Deep Learning for Continuous Multiple Time Series Annotations

Jian Huang
National Laboratory of Pattern Recognition, Institute of Automation Chinese Academy of Sciences, Beijing, China
jian.huang@nlpr.ia.ac.cn

Ya Li
National Laboratory of Pattern Recognition, Institute of Automation Chinese Academy of Sciences, Beijing, China
yli@nlpr.ia.ac.cn

Jianhua Tao
National Laboratory of Pattern Recognition, CAS Center for Excellence in Brain Science and Intelligence Technology, Institute of Automation Chinese Academy of Sciences, Beijing, China
jhtao@nlpr.ia.ac.cn

Zheng Lian
National Laboratory of Pattern Recognition, Institute of Automation Chinese Academy of Sciences, Beijing, China
lianzheng2016@ia.ac.cn

Mingyue Niu
National Laboratory of Pattern Recognition, Institute of Automation Chinese Academy of Sciences, Beijing, China
niumingyue2017@ia.ac.cn

Minghao Yang
National Laboratory of Pattern Recognition, Institute of Automation Chinese Academy of Sciences, Beijing, China
mhyang@nlpr.ia.ac.cn

ABSTRACT

Learning [1] from multiple annotations is an increasingly important research topic. Compared with conventional classification or regression problems, it faces more challenges because time-continuous annotations would result in noisy and temporal lags problems for continuous emotion recognition. In this paper, we address the problem by deep learning for continuous multiple time series annotations. We attach a novel crowd layer to the output layer of basic continuous emotion recognition system, which learns directly from the noisy labels of multiple annotators with end-to-end manner. The inputs of the system are multimodal features and the targets are multiple annotations, with the intention of learning an annotator-specific mapping. Our proposed method considers the ground truth as latent variables and multiple annotations are variant of ground truth by linear mapping. The experimental results show that our system can achieve superior performance and capture the reliabilities and biases of different annotators.

KEYWORDS

Continuous Emotion Recognition; Multiple Annotations; Deep Learning; Crowd Layer

AVEC'18, October 22, 2018, Seoul, Republic of Korea.

ACM ISBN 978-1-4503-5983-2/18/10...$15.00.
DOI: https://doi.org/10.1145/3266302.3266305

ACM Reference Format:
J. Huang, Y. Li, J. Tao, Z. Lian, M. Niu, and M. Yang. 2018. Multimodal Continuous Emotion Recognition with Data Augmentation Using Recurrent Neural Networks. In Audio/Visual Emotion Challenge and Workshop (AVEC'18), October 22, 2018, Seoul, Republic of Korea, 8 pages. https://doi.org/10.1145/3266302.3266305

1 INTRODUCTION

In recent years, emotion recognition has become more and more important in the research topic area. Automatic estimation of emotional state has a wide application in human-computer interaction, since it enables machines to well understand humans' spontaneous affective state just as human beings do [1]. Over the past decade, numerous research efforts have been made to build an effective and robust recognition model, leading to a great achievement [2][3][4].

Prior works propose a variety of feature sets and models for training an emotion recognition system. Audio modality plays a key role in emotion recognition especially in arousal dimension and various acoustic low-level descriptors (LLDs) [5] are proposed. The researchers also attempt to use deep audio features extracted from deep neural networks, for example DNN [6] and CNN [7]. Empirically, facial expressions have an important influence on valence dimension. Traditional visual features, such as Local Binary Gabor Patterns (LGBP) [8], geometric features, multi-scale dense SIFT, have achieved good performance in dimensional emotion recognition [9]. Besides, deep visual features extracted from VGG [10], ResNet [9], DenseNet [10] are widely applied. The AVEC 2015 [11] explore robust emotional models from physiological signals in comparison with the traditional audio-visual models. Keren et al. [12] exploit a series of convolutional and recurrent neural

networks to improve the performance of dimensional emotion recognition from physiological signals.

Dimensional emotion is continuous temporal dynamic process related closely to contextual information. Therefore, successful application of temporal information has a critical impact on performance improvement of emotion recognition. Chao et al. [2] use Deep Belief Network (DBN) with temporal pooling and multimodal-temporal fusion, which demonstrates that high level temporal fusion can improve performance. Wöllmer et al. [13] utilize Long Short-Term Memory Recurrent Neural Network (LSTM-RNN) to perform regression analysis on arousal and valence dimension, which improves recognition performance significantly. Compared with other models, LSTM-RNN can capture the temporal information of the emotional dimension better to achieve superior performance [4].

Benefited from the complementarity of different modalities, multimodal emotion fusion can achieve significant performance improvement. Feature level fusion and decision level fusion strategies are widely utilized [14]. Feature level fusion extracts the features from every modality separately and then concatenates them into feature vector for final emotion recognition [15]. However, feature level fusion suffers from the curse of dimensionality and demands a strict time synchrony between the modalities. Decision level fusion assumes each modality is independent and builds separate emotion recognition models, then combines the predictions of different modalities to train a second level model [16]. In this paper, we utilize these two fusion methods to compare their performance.

