal Perceived Labels

Yuanyuan Lei, *Member, IEEE,* and Houwei Cao, *Member, IEEE*

Abstract—This paper introduces a novel preference learning framework that simultaneously considers both the intended and the perceived labels while addressing the mismatches between them. Based on analyzing the discrepancies and agreements between the intended and the perceived labels in different modalities of audio-only, visual-only, and audio-visual, as well as the consistency among the perceptual ratings of all raters, we propose three sets of pair-wise ranking rules to generate multi-scale relevant scores for preference learning, scaling from sketchy manner to detailed manner. Three ranking models with support vector machine, deep neural networks, and gradient boosting decision trees are developed. Our results demonstrate that all three preference learning models significantly outperform the conventional classifiers baselines, and the LambdaMART model with gradient boosting decision trees achieves the best performance. The improvement from the preference learning models confirm the benefits of complementary information provided by different types of labels. We also observe additional improvement from the detailed 'complex ranking rules', particular with the best LambdaMART model, which suggests that we should treat intended and perceived labels in single-model & multi-modal differently. We further discuss the complementary of different ranking models, and obtain the best overall accuracy of 85.06% on CREMA-D dataset when combining the two best ranking models–LambdaMART and RankNet–together, which is significantly better than the 76.19% accuracy attained by the baseline models. Finally, we perform the cross-corpus emotion recognition experiments by training emotion rankers on CREMA-D and tested the ranking-based emotion classifier on the SAVEE dataset that do not have perceived labels annotated.

Index Terms—multimodal emotion recognition, preference learning, intended label, perceived label, RankSVM, RankNet, LambdaMART.

1 Introduction

Emotions are essential to human life. They directly influence human perception and behaviors, and have big impacts on our daily tasks, such as learning, social interaction, and rational decision-making. Automatic emotion recognition has found applications in many domains, including multimedia retrieval, social media analysis, human-computer and human-robot interaction, health care, etc. The emotional states of people vary over the course of conversations and these changes are expressed externally through a variety of channels, including facial expressions, voice, spoken words, body gestures, etc.

The developed approaches for emotion recognition are mostly *unimodal*, that is, the analysis is usually based on one modality out of facial expression, voice, text, etc. Although the recent advances in automatic emotion recognition through cues in individual modalities have been remarkable, emotion recognition is far from being a solved problem. Firstly, human emotion expressions are subtle and can be conveyed by a combination of several emotions. Emotions are expressed differently through different verbal and non-verbal channels, such as text, facial expression and voice. There exists a large inter-subject variability in emotion expression and perception. In our preliminary perception study on the CREMA-D database [1], we studied how people perceive acted emotion differently across the three modalities of audio-only, video-only, and audio-visual. We demonstrated that the emotion expressions in different channels are different, and the expressions in each channel can be perceived as mixtures of emotions. Multi-modal emotion recognition that simultaneously analyzes information from all modalities have a great potential to further improve the accuracy of unimodal analysis [2], [3].

Meanwhile, a major challenge for emotion recognition is the inconsistency between emotion labels, especially for multimodal emotion data. In supervised machine learning, the learned models can be considerably affected by the training data and the labels assigned to them. For emotion recognition tasks, due to the complex nature of emotions, obtaining the ground-truth labels describing the emotional content of a data sample is challenging. Many existing datasets only provide labels of the intended emotion, which is the target emotion during the emotion production/expression. Those labels may not precisely reveal the underlying emotion of a given recording. To address that, perceptual evaluations are conducted to annotate the perceived emotion in many studies, where each data sample is often annotated by multiple raters, and the ground-truth label is the consensus label obtained by some aggregation method, e.g., majority vote. However, the perception of emotion is subjective and different channels have different emotion expressiveness. As a result, for the same recording, more than one emotion can

- *Yuanyuan Lei is with the Department of Computer Science, Texas A&M University, TX 77840, USA. Email: yuanyuan@tamu.edu*
- *Houwei Cao is with the Department of Computer Science, New York Institute of Technology, NY 10023, USA. Email: hcao02@nyit.edu*

Manuscript received October 26, 2021

be perceived by different raters on different channels. The inter-rater and inter-channel agreement can be very low for some cases and the corresponding perceptual labels can be very inconsistent. How to learn with those inconsistent and subjective labels is still an open research question.

One common approach is to discard data without consensus label or with low inter-rater agreement. For example, authors of [4] have shown that emotion recognition can be improved considerably by taking into account annotator agreement and training the model on smaller but reliable dataset. In contrast to previous studies that relied on either intended or perceived emotion labels, we argue that both intended and perceived labels contain valuable information about how humans express and perceive emotions, which should be collectively mined to improve the accuracy of emotion recognition.

In this paper, we develop novel preference learning models for multimodal emotion recognition that address the mismatch between the perceived labels on different modalities, as well as the mismatch between the intended and perceived emotions in each modality. Our preference learning models consider the subjective nature of emotion and exploit the evaluations from multiple raters to better assess the underlying ordinal, mixture representation of emotion. They sort all data in a given sample with respect to the degree with which they convey a target emotion. Towards this goal, we make the following contributions.

1) We propose to calculate the relevance of each utterance to each target emotion based on the matchings between the target emotion and the intended and perceived emotion labels of the utterance. We develop three sets of ranking rules: a simple ranking rule that assigns the same weight to labels from all modalities, an intermediate rule that assigns larger weight to multi-modal labels than single-modal labels, and a complex ranking rule that assigns different weights to labels from different modalities.
2) For each target emotion, we train an emotion ranker using three *pairwise learning to rank models*, namely support vector machine, deep neural networks, and gradient boosting decision trees, to minimize the number of incorrectly ordered utterance pairs generated by the same speaker. The outputs of the emotion ranker are used to sort all utterances based on their relevances to the target emotion.
3) We further develop ranking-based multi-class emotion classifiers based on the results obtained from the six emotion rankers. The final emotion prediction is generated using the highest rank rule, second-level training, or model combination to take advantage of the complementary information of different labels, ranking rules and ranking models.
4) We conduct extensive evaluation of the proposed multimodal ranking rules and preference learning models on the CREMA-D dataset. Our results demonstrate that all three preference learning models significantly outperform the conventional classifiers baselines, and combining different ranking models can further improve the accuracy. Our results confirm the benefits of complementary information provided by different labels, and suggest that we should treat different labels differently.
5) We perform the cross-corpus emotion recognition experiments by training emotion rankers on CREMA-D and tested the ranking-based emotion classifier on the SAVEE dataset that do not have perceived labels annotated. Our results show that the ranking-based classifiers outperform the conventional supervised method by larger margin when the training and testing set are from different corpus, which further prove the effectiveness and generalization ability of the proposed ranking models.

The rest of the paper is organized as follows. We review the related work in Section 2. The dataset and features are introduced in Section 3. The preference learning framework is developed in Section 4, in the order of ranking rules for generating relevance scores, emotion rankers based on pairwise learning to rank models, and the final generation of multi-class prediction. The evaluation results and the cross-corpus experiments are discussed in Section 5. The paper is concluded in Section 6.

2 RELATED WORK

Most of the existing approaches to automatic human emotion analysis are aimed at the recognition of a small number of prototypical (basic) expressions of emotion, and have been trained and tested on posed or acted affective expressions [5]. Research on emotion recognition from cues expressed in facial expression has a long-standing tradition. Numerous prior studies followed Ekman's basic discrete emotion theory [6] and concentrated on emotion perception from facial cues. They have established that prototypical basic emotions can be universally recognized by different groups of people based on the activation of specific facial expressions [7]. There are three widely adopted approaches to the extraction of information related to facial expression: detection and tracking of facial feature points [8], fitting face models to characterize shape and/or appearance, such as active appearance models (AAMs) [9], and image analysis by basic functions, such as Gabor wavelets [10], or by texture descriptors, such as local binary patterns (LBP) [11]. The traditional paradigm of emotion recognition in speech is to extract acoustic features from the speech signal, then train classifiers on these representations, which when applied to a new utterance are able to determine its emotion content. Many acoustic features, such as prosodic features (e.g., pitch, energy, duration), spectral features (e.g., Mel-frequency cepstral coefficients (MFCC), Linear Prediction Cepstrum Coefficients (LPCC)), voice quality features (e.g. jitter and shimmer), etc. contribute to the transmission of emotional content in voice [12]. A variety of pattern recognition methods have been explored for automatic emotion recognition, such as Gaussian mixture models, hidden Markov models, support vector machines, regression, and neural networks [13]. In the recent years, the successes of deep neural networks (DNN) have spanned many domains, from speech recognition, machine translation, to computer vision. Recently, DNNs have been successfully applied in affective computing tasks, including end-to-end training,

and learning of discriminative features for speech and facial emotion recognition [14], [15].

Meanwhile, researchers have recently devoted more efforts to multimodal emotion analysis and recognition, in the hope that combining different information sources would lead to more accurate recognition. It is well-known that facial expressions convey more information about a subject's emotional state than changes in voice, which typically convey arousal [16], [17], [18]. Many studies investigate how different modalities is integrated, and the most common approaches are naive feature-level fusion (in which all features are combined together to learn a classifier) and decision-level fusion (in which a classifier for each modality is trained separately and their predictions are combined by rules) [19], [20]. However, learning from multiple modalities is not trivial. Different modalities may have different degrees of expressiveness for different emotions. Different expressions, which can be redundant or independent, complementary or contradictory, bring many challenges to the multimodal emotion recognition task [21], [22]. Many existing multimodal emotion recognition systems that applied the state-of-the-art feature-based and model-based fusion have only shown little or no improvement over unimodal systems. The improvements from multimodal fusion are much smaller on datasets of spontaneous emotions than those with acted emotions [23]. Savran et al. proposed an advanced multimodal emotion recognition system, which had shown great promise in recognition performance by combining textual, audio, and video modalities [21]. This was achieved by using a powerful temporal prediction model as the prior in Bayesian fusion as well as by incorporating uncertainties about the unimodal predictions [22].

Although preference learning frameworks have been widely used in many information retrieval applications from different types of data, e.g. in text, image, video, and music [24], [25], [26], [27], they have been applied to speech emotion recognition only very recently, e.g. [28], [29], [30], [31], [32], [33], [34], [35]. Cao et al. [33] first trained rankers by establishing a binary preference score based on the consensus labels. For example, for a ranker for *Anger* emotion, a sample labeled as *Anger* was always preferred over another sample labeled with other emotions. Lotfian et al. [31] considered the perceptual ratings from all individual evaluators to create a continuous relevance score for preference learning. Parthasarathy et al. [32] first applied qualitative agreement (QA) methods to estimate reliable labels from inconsistent annotations, and then used those labels for preference-learning. Jin et al. [36] developed a preference learning framework for speech emotion recognition within audio modality only. Different from previous works, we focused on developing the preference learning framework to more challenging multimodal emotion recognition, where we consider the discrepancies and consistencies between the intended and perceived labels in different modalities of audio-only, visual-only, and audio-visual. We assume that samples that have the consistent target emotion during the emotion expression and perception are more likely to convey the target emotion. Based on analyzing the agreement between the intended and perceived emotions, as well as the consistency among all perceptual ratings, we proposed various pairwise ranking rules to generate multi-scale relevant scores for preference learning. Our results demonstrated that the ranking rule treating different labels differently can outperform the ranking rule treating different labels equally. Also, this work considers multimodal analysis, and the discrepancy between labels obtained from different modalities was additionally addressed. Moreover, we further explore more advanced preference learning models including ranking with deep neural networks, and ranking with boosted decision trees. Finally, we combine the ranking scores of individual emotions for multi-class classification, and discuss the potential for improving prediction by combining various ranking and classification systems together.