However, emotions are naturally ambiguous [17]. It is a challenging task to describe and predict the emotional state accurately, due to the human observer's subjectivity when perceiving the emotional state of others [18]. In order to obtain reliable labels, it requires multiple annotators to perceive an emotional audio and/or video. Therefore, it is worth exploring how to process original multiple annotations to generate more reliable ground truth and improve the performance. Let $\boldsymbol{\mathcal{D}} = \{\boldsymbol{x_n}, \boldsymbol{y_n}\}_{\boldsymbol{n=1}}^{\boldsymbol{N}}$ be a dataset of size $\boldsymbol{N}$, where for each input feature $\boldsymbol{x_i} \in \mathcal{X}$ and $\boldsymbol{y_i} \in \mathcal{Y}$. We are given a vector of crowdsourced annotations $\boldsymbol{y_n} = \{\boldsymbol{y_n^r}\}_{\boldsymbol{r=1}}^{\boldsymbol{R}}$, with $\boldsymbol{y_n^r}$ representing the annotation provided by the $\boldsymbol{r^{th}}$ annotator in a set of $\boldsymbol{R}$ annotators. The task is to learn a function $f: \mathcal{X} \rightarrow \mathcal{Y}$ which generalizes well on unseen data.

The most straightforward way is to use heuristic metrics by performing majority voting or calculating the mean values among all available annotations [19][20]. Despite being easy to compute, these metrics may not provide an accurate representation for the ground truth. Actually, different annotators have different levels of expertise, it is essential to consider how reliable different annotators are. Besides, the annotation process is not only influenced by human factors (such as the subjectivity of annotators, their age, fatigue and stress) but also the fuzziness of the meaning associated with various labels related to human behaviors. The issue becomes more prominent when the task is temporal, as it renders the labeling procedure vulnerable to temporal lags caused by varying response times of annotators. Fig. 1 shows valence emotional time series annotations from six annotators along with video stills, which illustrates there exists many difficulties like spike noise, annotation lags and bias.

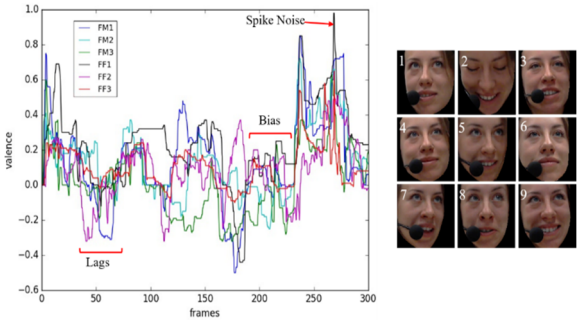


Figure 1: Valence annotations along with video stills.

The Audio-Visual Emotion Challenge (AVEC), an annual challenge since 2011, aims at promoting the development of multimedia processing and machine learning methods for automatic continuous emotion recognition in the wild. It provides a framework for non-acted spontaneous emotion recognitions and a fair benchmark to evaluate various emotion recognition methods. The AVEC 2018 Gold-standard Emotion Sub-Challenge (GES) [22] is a new task creating a single time series of emotion labels, usually referred as "gold-standard", from a pool of time-continuous annotations of dimensional emotions provided by several annotators.

In the following, Section 2 introduces the related works. Section 3 briefly introduces the database. Section 4 presents feature sets adopted in this challenge. The analysis and modeling of continuous emotion recognition are elucidated in Section 5. Section 6 convers the details of the entire experiments and results. Section 7 concludes this paper.

2 RELATED WORKS

Actually, learning from multiple annotations is not unique to continuous emotion recognition. In order to annotate large sets of data in a scalable manner, crowdsourcing [21] has established itself as an efficient and cost-effective solution. But it often requires aggregating labels from multiple noisy contributors with different levels of expertise.

Many approaches have been proposed to mitigate the effects of the noise and biases inherent to such heterogeneous sources of data from multiple annotators in different paradigms. One of the key early contributions is Dawid and Skene's work of [23], who propose an Expectation Maximization (EM) [24] algorithm to obtain point estimates of the error rates of patients given repeated but conflicting responses to medical questions. This work was the basis for many other variants. Whitehill et al. [25] extend their work by accounting for item difficulty in the context of image classification. Similarly, Ipeirotis et al. [26] propose to extract a single quality score for each worker that allows to prune low-quality workers.

Dawid and Skene's work [23] just focus on estimating the ground truth from multiple noisy labels. Smyth [27] propose to

first estimate the ground truth (without using the features) and then use the probabilistic ground truth to learn a classifier. Recent works show that jointly learning the classifier model and the annotators noise model relying on latent variable models generally leads to improved results. Raykar et al. [28] propose an approach for jointly learning the levels of expertise of different annotators and the parameters of a logistic regression classifier, by modeling the ground truth labels as latent variables. This work was later extended in [29] by considering the dependencies of the annotators' labels on the instances they are labeling. Rodrigues et al. [30] propose a new probabilistic model for supervised learning with multiple annotators where the reliability of the different annotators is treated as latent variables. Groot et al. [31] propose an extension of Gaussian processes to do regression in a multiple annotator setting.

When learning from multiple annotations in dimensional emotion recognition, some researches have addressed a few of these problems. Grimm et al. [32] introduced Evaluator Weighted Estimator (EWE), which considers inter-evaluator agreement to weight individual annotations and meanwhile filter out unreliable evaluators to improve the robustness of the results. Mariooryad et al. [33] align the annotator ratings by adjusting delays using mutual information between the features and every annotator's ratings. Feng et al. [34] present Canonical Time Warping (CTW) for accurate spatiotemporal alignment of facial expressions, which accommodates for subject variability and allows temporal local transformations. Following this work, Nicolaou et al. [35] assume that there is a latent space shared by annotator ratings and identify it using dynamic probabilistic Canonical Correlation Analysis (CCA) model with time warping. Another advanced derivation has recently been proposed for emotion recognition [36]. Motivated by Raykar's work [28], Gupta et al. [37] compute the ground truth by modeling annotators' specific distortions. They assume that the ground truth can be computed using the features based on a "feature mapping function", and the annotators process the latent ground truth based on annotator specific "distortion functions" to provide their ratings.

Another thought introduces emotion prediction uncertainty. Han et al. [38] point out that the uncertainty remains in the emotion target among several human annotators. They turn continuous emotion recognition task identified by a "hard" unique value to "soft" emotion prediction, with an additional target to indicate the uncertainty of human perception based on inter-rater disagreement level. Ting et al. [39] develop a multi-rater Gaussian Mixture Regression (GMR) model that incorporates multi-rater information to predict emotion uncertainty, under the assumption that multi-ratings reflect the uncertainty. On basis of [39], Ting et al. [40] explore emotional temporal dependencies of emotion uncertainty using Kalman filters. This work is realized by incorporating feedforward and backward Kalman filters into GMR to estimate the time-dependent label distribution that reflects the emotion uncertainty.

Recently, Guan et al. [41] propose an approach for training deep neural network that exploits information about the annotators. This approach is two-stage procedure, and first step is to model the multiple experts individually in the neural network, then while keeping their predictions fixed, independently learning average weights to combine them using backpropagation. Further, Rodrigues et al. [42] propose a novel general-purpose crowd layer to train deep neural networks directly from the noisy labels of multiple annotators. In this paper, we introduce the thought of crowd layer to model multiple time series annotations for continuous emotion recognition.

3 DATABASE

The challenge is evaluated on the RECOLA dataset [43]. In this dataset, the subjects are recorded by audio, video, electro-cardiogram (ECG) and electro-dermal activity (EDA) modalities. Spontaneous and naturalistic interactions are collected from 27 French-speaking subjects when they try to solve a collaborative task. This dataset is annotated by six French-speaking assistants in two emotion dimensions including arousal and valence. The organizers utilize a normalization technique based on the EWE [32] to get the gold standard as the ground truth. The individual ratings having six label values and the generated gold standard having only one label value are both available.

4 MULTI-MODAL EMOTION FEATURES

4.1 Acoustic Features

This challenge adopts eGeMAPS [5] as the baseline audio features. The eGeMAPS is an expert-knowledge based feature set consisting of 23 acoustic low-level descriptors (LLDs) such as energy, spectral and ceptral features, pitch, voice quality and micro-prosodic features. Overall, the acoustic baseline feature sets contain 88 dimensional features. The extraction of the LLDs and the computation of the functionals are done using the openSMILE toolkit [44]. Another segment-level representation of acoustic features, bag-of-audio-words (BoAW) are also provided. The bag-of-words are introduced for text features originally, but have been successfully applied to others modalities, such as the video and the audio domain [45]. In the BoAW framework, the LLDs over a certain segment are first quantized using predefined templates of a codebook of 'audio words', then a histogram of the audio words occurring in the corresponding segment is created. For the BoAW generation, the openXBOW toolkit [46] is employed.

4.2 Visual Features

For visual modality, there are four features sets. The first one is appearance feature set represented by the LGBP-TOP features set [8], whose main components are kept by PCA. The second one is geometric features, which are derived from facial landmarks. Details can be found in the baseline paper [22]. The organizers also provide pixel coordinates of 49 landmarks, namely Facial Action Units (AUs). The last one is bag-of-video-words (BoVW) based on AUs. The frames which fail to extract features are replaced with neighboring successful frame.

4.3 Psychological Features

For psychological modality, the features are extracted from ECG and EDA signals. The ECG signals are filtered by 5th order Butterwoth bandpass. Then the features like heart rate (HR) and its measure of variability (HRV), zero-crossing rate and other statistical data are extracted. For EDA, different features are extracted, including skin conductance response (SCR), skin conductance level (SCL), and relative statistics from EDA, SCR and SCL. In this paper, we combine the psychological features together.