3 Dataset and Features

3.1 CREMA-D

The dataset we use is *Crowd-sourced Emotional Multimodal Actors Dataset* (CREMA-D) [1], which is an audiovisual corpus collected to explore human emotion expression and perception behaviors in different modalities. It consists of facial and vocal emotional expressions in sentences spoken in a range of basic emotional states (Anger, Disgust, Fear, Happiness, Neutral, and Sadness). This corpus consists of 7, 442 clips (over 10 hours) of emotional sentences collected from 91 actors with diverse ethnic backgrounds. The task for the actors was to convey that they are experiencing a target emotion while uttering a given sentence. The intended emotion label is the target emotion given to the actors during recording. The categorical emotion labels and real-valued intensity values for the perceived emotion were also collected through crowd-sourced perceptual evaluations from 2, 443 raters in three modalities: audio, visual, and audiovisual. More than 95 percent of the clips in the database have 8 to 12 perceptual ratings. Thus for each clip, we have four dimensional emotion labels, including the three crowd-sourced perceived emotion labels which reflect human perception behaviors in different modalities, and the intended emotion label which is the target emotion the actor originally wanted to express.

TABLE 1
Percentage of samples for which the perceived emotion matches the intended emotion from each emotion category in CREMA-D dataset

Matching %	ANG	DIS	FEA	HAP	NEU	SAD	ALL
Audio	60.6	30	32	26	95.7	16.4	41.6
Video	65.4	63.4	51.6	95.6	91.8	33.4	64.3
Multi	74.7	74.4	64.8	94.8	95.7	32.3	72.2

Table 1 shows the percentage of samples for which the perceived emotion matches the intended emotion from each emotion category. It can be seen that there is a considerable mismatch between them due to the complex nature of emotions. The matching ratios between the intended and perceived emotion labels vary from 16.4% to 95.7% across different target emotions and different modalities. Specifically, we can see that for Anger, Disgust, and Fear, multi-modality perceived labels have more matches than the video perceived labels, and these two types of perceived labels both have more matches than the audio perceived labels. For Happy and Sad emotions, the matching percentages of video perceived labels and multi-modality perceived

TABLE 2
The descriptions and the corresponding facial muscles of the 17 selected action units.

Action Unit Number	Description	Facial Muscle
Action Unit 1	Inner brow raiser	Frontalis (Pars medialis)
Action Unit 2	Outer brow raiser	Frontalis (Pars lateralis)
Action Unit 4	Brow lowerer	Depressor glabellae, Depressor supercilli, Currugator supercilli
Action Unit 5	Upper lid raiser	Levator palpebrae superioris, Superior tarsal muscle
Action Unit 6	Cheek raiser	Orbicularis oculi (Pars orbitalis)
Action Unit 7	Lid tightener	Orbicularis oculi (Pars palpebralis)
Action Unit 9	Nose wrinkler	Levator labii superioris alaquae nasi
Action Unit 10	Upper lip raiser	Levator labii superioris, Caput infraorbitalis
Action Unit 12	Lip corner puller	Zygomatic major
Action Unit 14	Dimpler	Buccinator
Action Unit 15	Lip corner depressor	Depressor anguli oris (Triangularis)
Action Unit 17	Chin raiser	Mentalis
Action Unit 20	Lip stretcher	Risorius
Action Unit 23	Lip tightener	Orbicularis oris
Action Unit 25	Lips part	Depressor labii inferioris, Relaxation of mentalis, Orbicularis oris
Action Unit 26	Jaw drop	Masetter, Relaxation of temporalis and internal pterygoid
Action Unit 45	Blink	Relaxation of levator palpebrae, Contraction of orbicularis oculi (Pars palpebralis)

labels are similar, both higher than audio perceived labels. For Neutral, however, the three types of perceived labels have similar matching percentages. Overall, multi-modality perceived labels are more reliable than single-modality perceived label, and video perceived labels are more reliable than audio perceived labels.

In this study, different from the previous studies that relied on either the intended or the perceived emotion labels, we consider both of them. With the agreement and disagreement information of the intended label and three modalities perceived labels, we can model the consistency and inconsistency between human emotion expression and perception in different modalities, and further develop comprehensive emotion ranking rules for preference learning.

3.2 Multi-Modality Features

For each audiovisual clip in the CREMA-D dataset, we first extract the utterance-level acoustic features from audio channel, and facial features from video channel respectively, and then combine them together as our multi-modality features for the development of multimodal emotion rankers in preference learning.

3.2.1 Audio Acoustic Features

We used OpenSMILE toolkit [37] to extract *emobase* acoustic feature set which are wildely used for emotion recognition. *Emobase* feature set contains 52 Low Level Descriptors (LLDs) such as MFCCs, voice quality, fundamental frequency (F0), F0 envelope, LSP and intensity features along with their first and second order derivativesce. In addition, 19 High Level Statistical Functionals (min, max, range, argmin, argmax, mean, standard deviation, three quartile values, three inter-quartile range values, skewness, kurtosis, and intercept, slope, linear error, quadratic error in linear regression) are applied to the LLDs at the utterance level, resulting in a total number of 988 features.

3.2.2 Video Facial Features

Facial Action Units (AUs) are characterized by contractions of specific facial muscles that correspond to a displayed emotion. They have been widely used as features in facial expression analysis and emotion recognition [7], [8], [10]. In this study, we select 17 AUs that are commonly involved in the coding of the six basic emotions. The descriptions and the corresponding facial muscle of these 17 selected action units are listed in Table 2. Similar as acoustic analysis, we also extract the utterance-level AU features. Specifically, for each video frame, we first estimate the intensity of the selected 17 AUs by using the OpenFace facial behavior analysis toolkit [38], [39]. Those are the LLDs for video features. After that, we estimate 22 High Level Statistical Functionals (the 19 statistical functions used in the above acoustic analysis, and R-square, p-value, standard error of estimated slope in linear regression) at the utterance level. Thus, for each video clip, the utterance-level facial action unit features lie in the space of 374 dimensions.

3.2.3 Multi-Modality Features

Finally, we combine the 988 audio acoustic features and 374 video facial action unit features together as our multi-modality features. Z-score normalization has been applied across the entire dataset: each feature minus its mean and divided by its standard deviation. Particularly, if the standard deviation of a certain feature is zero, which means all the 7442 clips have the same value in this feature, then we just eliminate this useless feature. For example, the No. 1018 feature which represents the minimum value of AU20 (Lip stretcher) in all the 7442 clips equal to the same value 0, then this feature is supposed to be useless in differentiating emotions and should be removed. After the z-score normalization and feature elimination, the dimension of the multi-modality features reduces from 1362 to 1347.

4 METHODOLOGY

In this study, we are interested in building rank-based classifiers for emotion recognition. The ranking problem is to sort the utterances with respect to how much they convey a particular emotion. To train a ranker for a target emotion, we need to specify a set of pairs of instances so that, within each pair, one instance conveys the target emotion better than the other. The optimization problem of the ranker is to minimize the number of incorrectly ranked pairs. In this section,

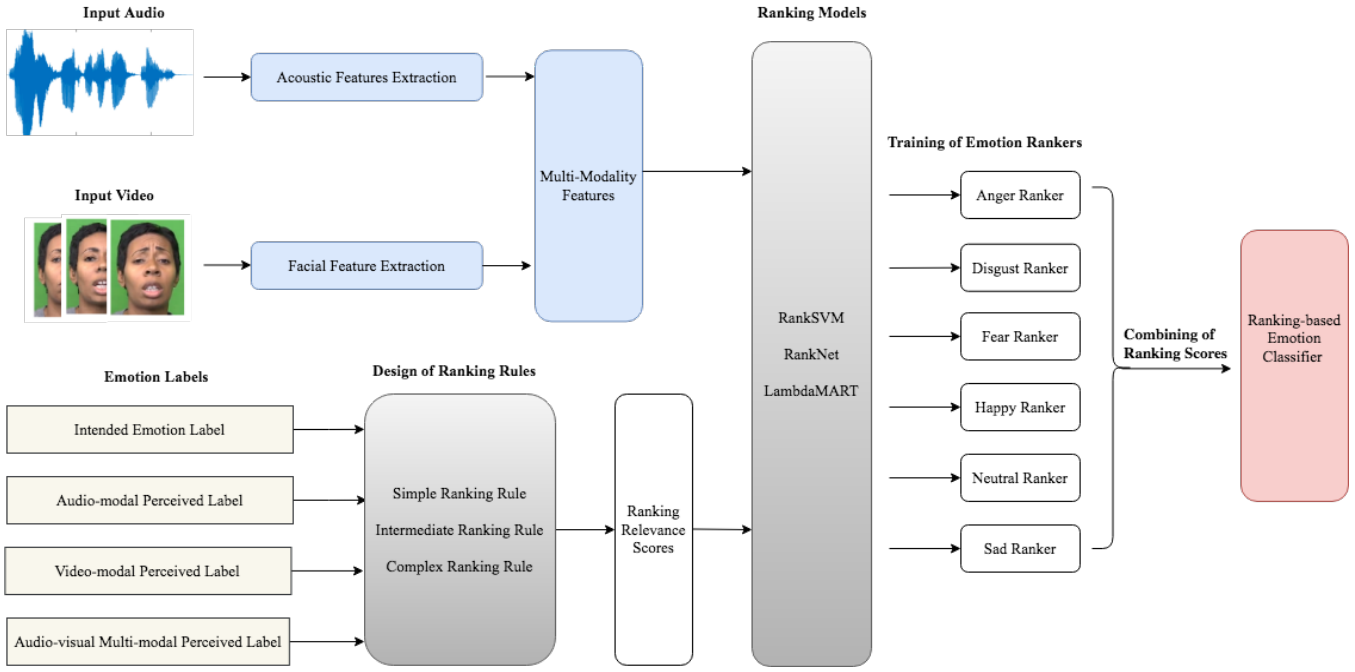

Fig. 1. The diagram of our ranking-based emotion classifier, including modules for features extraction, three ranking rules based on intended and multi-modal perceived labels, three ranking models, and the integrated ranking-based emotion classifier.

we first define various ranking rules by creating ranking relevance score for each observation based on the consensus among the intended emotion label and three perceived labels. Then we introduce three advanced ranking models that we use to build rankers for emotion analysis. Lastly, we discuss multi-class emotion recognition by integrating different ranking-based classifiers together. Figure 1 illustrates the diagram of our ranking-based emotion classifier, including the modules for features extraction, three ranking rules based on intended and multi-modal perceived labels, three ranking models, and integrating emotion rankers into ranking-based classifier.

4.1 Data Denotation

We first summarize the data denotation here: our data can be denoted as

$$\{x_i, s_i, I_i, P_i^{(a)}, P_i^{(v)}, P_i^{(m)}\}, i = 1 \cdots m,$$

where $m = 7442$ is the number of observations, $x_i \in R^d$ is d-dimensional multi-modality features, $d = 1347$ as stated above, $s_i \in \{1 \cdots 91\}$ is the speaker index, I_i is the intended emotion label, $P_i^{(a)}$ is the perceived label in audio single-modality, $P_i^{(v)}$ is the perceived label in video single-modality, $P_i^{(m)}$ is the perceived label in audiovisual multi-modality. Intended label I_i can be six values: `Anger`, `Disgust`, `Fear`, `Happy`, `Neutral` and `Sad`. In addition to the six basic emotion values, three perceived labels $P_i^{(a)}, P_i^{(v)}, P_i^{(m)}$ can also be `Ambiguous` because they are crowd-sourced vote results from 2,443 raters.

4.2 Ranking Rules based on Multi-dimensional Labels

The key point in developing ranking rules is how to determine whether an observation conveys the target emotion more intensely than another. In this section, we define ranking rules by creating ranking relevance score based on the agreement and disagreement information of the intended label and three-dimensional perceived labels. In each target emotion ranker, the higher ranking relevance score means the expression is more likely to convey the target emotion. For example, in the Anger ranker, the ranking relevance score of observation A is higher than observation B, means the speaker conveys more clear anger in utterance A than in utterance B. In mathematical representation, for a target emotion t, ranking rule is actually a mapping from four-dimensional labels $(I_i, P_i^{(a)}, P_i^{(v)}, P_i^{(m)})$ to a ranking relevance score $r_{i,t} \in R$, and after mapping, our data become $\{x_i, s_i, r_{i,t}\}, i = 1 \cdots m$. The design of mapping function is based on counting the influence factors of the four labels into ranking relevance score. Here, we developed three kinds of ranking rules by employing two different sets of weights in counting.

4.2.1 *Simple Ranking Rule*

We first propose **"simple ranking rule"** with the idea that the four emotion labels are of the same importance in identifying emotion class. The simple ranking rule considers the four label of the same weight in mapping ranking relevance score. To be detailed, in the target emotion ranker:

$$r_{i,t} = w_i^{(I)} r_{i,t}^{(I)} + w_i^{(a)} r_{i,t}^{(a)} + w_i^{(v)} r_{i,t}^{(v)} + w_i^{(m)} r_{i,t}^{(m)} \quad (1)$$

where $r_{i,t}^{(*)}$ is the indexes showing whether the corresponding label of sample i matches with the target emotion t. For example, in the Anger ranker, if $I_i = Anger$ then $r_{i,A}^{(I)} = 1$ otherwise $r_{i,A}^{(I)} = 0$, same for the other three labels and their indexes values. Although the three perceived labels can assume `Ambiguous` value other than the six basic emotions,

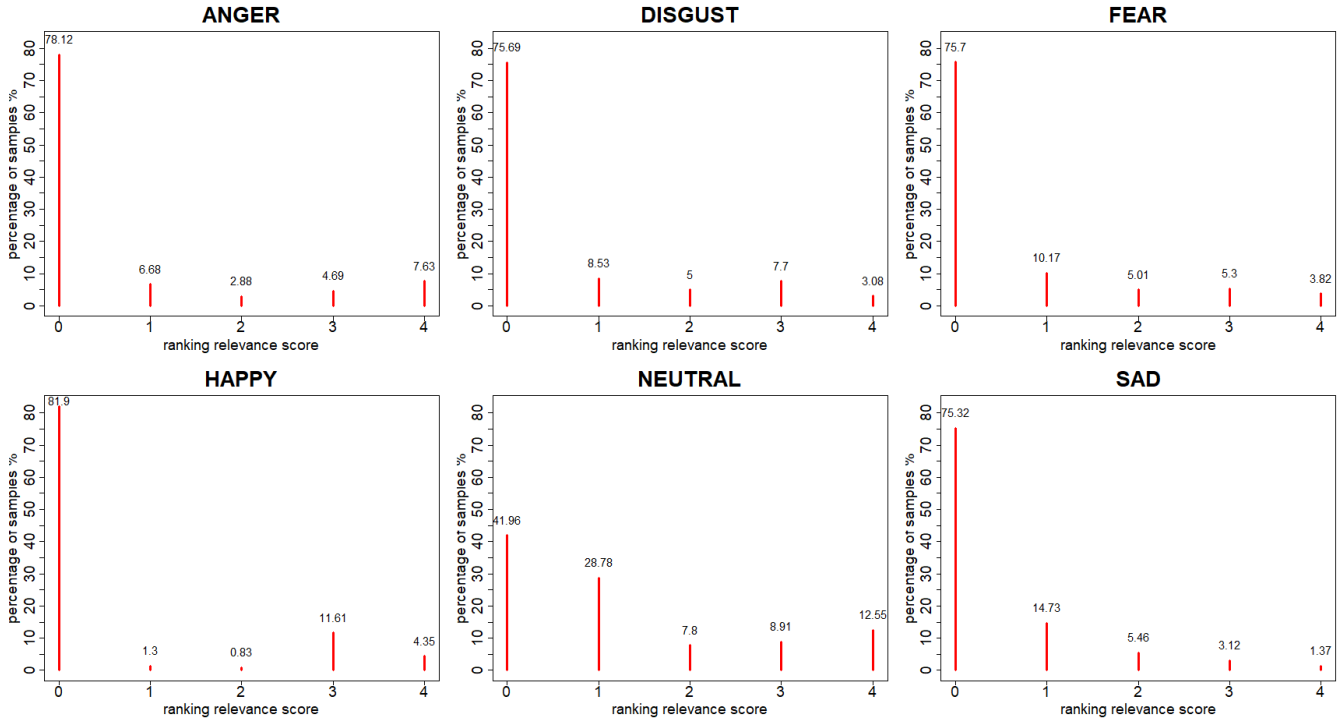

Fig. 2. Histogram of ranking relevance scores of Simple Ranking Rule on the CREMA-D dataset

there are only 5.6% of the whole data with `Ambiguous` perceived labels, and we just set their index value to zero for any target emotion in simple ranking rule.

In this way, the influence factor of a certain label is represented as its weight multiplied by its index value, and ranking rule is the mapping based on adding the influence factors of the four labels together into ranking relevance score. What's more, simple ranking rule uses the same set of weights for all i:

$$w_i = (w_i^{(I)}, w_i^{(a)}, w_i^{(v)}, w_i^{(m)}) = (1, 1, 1, 1) \tag{2}$$

In other words, the simple ranking rule assumes the four labels are equally important for counting influence factors. Intuitively, the samples with four labels all consistently matching with the target emotion are the most clear in conveying the target emotion. The fewer the number of the labels matching with the target emotion, the lower the value of ranking relevance score, the less clear in conveying the target emotion.

Figure 2 shows the histograms of the ranking relevance scores generated by simple ranking rule for the six emotions. The horizontal axis represents the ranking relevance score, which is the number of labels agree with the target emotion in simple ranking rule, and value 0 means none of the four labels agree with this target emotion. We can see that for Anger, Disgust, and Fear, they have similar distributions and the non-zero ranking relevance scores distributed quite evenly. For Happy emotion, samples with non-zero ranking scores mainly fall in the level 3 & 4, meaning the majority of them have high consensus with at least 3 labels matching with the target emotion. For Neutral and Sad, however, the distribution of samples with non-zero ranking score is skewed towards level 1, meaning there is a high degree of disagreement among the four labels. As a result, we can expected that it is easier for the system to distinguish Happy emotion than to recognize Neutral and Sad.

4.2.2 *Intermediate Ranking Rule*

We further extend the simple ranking rule into **"intermediate ranking rule"** by considering different importance on single-modal perceived labels and multi-modal perceived label. As illustrated in Table 1, in the CREMA-D database, the matching ratio between the intended and multi-modal perceived labels is significantly higher than the matching ratios with different single-modal perceived labels. Thus, we consider to assign higher weight to multi-modal label than single-modal labels. To be detailed, intermediate ranking rule also uses (1) to add the four labels' influence factors together as ranking relevance score, but it utilizes a different weight set for all i:

$$w_i = (w_i^{(I)}, w_i^{(a)}, w_i^{(v)}, w_i^{(m)}) = (3, 1, 1, 3) \tag{3}$$

where the audiovisual multi-modality perceived label has larger weight than the two single-modality perceived labels, and the intended label is of the same importance with the multi-modality perceived label. Moreover, the relevance values $r_{i,t}^{(*)}$ take perceived `Ambiguous` labels into consideration: in Anger ranker for example, if $P_i^{(*)} =$ `Anger` then $r_{i,A}^{(*)} = 2$, if $P_i^{(*)}$ is `Ambiguous`, confusing Anger with other emotions, then $r_{i,A}^{(*)} = 1$, if $P_i^{(*)}$ equals to an emotion other than `Anger`, then $r_{i,A}^{(*)} = 0$, where $P_i^{(*)}$ can be any one of the three perceived labels $P_i^{(a)}, P_i^{(v)}, P_i^{(m)}$; as for the intended label, if $I_i =$ `Anger` then $r_{i,A}^{(I)} = 2$ otherwise $r_{i,A}^{(I)} = 0$, the same for other emotion rankers.

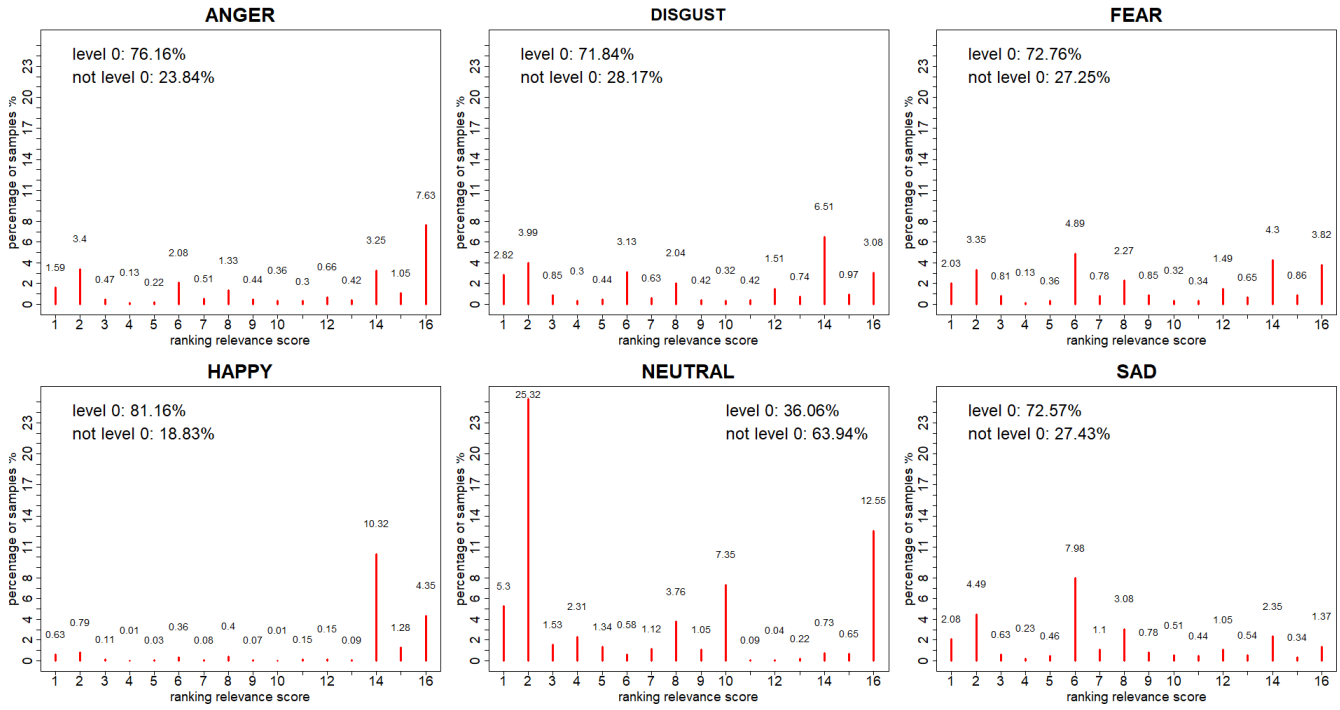

Fig. 3. Histogram of ranking relevance scores of Intermediate Ranking Rule on the CREMA-D dataset

Compared to the simple ranking rule that only have five different ranking relevance levels, intermediate ranking rule calculates the ranking relevance scores in a more meticulous way, with seventeen different levels. Figure 3 shows the histograms of them for six emotions respectively. The horizontal axis shows the ranking relevance score, and level 0 also means none of the four labels is as the same as this target emotion. We can also see that for Disgust and Fear, the distribution of non-zero ranking relevance scores is quite even. For Anger and Happy, the non-zero ranking scores tend to be higher level, representing more agreement among the four labels. For Sad emotion, the ranking scores tend to be lower level, showing more mismatches between the four labels. Unlike the other five emotions that the majority of samples fall in level 0, the distribution of ranking scores for Neutral is more uneven, showing there are more disagreement and mismatches among the intended and perceived labels in understanding Neutral emotion.