5 EMOTION RECOGNITION MODELS

The overview of system framework is shown in Fig. 2. The front part of our system is same as basic continuous emotion recognition system. The inputs are various features from audio, visual and psychological modalities. We adopt the LSTM-RNN based neural network with temporal pooling to achieve short level temporal modeling. Meanwhile, the factor of annotation delay is also considered. Then the dense layer maps the representation of LSTM layer to the output layer. We can get the predicted results for the output layer which are corresponding to the gold standard. The emphasis is that our system adds the crowd layer taking the output layer as input to learn from the noise labels of multiple annotators. The crowd layer learns an annotator-specific mapping from the output layer to the labels of the different annotators. The results from the crowd layer are corresponding to the multiple individual ratings. When the model training is finished, the results from the output layer are the estimates.

5.1 Temporal Model

In this paper, we utilize LSTM to model emotional temporal process because of its good ability of learning long-term dynamic information. Considering that every sample has dense labels, there exists redundant information among adjacent frames. Therefore, we utilize the temporal pooling to smooth the label and decrease label noises, same as the work [2][3]. The temporal pooling operation adds the window to average both the features and labels, which can get the statics of the successive frames to achieve short level temporal modeling. In view of the influence of annotation delay, we shift the features to the future for delay compensation.

5.2 Crowd Layer

The crowd layer is proposed by [42], which is a special type of neural network layer learning directly from the noisy labels of multiple annotators with end-to-end fashion. In this paper, the crowd layer is attached to the output layer. As a result, the former output layer becomes a bottleneck layer that is shared among all the annotators, as shown in Fig. 2. Actually, the system considers the ground truth as latent variables and multiple annotations as the variant of the ground truth with the assistant of the crowd layer. Therefore, we can take full advantage of multiple annotations to train emotional recognition models. The inputs of the whole system are multimodal features and the outputs are multiple individual ratings. When training the models, the crowd layer adjusts the gradients from the labels of different annotators according to their reliability. In doing so, the bottleneck layer of the network receives adjusted gradients from different annotators' labels which it aggregates and backpropagates further through the rest of the network. The function of the crowd layer is to account for unreliable annotators and correct systematic biases in their labeling. Moreover, all of that can be done naturally within the backpropagation framework through end-to-end manner. In this paper, the mapping from the output layer to the crowd layer is achieved by linear mapping, represented by $f(x) = Wx + b$, W is mapping matrix and b is bias.

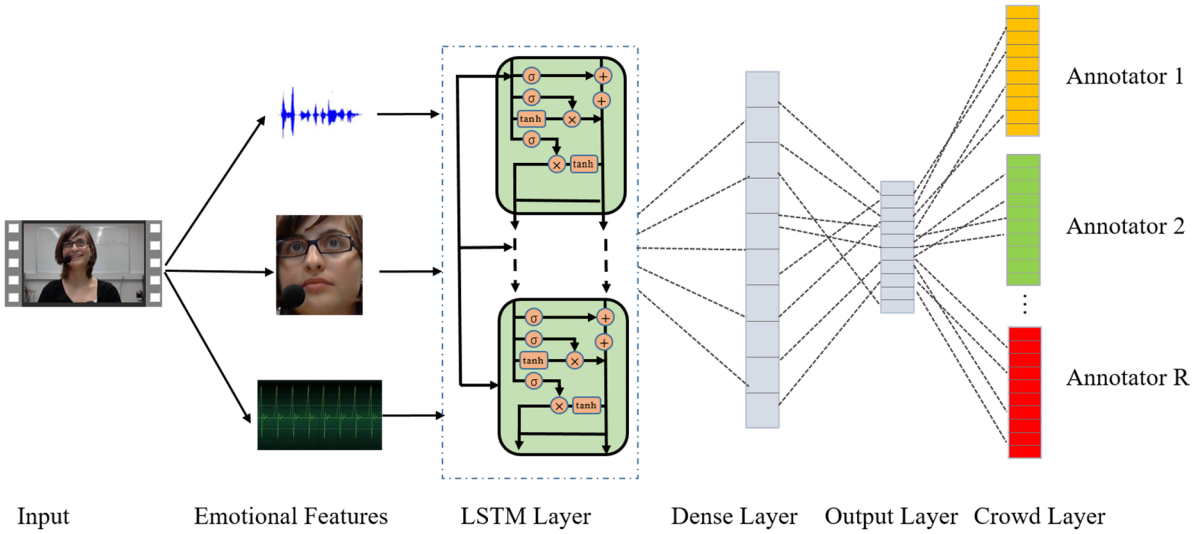


Figure 2: Overview of the proposed method. The features from different modalities are extracted. We utilize LSTM layer to achieve emotional temporal model. The dense layer maps the representation of LSTM layer to the output layer. Finally, the crowd layer takes the output layer as input to learn from the noise labels of multiple annotators.

5.3 Multimodal Fusion

Previous researches have verified that multimodal fusion can improve the performance of continuous emotion recognition. In this paper, we compare the performance of two fusion methods, namely feature level fusion and decision level fusion. Feature level fusion is achieved by concatenating different features in hidden layer. Suppose there are two input modalities represented by a_t, s_t separately. The fusion equation is below:

$$m_t = tanh(W_m [a_t, s_t] + b_m)$$

where W_m and b_m are the weight and bias in this layer.

We also consider decision level fusion. Each feature set is trained individually based on LSTM model. The estimates from different modalities are concatenated to obtain final emotion predictions using SVR in decision level fusion.