4.2.3 *Complex Ranking Rule*

We further design the **"complex ranking rule"** by considering different roles and importances for the four labels retrieved during emotion production and perception. Specifically, complex ranking rule views roles of the four labels in a meticulous way: Firstly, the intended label reveals human emotion expression behaviors, however the other three perceived labels show human emotion perception behaviors. These two kinds of behaviors should carry the same weight in identifying emotion class; Secondly, among the three perceived labels, the crowd-sourced vote results from multi-modality should be more reliable than the vote results from single-modality, thus audiovisual perceived label should have larger weight than audio or video perceived label; Thirdly, the data with perceived label `Ambiguous` in target emotion may be more intense than the data with the perceived label totally disagrees with the target emotion.

Complex ranking rule still uses (1) to add the four labels' influence factors together as ranking relevance score, but it counts the influence factors in a different way. Based on the assumptions stated in the previous paragraph, the complex ranking rule employs the weight set for all i:

$$w_i = (w_i^{(I)}, w_i^{(a)}, w_i^{(v)}, w_i^{(m)}) = (7, 2, 2, 3) \tag{4}$$

where the weight of the intended label which corresponds to the expression channel equals to the sum of the weights of three perceived labels which correspond to the perception channel, and audiovisual multi-modality perceived label has larger weight than the two single-modality perceived labels. Also, complex ranking rule takes perceived `Ambiguous` labels into consideration as well: if the perceived label equals to the target emotion, its corresponding relevance value is assigned as 2, if the perceived label is `Ambiguous`, its corresponding relevance value is assigned as 1, and if the perceived label equals to an emotion other than the target emotion, its corresponding relevance value is 0. As for the intended label, if the intended emotion equals to the target emotion, then $r_{i,t}^{(I)} = 2$ otherwise $r_{i,t}^{(I)} = 0$.

Compared to simple ranking rule and intermediate ranking rule, complex ranking rule differentiates emotion intensity level in a very detailed manner, with twenty-nine levels. Table 3 shows a comprehensive analysis for these twenty-nine ranking relevance scores in complex ranking rule, and Figure 4 shows the histograms of them for six emotions respectively. Similar as what we observed with the previous two ranking rules, we can see that for Anger, Disgust, and Fear, the distribution of non-zero ranking relevance scores is quite even. For Happy emotion, this distribution is skewed to higher relevance scores, meaning more labels agree with

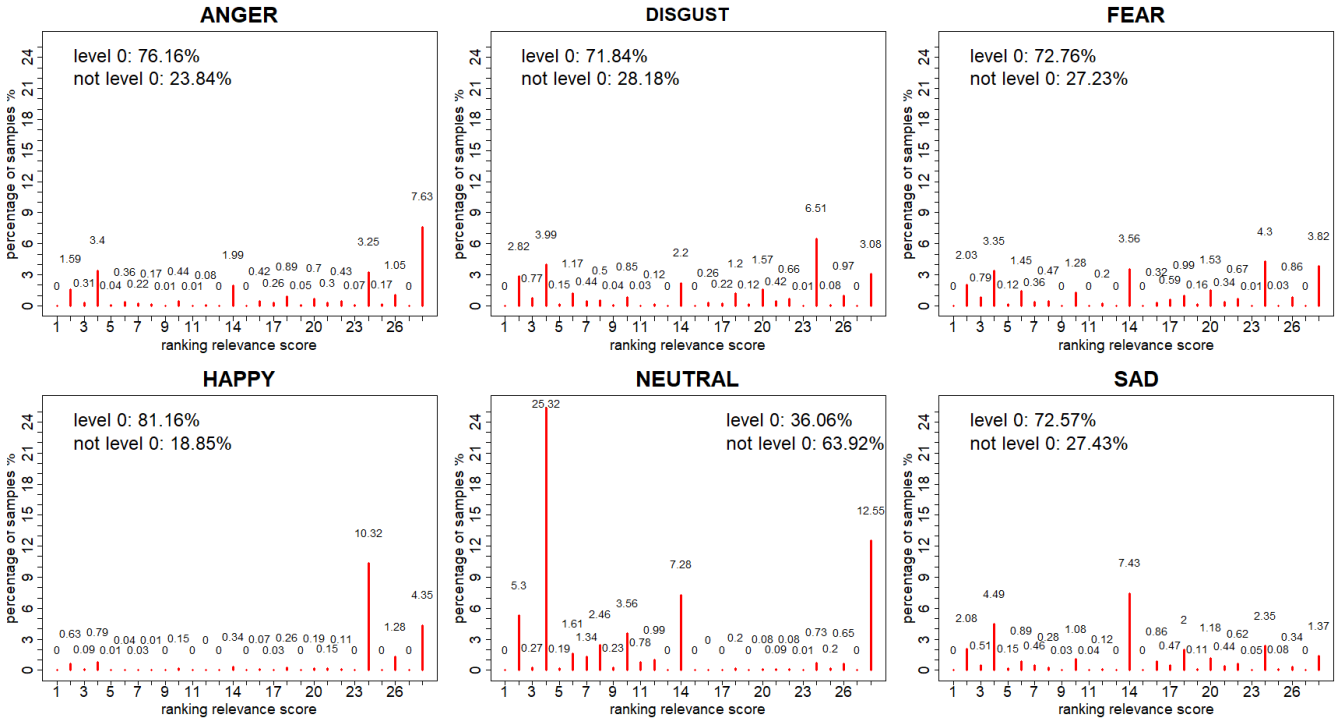

Fig. 4. Histogram of ranking relevance scores of Complex Ranking Rule on the CREMA-D dataset

TABLE 3
A comprehensive analysis for the twenty-nine ranking relevance scores in complex ranking rule

Ranking relevance score value	Which label matches with the target emotion
$[0, 4)$	none of the four labels
$[4, 6)$	one of single-modality perceived labels
$[6, 8)$	multi-modality perceived label
$[8, 10)$	both the audio perceived label and video perceived label
$[10, 14)$	one of single-modality perceived labels and multi-modality perceived label
14	only intended label, or all three perceived labels
$[14, 18)$	only intended label
$[18, 20)$	intended label and one of single-modality perceived labels
$[20, 22)$	intended label and multi-modality perceived label
$[22, 24)$	intended label and two single-modality perceived labels
$[24, 28)$	intended label, one of single-modality perceived labels, and multi-modality perceived label
28	all the four labels

the target emotion. For Neutral and Sad, however, this distribution is skewed to lower relevance scores, meaning more disagreement among the four labels. Thus, in complex ranking rule, it is also easier for the system to distinguish Happy, and harder to recognize Neutral and Sad. On the other hand, different from what have been found in the simple ranking rule that the distribution of samples with non-zero ranking level is skewed towards one or two levels, we can observe much wider distribution across the different relevance scores with the complex ranking rule, suggesting that the complex ranking rule can better capture the more complex and ambiguous emotion expressions.

4.3 Emotion Rankers using Learning-to-Rank Models

Through the mapping function of ranking rules, our data become $\{x_i, s_i, r_{i,t}\}, i = 1 \cdots m$, and now we are going to build emotion ranking models to predict ranking relevance score and sort the observations in the descending order of the predicted ranking level. We build the following three ranking models for analyzing the six basic emotions. Although these three ranking models are of different structures, they are all *pairwise learning to rank models* in preference learning.

- RankSVM (ranking with support vector machine)
- RankNet (ranking with deep neural networks)
- LambdaMART (ranking with gradient boosting decision trees)

These three pairwise ranking models are originally used in information retrieval tasks to rank search results, of which the task is to sort webpages returned by a search engine by their relevances to the query. For our task, due to the large inter-subject variability in emotion expressions, we only conduct pairwise ranking between clips generated by the same speaker. In other words, a query is defined by a speaker in the dataset. When training a ranker for a target emotion t, clips generated by the same speaker are

sorted based on their relevance scores $r_{i,t}$ defined in (1). We train pairwise ranking models to minimize the number of incorrectly ordered pairs generated by the same speaker. In testing, the outputs of the ranker for target emotion t are used to sort all clips generated by a speaker in terms of their relevances to the target emotion t.

4.3.1 RankSVM - Ranking with Support Vector Machine

RankSVM proposed by Joachims [40] is a classical ranking model using Support Vector Machine (SVM) in preference learning. The idea behind it is to transform the pairwise ranking problem to a binary classification problem on pairs with a partial ordering. For a given subject s, U_i is the i-th sample and U_j is the j-th sample (with $s_i = s_j = s$), and their feature vectors are x_i and x_j, respectively. Let $U_i \succ U_j$ denote the event that U_i has a higher relevance score to emotion t than U_j, i.e. $r_{i,t} > r_{j,t}$, and $\mathcal{P}$ denote the set of pairs of indices (i, j) for which $U_i \succ U_j$. The RankSVM optimization problem is formulated as:

$$\begin{aligned} &\min_{\omega,\xi} \frac{1}{2}\|\omega\|^2 + C \sum_{\{i,j\}\in\mathcal{P}} \xi_{i,j} \\ &s.t.\ \langle \omega, (x_i - x_j)\rangle \geq 1 - \xi_{i,j},\ \xi_{i,j} \geq 0 \quad \forall \{i, j\} \in \mathcal{P}, \end{aligned} \tag{5}$$

where ξ represents the slack variables and C is the parameter to trade-off margin size against training error. RankSVM learns the optimal weight vector $\hat{\omega}$ to minimize the objective function in equation (5) through training. In the testing process, if $\langle \hat{\omega}, (x_i - x_j)\rangle \geq 0$, then $U_i \succ U_j$. In this way, after subtracting feature vectors of a pair of samples, i.e. $x_i - x_j$, RankSVM converts the ranking problem into a binary classification problem on partially ordered pairs and solves it with SVM classification.

4.3.2 RankNet - Ranking with Deep Neural Networks

RankNet proposed by Burges [41] is another pairwise ranking model built on deep neural network. It deploys a probabilistic ranking cost function and uses gradient descent to update model parameters in neural networks. At a given point during training, RankNet maps an input feature vector $x \in R^d$ to a number $f(x) \in R$. For each pair of samples U_i and U_j in the set $\mathcal{P}$, RankNet computes their outputs $s_i = f(x_i), s_j = f(x_j)$, and maps the two outputs to a learned probability that U_i is ranked higher than U_j via a sigmoid function:

$$P_{ij} = P(U_i \succ U_j) = \frac{1}{1 + e^{-\sigma(s_i - s_j)}} \tag{6}$$

where σ determines the shape of the sigmoid. RankNet uses the classical cross-entropy cost function:

$$\begin{aligned} C_{ij} &= -\bar{P_{ij}} \log(P_{ij}) - (1 - \bar{P_{ij}}) \log(1 - P_{ij}) \\ &= \frac{1}{2}(1 - S_{ij})\sigma(s_i - s_j) + \log(1 + e^{-\sigma(s_i - s_j)}), \end{aligned} \tag{7}$$

where S_{ij} takes value of 1 if $U_i \succ U_j$ or -1 if $U_i \prec U_j$, $\bar{P_{ij}} = \frac{1}{2}(1 + S_{ij})$ is the known probability that U_i is ranked higher than U_j. During training process, RankNet updates the model parameters w_k in neural network using gradient descent:

$$\begin{aligned} w_k \to w_k - \eta \frac{\partial C}{\partial w_k} &= w_k - \eta \sum_{(i,j)\in\mathcal{P}} \left(\frac{\partial C_{ij}}{\partial s_i} \frac{\partial s_i}{\partial w_k} + \frac{\partial C_{ij}}{\partial s_j} \frac{\partial s_j}{\partial w_k} \right) \\ &= w_k - \eta \sum_{(i,j)\in\mathcal{P}} \frac{\partial C_{ij}}{\partial s_i} \left(\frac{\partial s_i}{\partial w_k} - \frac{\partial s_j}{\partial w_k} \right) \\ &= w_k - \eta \sum_{(i,j)\in\mathcal{P}} \lambda_{ij} \left(\frac{\partial s_i}{\partial w_k} - \frac{\partial s_j}{\partial w_k} \right), \end{aligned} \tag{8}$$

where $\lambda_{ij} = \sigma(\frac{1 - S_{ij}}{2} - \frac{1}{1 + e^{\sigma(s_i - s_j)}})$.

4.3.3 LambdaMART - Ranking with Boosted Decision Tree

LambdaMART proposed by Burges [42] as a pairwise ranking model based on gradient boosted decision trees, is a combination of MART and LambdaRank [43]. LambdaRank is an updated version of RankNet whose key idea is computing the desired gradients λ_i directly instead of deriving them from a cost as in (8):

$$\begin{aligned} w_k \to w_k - \eta \sum_{(i,j)\in\mathcal{P}} (\lambda_{ij} \frac{\partial s_i}{\partial w_k} - \lambda_{ij} \frac{\partial s_j}{\partial w_k}) \\ = w_k - \eta \sum_i \lambda_i \frac{\partial s_i}{\partial w_k} \\ where \quad \lambda_i = \sum_{j:(i,j)\in\mathcal{P}} \lambda_{ij} - \sum_{j:(j,i)\in\mathcal{P}} \lambda_{ij} \end{aligned} \tag{9}$$

for simplicity, we denote the above sum operation as

$$\sum_{(i,j)\leftrightarrow\mathcal{P}} \lambda_{ij} \equiv \sum_{j:(i,j)\in\mathcal{P}} \lambda_{ij} - \sum_{j:(j,i)\in\mathcal{P}} \lambda_{ij}$$

What's more, different from RankNet only optimizing for the number of pairwise errors, LambdaRank can also optimize for other measures that are either discontinuous or flat. For example, when optimizing for the Normalized Discounted Cumulative Gain (NDCG), LambdaRank just modifies the gradients by multiplying the size of the change in NDCG ($|\Delta_{NDCG}|$) given by swapping the ranking positions of U_i and U_j while leaving the other samples' ranking positions unchanged:

$$\begin{aligned} \lambda_{ij} = \frac{\partial C}{\partial s_i} = \frac{-\sigma}{1 + e^{\sigma(s_i - s_j)}} |\Delta_{NDCG}| \\ \lambda_i = \sum_{j:(i,j)\in\mathcal{P}} \lambda_{ij} - \sum_{j:(j,i)\in\mathcal{P}} \lambda_{ij} = \sum_{(i,j)\leftrightarrow\mathcal{P}} \lambda_{ij} \end{aligned} \tag{10}$$

LambdaRank and MART can be well combined as LambdaMART, since LambdaRank computes the gradients directly at any training point, while MART performs gradient descent using regression trees and each tree models the gradient. Within LambdaMART, user-chosen parameters are the number of trees N, the maximum number of leaves L, and the learning rate η. Each tree maps the input feature vector x_i to the tree output $f_k(x_i), k = 1...N$ and the final output is the weighted linear combination of trees outputs $\sum_{k=1}^{N} \alpha_k f_k(x_i)$. During training process, LambdaMART learns $\gamma_{lk}, l = 1...L, k = 1...N$ for each leaf modelling the gradients of the cost with respect to

the tree output with Newton method shown as below, and update the new tree output from the previous tree by $f_k(x_i) = f_{k-1}(x_i) + \eta \sum_l \gamma_{lk} I(x_i \in R_{lk})$, where $I(\cdot)$ is the indicator function and R_{lk} denotes the set of samples that falls in the lth leaf node of the kth tree.

$$\begin{aligned}\gamma_{lk} &= \frac{\sum_{x_i \in R_{lk}} \frac{\partial C}{\partial s_i}}{\sum_{x_i \in R_{lk}} \frac{\partial^2 C}{\partial s_i^2}} = \frac{\sum_{x_i \in R_{lk}} \lambda_i}{\sum_{x_i \in R_{lk}} \frac{\partial \lambda_i}{\partial f_{k-1}(x_i)}} \\ &= \frac{\sum_{x_i \in R_{lk}} \sum_{(i,j) \leftrightarrow \mathcal{P}} \frac{-\sigma}{1+e^{\sigma(s_i - s_j)}} |\Delta_{NDCG}|}{\sum_{x_i \in R_{lk}} \sum_{(i,j) \leftrightarrow \mathcal{P}} \sigma^2 \frac{e^{\sigma(s_i - s_j)}}{1+e^{\sigma(s_i - s_j)}} |\Delta_{NDCG}|}\end{aligned} \tag{11}$$

LambdaMART updates only a few parameters for the current leaf nodes at a time using all the data, compared to LambdaRank that updates all the parameters after each query is screened. Thus, LambdaMART, as a boosted tree version of LambdaRank, shows better results than LambdaRank and the original RankNet.

4.4 Ranking-based Emotion Recognition

Six basic emotion rankers were constructed using different types of ranking rules and preference-learning models described above. After that, we further develop ranking-based emotion classifiers based on the results obtained from the six emotion rankers. Within each emotion ranker, all the samples generated by a specific speaker s whose data were not used in the training process were mapped to predicted ranking scores in the testing process. The higher the predicted ranking scores means the more intensity in expressing the emotion. After sorting the predicted ranking scores in decreasing order, each sample of the speaker s has its predicted rank. Suppose this speaker s has m_s samples, for the k-th sample, the predicted ranks by the six emotion rankers are:

$$\{A_k, D_k, F_k, H_k, N_k, S_k, \} \tag{12}$$

where $k = 1...m_s$ and the six predicted ranks can take the integer value between 1 and m_s. We develop ranking-based emotion classifiers with three ideas: *highest rank rule, second level training, and combining policy*.

Our ranking-based emotion classifiers closely rely on the performance of the six emotion rankers. Ideally, if a sample is ranked high by a ranker for emotion t, and ranked low by the other emotion rankers, one should pick t as the final classification decision with high confidence. However, if there is no clear distinction between the ranks from all the rankers, a tougher decision has to be made to achieve high accuracy in the final classification. We divide the whole classification task into six separated subtasks of recognizing the intensity of each emotion and focus on learning only one emotion feature at a time. This makes our ranking-based classifiers have the potential to outperform the traditional classifiers that just apply the classification algorithms directly.

4.4.1 Highest Rank Rule

The *highest rank rule* directly picks the emotion of which the predicted rank is the highest as the final classification decision. For example, if the predicted rank value by the Anger ranker A_k is smaller[1] than the other five predicted ranks, we pick Anger as the final classification decision. If there are more than one predicted ranks of the same smallest value, we just randomly pick one of these emotions as the final decision.

1. Smaller rank value means higher rank

TABLE 4
Parameters used in LambdaMART models. **A:** Anger, **D:** Disgust, **F:** Fear, **H:** Happy, **N:** Neutral, **S:** Sad; **R1:** Simple ranking rule, **R2:** Intermediate ranking rule, **R3:** Complex ranking rule.

	Number of trees	Max number of leaves	Learning rate
A-R1	140	21	0.08
A-R2	220	21	0.09
A-R3	280	17	0.09
D-R1	90	45	0.08
D-R2	180	55	0.1
D-R3	220	55	0.09
F-R1	140	35	0.07
F-R2	300	45	0.1
F-R3	600	60	0.1
H-R1	100	15	0.01
H-R2	100	15	0.01
H-R3	250	20	0.03
N-R1	100	50	0.09
N-R2	1000	60	0.09
N-R3	3000	55	0.1
S-R1	130	27	0.09
S-R2	500	35	0.09
S-R3	2000	29	0.09

4.4.2 Second Level Training

Second level training builds a second-level learning model by transforming the six predicted ranks as the second-level features, and makes the classification decisions using the second-level model outputs. The second-level feature vector

$$(A_k^V, D_k^V, F_k^V, H_k^V, N_k^V, S_k^V), k = 1...m_s \tag{13}$$

is transformed element-wise from the vector of the predicted ranks in (12) using a linear function:

$$f(x) = \frac{m_s - x}{m_s - 1}, \tag{14}$$

so that the highest rank (rank 1) is converted to feature value of 1, and the lowest rank (rank m_s) is converted to feature value of 0. The second-level learning model we use is Support Vector Machine (SVM) with radial kernel, which maps the six original predicted ranks to the final classification decision with a non-linear model.

4.4.3 Model Combination

Finally, we consider the combination of multiple models, aiming to take advantage of the complementary information of multiple ranking models to increase the classification accuracy. Here, we combine predictions from the classifiers by taking the average of the posterior probabilities for each emotion class produced by all the classifiers. The emotion class with the highest average posterior probability is predicted as the class for each sample in the test data.

5 Experimental Results

In this section we evaluate the effectiveness of the proposed multimodal ranking rules and preference learning models for emotion recognition on the CREMA-D dataset. Throughout this section, we use the following abbreviations: Target

TABLE 5
R-precision for three ranking models with three ranking rules. **M1:** RankSVM, **M2:** RankNet, **M3:** LambdaMART; **R1:** Simple ranking rule, **R2:** Intermediate ranking rule, **R3:** Complex ranking rule.

%	Anger	Disgust	Fear	Happy	Neutral	Sad
M1-R1	78.03	72.87	60.72	85.99	63.58	56.51
M1-R2	78.98	72.39	61.27	85.21	63.68	57.31
M1-R3	78.74	72.56	61.59	85.62	64.15	57.63
M2-R1	86.30	81.06	68.21	**94.42**	75.81	65.43
M2-R2	**86.86**	81.05	68.13	94.26	75.08	62.73
M2-R3	86.77	81.69	**69.64**	94.35	76.99	65.09
M3-R1	85.45	82.94	67.11	91.83	75.76	64.33
M3-R2	85.85	84.43	68.14	90.96	80.26	64.57
M3-R3	85.92	**84.82**	68.52	93.87	**80.52**	**66.76**

TABLE 6
$Precision@K$ for which we retrieved all the target emotion samples for three ranking models with three rules. **M1:** RankSVM, **M2:** RankNet, **M3:** LambdaMART; **R1:** Simple ranking rule, **R2:** Intermediate ranking rule, **R3:** Complex ranking rule.