6 EXPERIMENTS AND ANALYSIS

6.1 Experimental setup

We use Ɛ-insensitive loss function to ignore small errors and assign absolute value loss to large errors, which is more suitable than the other loss functions [47]. Adadelta [48] optimization algorithm is utilized. Weight decay in the dense layer is also applied to prevent over-fitting. The hyper-parameters, the number of hidden layer nodes along with the delay time and the window length of temporal pooling, are chosen based on the performance on the development set by random combination. The maximum training epochs are 70. We also use dropout after LSTM with the rate 0.5. The results are evaluated using the concordance correlation coefficient (CCC), which combines the Pearson correlation coefficient of two times series with mean square error.

6.1 Unimodal Emotion Prediction

Firstly, we compare the performance of different features. The experiments utilize basic emotion recognition model, which includes the input layer, LSTM layer, dense layer and output layer without the crowd layer. Their inputs are mono-modality features and targets are the gold standard. The CCC performance is calculated between the predicted results from the output layer and the ground truth on the development set. Every feature is trained to select its own model respectively and the experimental results are shown in Table 1. The BoAW features obtain best performance 0.792 in arousal dimension and geometric features obtain best performance 0.615 in valence dimension. The BoAW features achieve comparable performance with the eGeMAPS features in arousal dimension, while the BoVW features don't obtain good performance in valence dimension. It is worth noting that the psychological features achieve better performance in valence dimension than arousal dimension. In total, the performance of arousal dimension is better than valence dimension, which is similar to previous works [13].

Next, we conduct experiments on different targets. As described above, the first one is basic emotion recognition model whose inputs are the emotional features and targets are the gold standard, represented by "Gold Standard" model. The second one is the model with crowd layer introduced in Fig. 2, whose inputs are the emotional features and targets are the individual ratings, represented by "Individual" model. It is important to note that, for "Individual" model, the CCC performance is also calculated between the results from the output layer (not from the crowd layer) and the ground truth on the development set after obtaining the trained models. Finally, the third one is similar to the second one, while the difference is that the gold standard and the individual ratings are both regraded as the targets, represented by "Multi-target" model. The performance evaluation method is same as the second one.

The experimental results, shown in Table 2, indicate that the performance of the "Gold Standard" model is better than the "Individual" model in audio and visual modalities. The performance of the "Multi-target" model is in the middle. It is understandable because the individual ratings contain original noisy labels of multiple annotations and the performance evaluation is based on the gold standard. Whereas, we expect the model learning from multiple individual ratings accounts for unreliable annotators and corrects systematic biases, thus having more generalization ability. Remarkably, for psychological features, the "Multi-target" model achieves best performance in arousal dimension and the "Individual" model achieves best performance in valence dimension. Compared with Table 1, the performance of psychological features is improved in both arousal and valence dimension, while not for audio and visual modalities.

Table 1: CCC performance of different features using basic emotion recognition model on the development set.

Features	Arousal	Valence
eGeMAPS	0.790	0.503
BoAW	**0.792**	0.332
Appearance	0.619	0.531
Geometric	0.598	**0.615**
AU	0.427	0.596
BoVW	0.238	0.469
Physio	0.233	0.378

Table 2: CCC performance of different modalities with different targets on the development set.

	Target	Arousal	Valence
eGeMAPS	Gold Standard	0.790	0.503
	Individual	0.767	0.466
	Multi-target	0.791	0.450
Geometric	Gold Standard	0.598	0.615
	Individual	0.520	0.607
	Multi-target	0.534	0.612
Physio	Gold Standard	0.233	0.378
	Individual	0.217	0.406
	Multi-target	0.243	0.421

Table 3: The CCC performance of different targets with feature level fusion for arousal and valence dimension on the development set and testing set.

Dimension	Target	Feature Combination	Development			Testing		
			RMSE	PCC	CCC	RMSE	PCC	CCC
Arousal	Gold Standard	eGeMAPS+ BoAW+ AU	0.116	0.808	0.805	0.138	0.659	**0.651**
	Individual	eGeMAPS+ BoAW+ AU	0.120	0.797	0.795	0.137	0.656	0.645
Valence	Gold Standard	Geometric+ AU+ eGeMAPS+ BoVW	0.099	0.655	0.651	0.108	0.592	**0.577**
	Individual	Geometric+ AU+ Appearance+ eGeMAPS+ BoVW+ Physio+ BoAW	0.101	0.644	0.641	0.109	0.575	0.564

Table 4: The CCC performance of different targets with decision level fusion for arousal and valence dimension on the development set and testing set.