%	Anger	Disgust	Fear	Happy	Neutral	Sad
M1-R1	51.73	46.05	35.24	63.09	46.37	35.95
M1-R2	53.45	47.30	35.17	65.97	46.99	36.74
M1-R3	53.85	47.82	35.68	66.02	46.55	36.78
M2-R1	67.44	58.50	39.67	86.91	56.76	41.24
M2-R2	**68.04**	58.63	39.76	85.29	56.39	41.10
M2-R3	65.18	57.51	38.81	**87.46**	57.74	42.18
M3-R1	66.93	64.32	38.24	78.25	58.49	44.68
M3-R2	67.56	**65.97**	**41.47**	75.75	65.22	45.75
M3-R3	68.02	65.94	40.33	86.24	**67.28**	**47.99**

TABLE 7
The classification accuracy of different baseline models. **Multi:** SVM classifier with multi-modality features, **Audio:** SVM classifier with audio acoustic feature, **Video:** SVM classifier with video facial features, **NN:** baseline classifier with neural network, **Tree:** baseline classifier with gradient boosted decision tree. **Rand:** random baseline

%	Overall	ANG	DIS	FEA	HAP	NEU	SAD
Multi	75.72	**86.01**	75.73	60.10	89.39	**79.46**	64.15
Audio	58.83	77.06	50.16	44.48	55.25	63.71	63.01
Video	64.76	67.76	72.25	53.19	88.93	61.40	44.55
NN	**76.19**	83.88	**76.00**	**63.04**	87.81	78.38	**68.34**
Tree	72.96	83.09	75.81	51.82	**89.40**	74.75	63.10
Rand	16.94	17.47	18.08	17.47	17.46	14.81	16.05

Basic Emotions, **A:** Anger, **D:** Disgust, **F:** Fear, **H:** Happy, **N:** Neutral, **S:** Sad; Ranking Rules, **R1:** Simple ranking rule, **R2:** Intermediate ranking rule, **R3:** Complex ranking rule; Ranking Models, **M1:** RankSVM, **M2:** RankNet, **M3:** LambdaMART.

In order to confirm the stability and speaker independence of the obtained classifiers, we performed all the experiments using leave-one-subject-out (LOSO) paradigm. In this form of cross-validation, all samples from a given speaker are used as a test set for a model trained on the data from all the other speakers, and the process is repeated for all speakers. The overall performance of the classifier is evaluated by combining the predictions from all test folds and computing the overall accuracy for the entire dataset. In our study, we used SVM rank toolkit [44] to implement the RankSVM model and LightGBM [45] to implement the LambdaMART model. The parameters used in LambdaMART (number of trees N, maximum number of leaves L, learning rate η) are optimized for Spearman's rank-order correlation in five-fold subject-independent validation, and are summarized in Table 4. Our RankNet structure is three-layer with 256 & 64 units in each hidden layer, ReLU as the activation functions, 256 batch size, 35 epochs for simple ranking rule, 45 epochs for intermediate ranking rules and 50 epochs for complex ranking rule. For all the experiments, similar to other studies used the CREMA-D dataset, we use the intended label as the ground-truth.

First, we analyze the performance of individual emotion rankers in Section 5.1. Six speaker-independent emotional rankers were constructed, one for each basic emotion. Next, we examine the accuracy of different approaches that incorporate the results from individual rankers to perform multi-class prediction for each utterance in Sections 5.2.1 and 5.2.2. We further discuss the performance of model combinations in Section 5.2.3. The key observations from the experimental results are summarized in Section 5.3. Finally, we also perform the cross-corpus experiments to evaluate the generalization ability of our ranking-based emotion classifier on other multi-modal emotion dataset.

5.1 Evaluation of Emotion Rankers

We first analyze the performance of the six emotion rankers under different ranking rules and ranking models. Generally speaking, for a good ranker, samples of the target emotion should be ranked higher than samples of any other emotions. We first look at precision at k, which is widely used to evaluate the performance of ranking models. It is defined as the percentage of utterances in the top k utterance list ordered by the decreasing ranking scores that indeed express the target emotion. Ideally, precision at k should achieve 100% for k less than the total number of target emotion samples of this speaker, and then decrease steadily. Fig. 5 shows the precision at different k values for the three ranking models with two ranking rules.

Meanwhile, in an idealized situation where we know the number of utterances that convey the target emotion for each speaker, we can measure the R-precision at different R values for different speakers. For example, if we knew that a speaker said 10 utterances in an angry manner, we can look at the precision at $R = 10$ for that speaker and see how many of the top ranked ten utterances were indeed anger utterances. Besides, we also consider the precision at the smallest k value that allows us to retrieve all the samples whose ground-truth is the target emotion, i.e., the smallest k achieving 100% recall, denoted as $Precision@K$. The better the ranker model is, the larger these two metrics will be. Table 5 and Table 6 shows the R-precision and $Precision@K$ results for the six emotion rankers under three ranking models with two ranking rules.

We first compare the three preference learning models. From Fig 5, Table 5, and Table 6, we can first see that the advanced RankNet and LambdaMART models have significantly better performance than the RankSVM model on all the six emotions with all the three ranking rules. We further compare the RankNet and LambdaMART models and observe that the LambdaMART model has slightly better performance than RankNet but there is no significant improvement between them except on Disgust and Neutral.

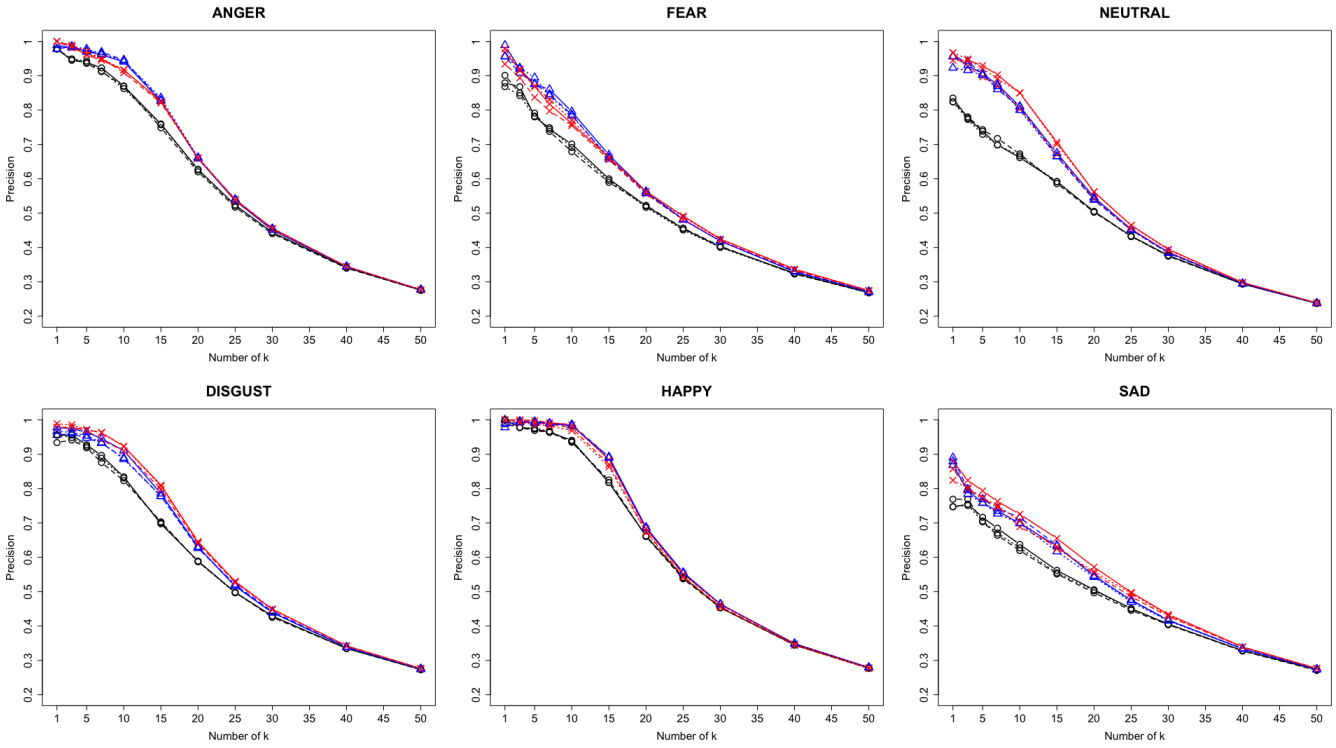

Fig. 5. Precision at k for the six emotion rankers on the CREMA-D dataset. Line type represents three ranking rules (Dashed line: Simple ranking rule, Dotted line: Intermediate ranking rule, Solid line: Complex ranking rule). Color represents three ranking models (Black: RankSVM, Blue: RankNet, Red: LambdaMART)

We then turn to study the impact of ranking rules. We notice that the complex ranking rule performs slightly better than simple ranking rule and intermediate ranking rule with all the three ranking models. However, after performing significance test with student t-test of 95% confidence, there are no significant improvement of complex ranking rule compared to the other two ranking rules, except for the LambdaMART model on Happy, Neutral, Sad. We can interpret this observation as LambdaMART model can learn out small difference in emotion intensity for these three emotions while the other two models cannot, thus producing better results with complex ranking rule than with the other two ranking rules.

5.2 Multi-class Emotion Classification

Next, we combine ranking scores from different preference learning models to form the final multi-class emotion classification. In order to investigate the effectiveness of preference learning models in emotion recognition, we will compare the developed ranking-based multi-class classifiers to the conventional supervised methods.

We introduce three types of conventional classifiers as our baselines: SVM, deep neural network (NN), and gradient boosted decision tree (Tree). The same multi-modality features used in training the preference learning models are used in the baselines. The SVM classifier was trained with radial basis kernels. The neural network classifier (NN) has the same structure as the RankNet: three layers with 256 & 64 units in each hidden layer. The gradient boosted decision tree classifier (Tree) has the same structure as the LambdaMART.

We report the emotion recognition accuracy of the three baseline models in Table 7. For the sake of comparison, we also trained the conventional uni-modal SVM classifiers with acoustic features and video facial features respectively, and report the results in Table 7 as well. As expected, using multi-modality features can lead to higher accuracy on all the six emotions than only using single modality features. We also include a random baseline that randomly pick one emotion and report its accuracy in the last row of Table 7.

5.2.1 *Highest Rank Rule*

The multi-class emotion classification accuracy of ranking-based classifiers with the highest rank rule are summarized in Table 8. We can first see that both RankNet (M2) and LambdaMART (M3) model have significantly improvement on all six emotions with both ranking rules compared to RankSVM (M1) model. The LambdaMART model achieves the best overall accuracy of 82.21%, which is slightly better than the 80.57% overall accuracy obtained by RankNet model, and it shows significant improvement on predicting Disgust, Neutral, and Sad. Moreover, the proposed ranking-based emotion classifiers has significantly higher accuracy on all the six emotions compared to the best conventional classifier baseline with 76.19% overall accuracy as shown on the last three rows (copy from the Table 7 for easier comparison).

On the other hand, we can also see that the complex ranking rule (R3) performs only slightly better than the simple ranking rule (R1) and the intermediate ranking rule (R2) with RankSVM and RankNet models, and there is no significant improvement. However, the benefit of complex

TABLE 8
The classification accuracy of ranking-based emotion classifier with the Highest Rank Rule. **M1:** RankSVM, **M2:** RankNet, **M3:** LambdaMART; **R1:** Simple ranking rule, **R2:** Intermediate ranking rule, **R3:** Complex ranking rule. The last three rows include the baselines with SVM, neural network, and gradient boosted decision tree for comparison.