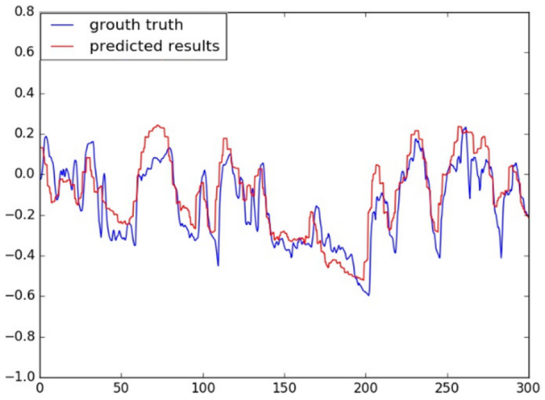
Dimension	Target	Feature Combination	Development			Testing		
			RMSE	PCC	CCC	RMSE	PCC	CCC
Arousal	Gold Standard	eGeMAPS+ BoAW+ AU	0.117	0.810	0.805	0.138	0.663	0.655
	Individual	eGeMAPS+ BoAW+ AU	0.111	0.827	0.822	0.130	0.704	**0.697**
Valence	Gold Standard	Geometric+ AU+ eGeMAPS+ Physio+ BoAW	0.092	0.681	0.655	0.099	0.632	**0.597**
	Individual	Geometric+ AU+ eGeMAPS	0.092	0.679	0.647	0.103	0.588	0.522

6.2 Multi-modal Emotion Prediction

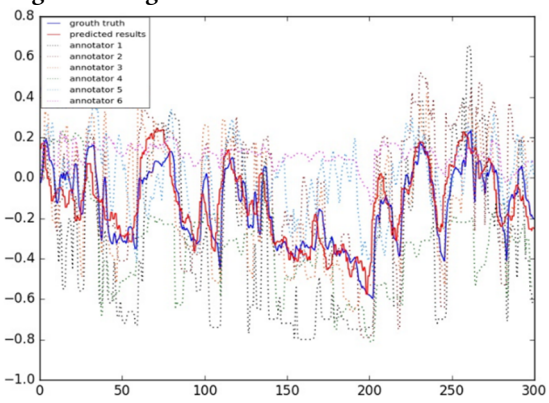
To obtain better system performance, we utilize multimodal fusion including feature level fusion and decision level fusion. Firstly, every feature set is trained separately to obtain its own best model. In order to select an optimal combination of feature sets, a greedy feature selection strategy is utilized. Firstly, we rank the feature sets according to their CCC performance on the development set. Then, the feature sets are added sequentially to the combination in order, and the feature set is retained if the performance of the development set increases, otherwise abandoned. The feature combination with the highest value is selected. Same feature selection strategy is utilized in both feature level fusion and decision level fusion, which is introduced in Section 5.3.

According to the described method above, the final feature combinations of different models for feature level fusion are shown in Table 3. We notice that the optimal feature combination of arousal dimension is similar to two models, including the eGeMAPS features, BoAW features and AU features which verities the effectiveness of acoustic features on arousal dimension. In valence dimension, the features combination of the "Gold Standard" model includes the Geometric features, AU features, BoVW features and eGeMAPS features. Whereas, the feature combination of the "Individual" model utilizes all available features. Table 3 also lists the performance of development set and testing set calculated by corresponding ground truth. The experimental results show that the performance of the "Gold Standard" models is better than the "Individual" model on testing set for arousal and valence dimensions. Unfortunately, these results don't live up to our expectations.

Then we make experiments with decision level fusion, as shown in Table 4. Notice that the optimal feature combination of arousal dimension is similar under all situations. In arousal dimension, the performance of the "Individual" model is better



(a) The predictions of the "Ground Standard" model against the ground truth.



(b) The predictions of the "Individual" model and corresponding ground truth against the ground truth.

Figure 3: The visualization of the predictions produced by different models against the ground truth and individual ratings in arousal dimension.

than the "Gold Standard" model, even better than the best performance of feature level fusion. However, the feature combination of valence dimension is different under different situations. For the "Gold Standard" model, the decision level fusion adds Phsio features compared with feature level fusion. The performance of decision level fusion is better than feature level fusion. Due to the limit of five trials, we don't obtain the experimental results of the "Multi-target" model on the testing set.

In general, our systems achieve superior performance. The performance of decision level fusion is better than feature level fusion. Our proposed method, the "Individual" model with crowd layer achieve best performance 0.697 in arousal dimension. As shown in Fig. 3, we take an example of development set to visualize the predicted results in arousal dimension. Fig. 3(a) shows the predictions using the "Ground Standard" model against the ground truth. Fig. 3(b) show the predictions using the "Individual" model and corresponding individual ratings against the ground truth. The results indicate that the "Individual" model can capture the reliability and biases of different annotators, resulting in performance improvement. The "Ground Standard" model achieves best performance 0.597 in valence dimension, which is also competitive results. This paper proposes new thought to deep learning from multiple annotations directly, with the addition of a novel crowd layer to improve the performance of continuous emotion recognition

7 CONCLUSIONS

This paper utilizes the crowd layer to train deep neural networks with end-to-end fashion, which enables us to directly learn from the labels of multiple emotional time series annotations. The system considers the ground truth as latent variables and multiple annotations as the variant of ground truth by linear mapping in deep neural network. Therefore, the crowd layer learns an annotator-specific mapping from the output layer to the labels of the different annotators. Despite its simplicity, the crowd layer is able to adjust the errors gradients that are backpropagated during training accordingly. The system adopts LSTM as basic emotion regression model by taking the factor of annotation delay and temporal pooling into consideration meanwhile. The experimental results show that acoustic features achieve better performance in arousal dimension, visual features achieve better performance in valence dimension and psychological features are useful for performance improvement. Decision level fusion achieves better performance than feature level fusion in arousal and valence dimension. Our proposed method achieves better performance in arousal dimension, which verifies its ability of capturing the reliability and biases of different annotators. The performance of valence dimension also achieves competitive results. This paper proposes new thought to deep learning from multiple annotations directly to improve the performance of continuous emotion recognition.

ACKNOWLEDGMENTS

This work is supported by the National Key Research & Development Plan of China (No. 2016YFB1001404), and the National Natural Science Foundation of China (NSFC) (NO.61425017, No.61773379, No.61332017, No.61603390, No.61771472) and the Major Program for the National Social Science Fund of China (13&ZD189).

REFERENCES

[1] R. Cowie, C. E. Douglas, N. Tsapatsoulis, et al. 2001. Emotion recognition in human-computer interaction. IEEE Signal processing magazine, 18(1): 32-80.
[2] L. Chao, J. Tao, M. Yang, et al. 2014. Multi-scale temporal modeling for dimensional emotion recognition in video. Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge. ACM, 11-18.
[3] J. Huang, Y. Li, J. Tao, et al. 2017. Continuous Multimodal Emotion Prediction Based on Long Short Term Memory Recurrent Neural Network. Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge. ACM, 11-18.
[4] S. Chen, Q. Jin, J. Zhao, et al. 2017. Multimodal Multi-task Learning for Dimensional and Continuous Emotion Recognition. Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge. ACM, 19-26.
[5] F. Eyben, K. R. Scherer, B. W. Schuller, et al. 2016. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. IEEE Transactions on Affective Computing, 7(2): 190-202.
[6] A. Popková, F. Povolný, P. Matějka, et al. 2016. Investigation of Bottle-Neck Features for Emotion Recognition. International Conference on Text, Speech, and Dialogue. Springer International Publishing, 426-434.
[7] Y. Aytar, C. Vondrick, A. Torralba. 2016. Soundnet: Learning sound representations from unlabeled video. Advances in Neural Information Processing Systems, 892-900.
[8] T. R. Almaev, M. F. Valstar. 2013. Local gabor binary patterns from three orthogonal planes for automatic facial expression recognition. Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on. IEEE, 356-361.
[9] B. Sun, S. Cao, L. Li, et al. 2016. Exploring multimodal visual features for continuous affect recognition. Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge. ACM, 83-88.
[10] S. Chen, Q. Jin, J. Zhao, et al. 2017. Multimodal Multi-task Learning for Dimensional and Continuous Emotion Recognition. The Workshop on Audio/visual Emotion Challenge. ACM, 19-26.
[11] F. Ringeval, B. Schuller, M. Valstar, et al. 2015. AVEC 2015: The 5th international audio/visual emotion challenge and workshop. Proceedings of the 23rd ACM international conference on Multimedia. ACM, 1335-1336.
[12] G. Keren, T. Kirschstein, E. Marchi, et al. 2017. End-to-end learning for dimensional emotion recognition from physiological signals. Multimedia and Expo (ICME), 2017 IEEE International Conference on. IEEE, 985-990.
[13] M. Wöllmer, M. Kaiser, F. Eyben, et al. 2013. LSTM-Modeling of continuous emotions in an audiovisual affect recognition framework. Image and Vision Computing, 31(2): 153-163.
[14] H. Gunes, M. Pantic, A. S. Ashour. 2010. Automatic, Dimensional and Continuous Emotion Recognition. International Journal of Synthetic Emotions, 1(1):68-99.
[15] R. Viktor, A. Sankaranarayanan, S. Shirin, K. Rohit, et al. 2012. Emotion recognition using acoustic and lexical features. In Thirteenth Annual Conference of the International Speech Communication Association.
[16] H. Zhaocheng, D. Ting, C. Nicholas, et al. 2015. An investigation of annotation delay compensation and output-associative fusion for multimodal continuous emotion prediction. In Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge, 41-48.
[17] E. Mower, A. Metallinou, C. C. Lee, et al. 2009. Interpreting ambiguous emotional expressions. Affective Computing and Intelligent Interaction and Workshops, ACII 2009, 3rd International Conference on. IEEE, 1-8.
[18] J. Brendan, B. Subhabrata, and F. C. Shih. 2014. Predicting viewer perceived emotions in animated GIFs. In Proc. ACM International Conference on Multimedia. Orlando, FL, 213–216.
[19] L. Ya, T. Jianhua, S. Björn, et al. 2016. MEC 2016: The multimodal emotion recognition challenge of CCPR 2016. In Proc. Chinese Conference on Pattern Recognition. Chengdu, China, 667–678
[20] S. Björn. 2015. Speech analysis in the big data era. In Proc. International Conference on Text, Speech, and Dialogue. Pilsen,Czech Republic, 3–11.
[21] T. S. Sindlinger. 2010. Crowdsourcing: Why the Power of the Crowd is Driving the Future of Business.
[22] R. Fabien, B. Schuller, M. Valstar, et al. 2018. AVEC 2018 Workshop and Challenge: Bipolar Disorder and Cross-Cultural Affect Recognition.