%	Overall	ANG	DIS	FEA	HAP	NEU	SAD
M1-R1	73.58	80.16	73.65	63.79	89.38	79.70	55.66
M1-R2	74.01	80.39	73.74	63.87	90.17	79.81	56.92
M1-R3	74.29	80.87	74.53	63.96	90.65	79.54	56.91
M2-R1	80.04	84.36	82.30	69.70	93.25	85.80	65.69
M2-R2	80.04	86.15	80.32	**70.48**	93.56	86.09	64.48
M2-R3	80.57	87.02	82.70	69.79	**93.63**	85.53	65.43
M3-R1	80.16	87.03	83.64	66.56	93.32	86.96	64.36
M3-R2	80.85	86.87	84.36	67.44	92.38	88.35	66.79
M3-R3	**82.21**	**87.34**	**85.31**	69.56	93.32	**88.96**	**69.71**
SVM	75.72	86.01	75.73	60.10	89.39	79.46	64.15
NN	76.19	83.88	76.00	63.04	87.81	78.38	68.34
Tree	72.96	83.09	75.81	51.82	89.40	74.75	63.10

TABLE 9
The classification accuracy of ranking-based emotion classifier with Second Level Training. **M1:** RankSVM, **M2:** RankNet, **M3:** LambdaMART; **R1:** Simple ranking rule, **R2:** Intermediate ranking rule, **R3:** Complex ranking rule.

%	Overall	ANG	DIS	FEA	HAP	NEU	SAD
M1-R1	76.60	82.27	80.26	64.03	90.18	74.12	68.33
M1-R2	76.54	83.59	80.03	63.33	90.88	73.44	67.47
M1-R3	76.97	83.92	80.35	63.65	90.73	75.29	67.63
M2-R1	81.95	87.80	85.15	**72.78**	93.55	78.89	73.08
M2-R2	81.44	87.95	84.04	70.97	93.39	81.10	71.10
M2-R3	82.01	87.88	84.77	70.43	**94.11**	81.00	73.70
M3-R1	81.57	87.49	**86.80**	66.02	92.30	81.88	74.92
M3-R2	82.12	87.89	85.70	69.72	91.75	83.62	74.20
M3-R3	**82.93**	**88.76**	86.73	69.24	92.54	**84.72**	**75.78**

ranking rule is successfully shown with the LambdaMART model, where the complex ranking rule performs significantly better than the other two ranking rules, especially on Disgust, Fear, Neutral, and Sad. This suggests that the LambdaMART model is more sensitive and has stronger ability to learn out small difference in emotion intensity.

5.2.2 Second-Level Training

The emotion classification accuracy of ranking-based classifiers with second-level model training are summarized in Table 9. Similar to what we found in the classifiers with the highest ranking rule, both RankNet and LambdaMART have significantly higher accuracy than RankSVM on all the six emotions for both ranking rules. However, there is no significant difference between RankNet and LambdaMART. LambdaMART performs better on Anger, Disgust, Neutral, Sad, while RankNet performs better on Fear and Happy. We also notice that the complex ranking rule performs slight better than the simple ranking rule and the intermediate ranking rule, but there is no significant difference except for the LambdaMART model on Fear and Neutral.

As what we expect, the ranking-based emotion classifier with second-level training has significantly higher accuracy on all the six emotions than the baseline, and the best overall accuracy attained is 82.93%. Compared with Table 8, ranking-based emotion classifier with second-level training significantly improved the overall accuracy and the accuracy on Disgust & Sad compared to ranking-based emotion classifier with the highest rank rule.

TABLE 10
The classification accuracy of ranking-based emotion classifier with Model Combination. **R1:** Simple ranking rule, **R2:** Intermediate ranking rule, **R3:** Complex ranking rule.

%	Overall	ANG	DIS	FEA	HAP	NEU	SAD
Combine RankNet and RankSVM							
R1	81.65	87.71	84.52	70.72	94.27	81.29	71.34
R2	82.17	88.81	84.91	70.88	94.27	81.27	72.69
R3	82.18	88.88	85.23	70.98	94.58	80.91	72.28
Combine LambdaMART and RankSVM							
R1	82.46	88.67	86.25	69.09	93.71	83.43	73.71
R2	83.21	89.69	87.44	70.20	93.64	84.27	74.12
R3	83.70	90.08	87.28	71.37	94.89	83.99	74.59
Combine LambdaMART and RankNet							
R1	84.21	89.85	**88.14**	71.84	94.81	84.48	76.16
R2	84.04	90.08	86.49	73.02	93.79	86.02	75.05
R3	**85.06**	89.84	87.59	**73.88**	**95.37**	86.14	**77.65**
Combine RankSVM & RankNet & LambdaMART							
R1	83.88	89.68	87.51	71.43	94.58	84.29	75.84
R2	84.41	90.62	87.59	72.38	95.13	84.77	76.00
R3	84.74	90.71	87.28	73.10	95.13	85.42	76.86

5.2.3 Model Combination

Finally, we consider the direct combination of different ranking-based classifiers together into a single prediction. The results of the overall accuracy and the accuracy on each emotion from model combination are summarized in Table 10. We found that when we combine the conventional baseline with various ranking-based models, the combined systems significantly outperform the conventional baseline, but do not show substantial improvements over the ranking-based systems. This may be because the conventional baseline classifiers perform significantly worse than the ranking-based systems for almost all the emotion classes, and so the combination of predictions is not beneficial.

However, the performance clearly improves when two or more ranking-based classifiers are combined by taking the advantage of complementary information provided by different ranking models. The best overall accuracy is attained at 85.06% when the two best classifiers–LambdaMART and RankNet using Complex ranking rule–are combined, which is significantly better than the conventional baselines and all different types of single preference learning models. The combined system of LambdaMART and RankNet exhibits much higher recall rate on all the emotion classes including Anger, Disgust, Fear, Happy, Sad, and maintains a good recall rate on Neutral.

5.3 Discussion

From the experimental results in Section 5.1 and 5.2, we now have the following observations:

Comparison of the three ranking rules: Complex ranking rule performs better than the simple ranking rule and intermediate ranking rules, but there is no significant difference except for LambdaMART on non-positive emotions: Disgust, Fear, Neutral, and Sad.

Comparison of three ranking models: Both RankNet and LambdaMART models perform significantly better than RankSVM model on all the six emotions for all ranking

TABLE 11
The classification accuracy of ranking-based emotion classifier while training on CREMA-D and testing on SAVEE dataset. **M1:** RankSVM, **M2:** RankNet, **M3:** LambdaMART

%	Overall	ANG	DIS	FEA	HAP	NEU	SAD
Cross Corpus Baseline							
Rand	16.90	13.33	1.67	11.67	23.33	25.00	18.33
Multi	14.29	**98.33**	00.00	1.67	00.00	00.00	00.00
Highest Rank Rule							
M1	29.52	31.67	26.67	28.33	23.33	37.50	21.67
M2	43.10	48.33	43.33	30.00	85.00	33.33	28.33
M3	**59.05**	70.00	71.67	36.67	**100.00**	**50.83**	33.33
Second Level Training							
M1	29.29	38.33	35.00	35.00	23.33	30.83	11.67
M2	44.76	51.67	53.3	**43.33**	75.00	30.00	30.00
M3	58.33	66.67	**75.00**	40.00	**100.00**	40.00	**46.67**
Within Corpus Training & Testing							
LOSO	64.29	75.00	43.33	58.33	80.00	80.00	33.33

rules. Between LambdaMART and RankNet, the LambdaMART model performs slightly better than RankNet. Moreover, the LambdaMART model has stronger ability to learn out small differences in emotion intensity and is more sensitive on non-positive emotions (Disgust, Fear, Neutral, Sad), which is the reason why the Complex ranking rule is significantly better than the Simple ranking rule when used with LambdaMART on these emotions. Also, the LambdaMART model is far quicker than RankNet (20X speedup).

Comparison of conventional SVM with ranking-based classifiers: All the three ranking-based emotion classifiers have significant accuracy improvements on all the six basic emotions over the baselines that applied SVM classifier directly. Ranking-based emotion classifier with second-level training improves the overall accuracy and achieves significant accuracy improvement on Disgust & Sad compared to ranking-based classifiers with highest rank rule. Combination of multiple ranking-based classifiers further improves the prediction accuracy.

5.4 Cross-corpus Evaluation

We also perform the cross-corpus emotion recognition experiments and discuss the effectiveness and generalization ability of our ranking-based emotion classifier on other multimodal emotion dataset.

The *Surrey Audio-Visual Expressed Emotion* (SAVEE) dataset [46] is selected for the cross-corpus evaluation. SAVEE contains audio-visual recordings from 4 male subjects in 7 different emotions: Anger, Disgust, Fear, Happy, Neutral, Sad, and Surprise. Here, we remove the Surprise emotion which do not exist in CREMA-D, and use the rest of 420 recording clips (105 audio-visual clips for each subject) of other six basic emotions. Different from the CREMA-D dataset, the SAVEE dataset do not have perceived labels in either single-modality or multi-modality annotated.

We first trained the six emotion rankers on CREMA-D, and then evaluated the ranking-based classifier on the unseen SAVEE dataset. The audio and video feature extracted for SAVEE is the same as the features used for CREMA-D in all previous experiments. The complex ranking rule, which is the rule of the best performance, is selected for the cross-corpus evaluation. To test the generalization ability of our developed approach, SAVEE dataset only serves as testing set and none of clips in SAVEE participate in the training phase. In this experimental setting, the six emotion rankers are trained on CREMA-D with both intended and multi-modal perceived labels considered, and the aggregated ranking-based classifier is evaluated on SAVEE dataset.

Table 11 shows the overall accuracy and accuracy for each emotion while training on CREMA-D and testing on SAVEE. The first row represents the random baseline that randomly pick one emotion as the predicted result. The second row is the conventional SVM classifier baseline with audiovisual features under the cross-corpus setting. Results for RankSVM (M1), RankNet (M2), LambdaMART (M3) ranking models with Highest Rank Rule, and Second Level Training, are reported in the following rows. We also provide the results of the standard leave-one-subject-out (LOSO) within-corpus evaluation in the last row of the table, which shows the in-domain performance where the training and testing data are both from the SAVEE dataset.

Observing the results in Table 11, we can see that the conventional supervised classifier performs extremely bad when the training set and testing set are from different corpus/domain. The svm classifier tends to recognize almost all the samples in SAVEE into Anger after learning on CREMA-D. We can interpret this phenomenon as the data distribution in the training and testing set is quite dissimilar, resulting in the lack of good generalization ability of the conventional supervised method. However, all the ranking-based emotion classifier perform noticeably better than the baselines. LambdaMART exhibits the best performance, exceeding the baseline by 42.15% in the overall accuracy. RankNet is the second best, significantly improving the overall accuracy by 27.86%, and RankSVM can also bring 12.62% increase. Compared to the conventional supervised method, our ranking-based emotion classifier has noticeably better generalization ability on other dataset. Although under the more strict cross-corpus setting where the ranker wasn't trained on any in-domain samples, our best ranking-based model can still approach closely to the in-domain within corpus performance.