Proceedings of the 8th International Workshop on Audio/Visual Emotion Challenge. ACM 2018.
[23] A. P. Dawid, A. M. Skene. 1979. Maximum likelihood estimation of observer error-rates using the EM algorithm. Applied statistics, 20-28.
[24] P. D. Arthur, M. L. Nan, B. R. Donald. 1977. Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society. Series B (Methodological).
[25] J. Whitehill, T. Wu, J. Bergsma, et al. 2009. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. Advances in neural information processing systems, 2035-2043.
[26] P. G. Ipeirotis, F. Provost, J. Wang. 2010. Quality management on amazon mechanical turk. Proceedings of the ACM SIGKDD workshop on human computation, 64-67.
[27] P. Smyth, U. Fayyad, M. Burl, et al. 1995. Learning with probabilistic supervision. Computational learning theory and natural learning systems, 3: 163-182.
[28] V. C. Raykar, S. Yu, L. H. Zhao, et al. 2010. Learning from crowds. Journal of Machine Learning Research, 11(Apr): 1297-1322.
[29] Y. Yan, R. Rosales, G. Fung, et al. 2014. Learning from multiple annotators with varying expertise[J]. Machine learning, 95(3): 291-327.
[30] F. Rodrigues, F. Pereira, B. Ribeiro. 2013. Learning from multiple annotators: distinguishing good from random labelers. Pattern Recognition Letters, 34(12): 1428-1436.
[31] P. Groot, A. Birlutiu, T. Heskes. 2011. Learning from multiple annotators with Gaussian processes. International Conference on Artificial Neural Networks. Springer, Berlin, Heidelberg, 159-164.
[32] M. Grimm, K. Kroschel. 2005. Evaluation of natural emotions using self-assessment manikins. Automatic Speech Recognition and Understanding, 2005 IEEE Workshop on. IEEE, 381-385.
[33] M. Soroosh and B. Carlos. 2015. Correcting time-continuous emotional labels by modeling the reaction lag of evaluators. Affective Computing, IEEE Transactions on, vol. 6, no. 2, pp. 97–108.
[34] Z. Feng and D. Fernando. 2009. Canonical time warping for alignment of human behavior. In Proc. Advances in Neural Information Processing Systems. Vancouver, Canada, 2286–2294.
[35] A. N. Mihalis, P. Vladimir, P. Maja. 2012. Dynamic probabilistic cca for analysis of affective behavior. in Computer Vision–ECCV 2012, pp. 98–111. Springer.
[36] Z. Feng and D. Fernando. 2016. Generalized canonical time warping. IEEE Transactions on Pattern Analysis and Machine Intelligence 38, 2, 279–294.
[37] R. Gupta, K. Audhkhasi, Z. Jacokes, et al. 2018. Modeling multiple time series annotations as noisy distortions of the ground truth: An Expectation-Maximization approach. IEEE transactions on affective computing, 9(1): 76.
[38] J. Han, Z. Zhang, M. Schmitt, et al. 2017. From Hard to Soft: Towards more Human-like Emotion Recognition by Modelling the Perception Uncertainty. Proceedings of the 2017 ACM on Multimedia Conference, 890-897.
[39] T. Dang, V. Sethu, J. Epps, et al. 2017. An Investigation of Emotion Prediction Uncertainty Using Gaussian Mixture Regression. Proc. Interspeech 2017, 1248-1252.
[40] T. Dang, S. Vidhyasaharan, A. Eliathamby. 2018. Dynamic multi-rater Gaussian mixture regression incorporating temporal dependencies of emotion uncertainty using kalman filters. ICASSP 2018
[41] M. Y. Guan, V. Gulshan, A. M. Dai, et al. 2017. Who said what: Modeling individual labelers improves classification. arXiv preprint arXiv:1703.08774.
[42] F. Rodrigues, F. Pereira. 2017. Deep learning from crowds. arXiv preprint arXiv:1709.01779.
[43] F. Ringeval, A. Sonderegger, J. Sauer, D. Lalanne. 2013. Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In Proc. Of EmoSPACE, FG, Shanghai, China.
[44] F. Eyben, M. Wöllmer, B. Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. Proceedings of the 18th ACM international conference on Multimedia. ACM, 1459-1462.
[45] F. Ringeval, B. Schuller, M. Valstar, et al. 2017. AVEC 2017: Real-life Depression, and Affect Recognition Workshop and Challenge. Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge. ACM, 3-9. 3.2 Visual Features.
[46] S. Maximilian, B. Schuller. 2016. openXBOW – Introducing the Passau open-source crossmodal Bag-of-Words toolkit. preprint arXiv:1605.06778.
[47] L. Chao, J. Tao, M. Yang, et al. 2015. Long short term memory recurrent neural network based multimodal dimensional emotion recognition. Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge. ACM, 65-72.
[48] M. D. Zeiler. 2012. ADADELTA: an adaptive learning rate method. arXiv preprint arXiv:1212.5701.