6 Conclusion

This paper introduced a novel preference learning framework that simultaneously considers both intended and perceived labels while addressing the mismatches between them. By analyzing the the discrepancies and agreements between the intended and perceived labels in different modalities of audio-only, visual-only, and audio-visual, as well as the consistency among perceptual ratings of all raters, we proposed three sets of pair-wise ranking rules to generate multi-scale relevant scores for preference learning. Emotion rankers using three learning-to-rank models of support vector machine, deep neural networks, and gradient boosting decision trees were developed. Our results demonstrated that all three preference learning models significantly outperform the conventional baseline classifiers. The improvement from the preference learning models confirm the benefits of complementary information provided by different types of labels. We also observed additional improvement from the complex ranking rule, particular with the best LambdaMART model, which suggests that we

should treat different labels differently. We further discussed the complementary of the different ranking models, and obtained the best overall accuracy of 85.06% when combining the two best ranking models–LambdaMART and RankNet–together. This is significantly better than the 76.19% accuracy attained by the best baseline model. Finally, we perform the cross-corpus emotion recognition experiments by training emotion rankers on CREMA-D and tested the ranking-based emotion classifier on the SAVEE dataset. Our results show that the ranking-based classifiers outperform the conventional supervised method by larger margin when the training and testing set are from different corpus, which further prove the effectiveness and generalization ability of the proposed ranking models.

ACKNOWLEDGMENTS

This work is partially supported by the NSF of USA under EAGER Grant IIS-2034791.

REFERENCES

[1] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "Crema-d: Crowd-sourced emotional multimodal actors dataset," *IEEE transactions on affective computing*, vol. 5, no. 4, pp. 377–390, 2014.

[2] C. Busso, S. Yildirim, M. Bulut, C. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan, "Analysis of emotion recognition using facial expressions, speech and multimodal information," in *ICMI'04 - Sixth International Conference on Multimodal Interfaces*, 2004, pp. 205–211.

[3] N. Bosch, "Multimodal affect detection in the wild: Accuracy, availability, and generalizability," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, Seattle, WA, USA, November 09 - 13, 2015*, Z. Zhang, P. Cohen, D. Bohus, R. Horaud, and H. Meng, Eds. ACM, 2015, pp. 645–649.

[4] H. Cao, A. Savran, R. Verma, and A. Nenkova, "Acoustic and lexical representations for affect prediction in spontaneous conversations," *Computer speech & language*, vol. 29, no. 1, pp. 203–217, 2015.

[5] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal processing magazine*, vol. 18, no. 1, pp. 32–80, 2001.

[6] P. Ekman, "An argument for basic emotions," *Cognition & emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

[7] G. Donato, M. S. Bartlett, J. C. Hager, P. Ekman, and T. J. Sejnowski, "Classifying facial actions," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 21, no. 10, p. 974, 1999.

[8] M. Pantic and M. S. Bartlett, "Machine analysis of facial expressions," in *Face recognition*. InTech, 2007.

[9] J. Nicolle, V. Rapp, K. Bailly, L. Prevost, and M. Chetouani, "Robust continuous prediction of human emotions using multiscale dynamic cues," in *Proceedings of the 14th ACM international conference on Multimodal interaction*. ACM, 2012, pp. 501–508.

[10] A. Savran, B. Sankur, and M. T. Bilge, "Regression-based intensity estimation of facial action units," *Image and Vision Computing*, vol. 30, no. 10, pp. 774–784, 2012.

[11] B. Schuller, M. Valster, F. Eyben, R. Cowie, and M. Pantic, "Avec 2012: the continuous audio/visual emotion challenge," in *Proceedings of the 14th ACM international conference on Multimodal interaction*. ACM, 2012, pp. 449–456.

[12] B. Schuller, B. Vlasenko, F. Eyben, G. Rigoll, and A. Wendemuth, "Acoustic emotion recognition: A benchmark comparison of performances," in *Automatic Speech Recognition & Understanding, 2009. ASRU 2009. IEEE Workshop on*. IEEE, 2009, pp. 552–557.

[13] D. Ververidis and C. Kotropoulos, "Emotional speech recognition: Resources, features, and methods," *Speech communication*, vol. 48, no. 9, pp. 1162–1181, 2006.

[14] B. Schuller, "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends," *Communications of the ACM*, vol. 61, pp. 90–99, 04 2018.

[15] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE Transactions on Affective Computing*, vol. PP, 04 2018.

[16] O. Pierre-Yves, "The production and recognition of emotions in speech: features and algorithms," *International Journal of Human-Computer Studies*, vol. 59, no. 1-2, pp. 157–183, 2003.

[17] Y.-L. Tian, T. Kanade, and J. F. Cohn, "Facial expression analysis," in *Handbook of face recognition*. Springer, 2005, pp. 247–275.

[18] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 1, pp. 39–58, 2009.

[19] C. A. Corneanu, M. O. Sim. F. Cohn, and S. E. Guerrero, "Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 8, pp. 1548–1568, 2016.

[20] T. Senechal, V. Rapp, H. Salam, R. Seguier, K. Bailly, and L. Prevost, "Facial action recognition combining heterogeneous features via multikernel learning," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 42, no. 4, pp. 993–1005, 2012.

[21] A. Savran, H. Cao, M. Shah, A. Nenkova, and R. Verma, "Combining video, audio and lexical indicators of affect in spontaneous conversation via particle filtering," in *Proceedings of the 14th ACM international conference on Multimodal interaction*. ACM, 2012, pp. 485–492.

[22] A. Savran, H. Cao, A. Nenkova, and R. Verma, "Temporal bayesian fusion for affect sensing: Combining video, audio, and lexical modalities." *IEEE Trans. Cybernetics*, vol. 45, no. 9, pp. 1927–1941, 2015.

[23] S. D'Mello and J. Kory, "Consistent but modest: A meta-analysis on unimodal and multimodal affect detection accuracies from 30 studies," in *Proceedings of the 14th ACM International Conference on Multimodal Interaction*, ser. ICMI '12. ACM, 2012, pp. 31–38.

[24] B. Pang, L. Lee *et al.*, "Opinion mining and sentiment analysis," *Foundations and Trends in Information Retrieval*, vol. 2, no. 1–2, pp. 1–135, 2008.

[25] J. J. M. Kierkels, M. Soleymani, and T. Pun, "Queries and tags in affect-based multimedia retrieval," in *Proceedings of the 2009 IEEE International Conference on Multimedia and Expo, ICME 2009, June 28 - July 2, 2009, New York City, NY, USA*, 2009, pp. 1436–1439.

[26] Y. Yang and H. H. Chen, "Ranking-based emotion recognition for music organization and retrieval," *IEEE Trans. Audio, Speech & Language Processing*, vol. 19, no. 4, pp. 762–774, 2011.

[27] M. Soleymani, G. Chanel, J. J. M. Kierkels, and T. Pun, "Affective ranking of movie scenes using physiological signals and content analysis," in *Proceedings of the 2nd ACM Workshop on Multimedia Semantics, MS 2008, Vancouver, British Columbia, Canada, October 31, 2008*, 2008, pp. 32–39.

[28] H. P. Martínez, G. N. Yannakakis, and J. Hallam, "Don't classify ratings of affect; rank them!" *IEEE Transactions on Affective Computing*, vol. 5, no. 3, pp. 314–326, 2014.

[29] G. N. Yannakakis and H. P. Martínez, "Ratings are overrated!" *Front. ICT*, vol. 2015, 2015.

[30] R. Lotfian and C. Busso, "Practical considerations on the use of preference learning for ranking emotional speech," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2016, Shanghai, China, March 20-25, 2016*, 2016, pp. 5205–5209.

[31] ——, "Retrieving categorical emotions using a probabilistic framework to define preference learning samples," in *Interspeech 2016, 17th Annual Conference of the International Speech Communication Association, San Francisco, CA, USA, September 8-12, 2016*, 2016, pp. 490–494.

[32] S. Parthasarathy and C. Busso, "Preference-learning with qualitative agreement for sentence level emotional annotations," in *Interspeech 2018, 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, 2-6 September 2018.*, 2018, pp. 252–256.

[33] H. Cao, R. Verma, and A. Nenkova, "Combining ranking and classification to improve emotion recognition in spontaneous speech," in *INTERSPEECH 2012, 13th Annual Conference of the International Speech Communication Association, Portland, Oregon, USA, September 9-13, 2012*, 2012, pp. 358–361.

[34] ——, "Speaker-sensitive emotion recognition via ranking: Studies on acted and spontaneous speech," *Computer Speech & Language*, vol. 29, no. 1, pp. 186–202, 2015.

[35] G. N. Yannakakis, R. Cowie, and C. Busso, "The ordinal nature of emotions," in *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2017, pp. 248–255.
[36] Z. Jin and H. Cao, "Development of emotion rankers based on intended and perceived emotion labels," in *Interspeech*, 2019.
[37] F. Eyben, F. Weninger, F. Groß, and B. W. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *ACM Multimedia Conference, MM '13, Barcelona, Spain, October 21-25, 2013*, 2013, pp. 835–838.
[38] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "Openface 2.0: Facial behavior analysis toolkit," *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pp. 59–66, 2018.
[39] T. Baltrusaitis, M. Mahmoud, and P. Robinson, "Cross-dataset learning and person-specific normalisation for automatic action unit detection," *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, vol. 06, pp. 1–6, 2015.
[40] T. Joachims, "Optimizing search engines using clickthrough data," in *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '02. Association for Computing Machinery, 2002, p. 133–142.
[41] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender, "Learning to rank using gradient descent," in *Proceedings of the 22nd International Conference on Machine Learning*, ser. ICML '05. Association for Computing Machinery, 2005, p. 89–96.
[42] C. J. Burges, "From ranknet to lambdarank to lambdamart: An overview," Tech. Rep. MSR-TR-2010-82, June 2010.
[43] C. Burges, R. Ragno, and Q. Le, "Learning to rank with nonsmooth cost functions," 01 2006, pp. 193–200.
[44] T. Joachims, "Training linear svms in linear time," in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '06. Association for Computing Machinery, 2006, p. 217–226.
[45] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc., 2017, pp. 3146–3154.
[46] S. Haq and P. J. B. Jackson, "Speaker-dependent audio-visual emotion recognition," in *Proc. Auditory-Visual Speech Processing*, 2009, pp. 53–58.

Yuanyuan Lei is a PhD in Computer Science student in the Department of Computer Science and Engineering at Texas A&M University (TAMU). She obtained her Master degree in Statistics from Columbia University in 2019, and Bachelor degree in both Mathematics and Computer Science from the University of Science and Technology of China in 2017. Her main areas of research are natural language processing, information extraction, text mining, speech recognition, machine learning, deep learning, and their applications in social science, multi-modality affective computing, human-centered data science.

Houwei Cao is an Associate Professor in the Department of Computer Science at New York Institute of Technology (NYIT). She was an adjunct professor at the Computer Science and Engineering Department of the Tandon School of Engineering of New York University before joining NYIT. She obtained her PhD degree in Electronic Engineering from the Chinese University of Hong Kong in 2011, and was a postdoctoral fellow at University of Pennsylvania from 2011 to 2014. Her main areas of research are signal processing, machine learning, data mining and their applications in human-centric data analytics, with emphasis on developing computational methods, algorithms, and models for speech recognition, natural language processing, multimodal affective computing, social network analysis, and healthcare information systems. She won the audio-visual emotion recognition challenge (AVEC) in 2012. Her research has been supported by the NSF, Northrop Grumman, and NYIT. Dr. Cao is a member of International Speech Communication Association (ISCA), the Association for the Advancement of Affective Computing (AAAC), and IEEE. She has served as program committee members and/or reviewers for more than ten journals and conferences in speech and language processing, affective computing, and computer vision.