Real-Time Video Emotion Recognition based on Reinforcement Learning and Domain Knowledge

Ke Zhang, Yuanqing Li, Jingyu Wang, *Member, IEEE,* Erik Cambria, *Fellow, IEEE,* Xuelong Li, *Fellow, IEEE,*

***Abstract*—Multimodal emotion recognition in conversational videos (ERC) develops rapidly in recent years. To fully extract the relative context from video clips, most studies build their models on the entire dialogues which make them lack of real-time ERC ability. Different from related researches, a novel multimodal emotion recognition model for conversational videos based on reinforcement learning and domain knowledge (ERLDK) is proposed in this paper. In ERLDK, the reinforcement learning algorithm is introduced to conduct real-time ERC with the occurrence of conversations. The collection of history utterances is composed as an emotion-pair which represents the multimodal context of the following utterance to be recognized. Dueling deep-Q-network (DDQN) based on gated recurrent unit (GRU) layers is designed to learn the correct action from the alternative emotion categories. Domain knowledge is extracted from public dataset based on the former information of emotion-pairs. The extracted domain knowledge is used to revise the results from the RL module and is transformed into other dataset to examine the rationality. The experimental results on datasets show that ERLDK achieves the state-of-the-art results on weighted average and most of the specific emotion categories.**

***Index Terms*—Multimodal Emotion Recognition, Reinforcement Learning, Domain Knowledge, Real-time Video Conversation.**

I. Introduction

EMOTION recognition technology keeps a high speed of development due to the continue attention from researchers. In order to better adapt to daily application scenarios, like smart home, mental illness care, education aids and car-hailing services [1], many recent studies are not satisfied with simply classifying a single utterance, sentence or article using traditional emotion categorization models. The analysis of dialogues, especially conversational videos, has become increasingly popular [2].

However, it is not appropriate to perform emotion recognition in conversational videos (ERC) by just transforming algorithms and results from original emotion recognition issues [3]. The emotion states always change flexibly and frequently in ERC with contents and focuses of the present dialogue and speakers. People are inclined to express their emotions implicitly through some habitual ways during interpersonal conversations, which are different from the official circumstances [4]. Misunderstandings can be caused even between humans if they were unfamiliar with each other, not to mention computers. Therefore, recognitions on article level or paragraph level are imprecisely and lack of meaning because of the too long span [5]. Motivated by these challenges, a large number of methods and theories for ERC have emerged in recent years [6].

The unique characteristics of ERC are summarized as context dependence, persistence and contagiousness [7] that the static and dynamic flows in a daily dialogue are both included. The context dependence emphasizes the reliance on contextual information other than single utterance or sentence. The persistence means that the interlocutors are inclined to maintain their original emotional states during the conversations and the third, contagiousness, describes the emotion states of interlocutors are interactive which can be influenced or be pushed into a different new state by others. However, it is not easy for computer algorithms to give most appropriate considerations to all these three characteristics of ERC like humans [8]. To capture the contextual emotional changes as much as possible during actual research processes, majority studies choose to build models on whole dialogue range [9] and extra layers are designed to track the flow of emotion states of each interlocutor throughout the conversations [10]. However, rare researches pay the same attention on the real-time performance of ERC as on improving the recognition accuracy, despite the many benefits of real-time capability like helping the computer show better responses according to the emotion of speakers on current time spot. Firstly, to maintain the real-time performance and include contextual information simultaneously [11], the concept of emotion-pair is defined as the smallest emotion stage unit of a conversation.

Motivated by the issues mentioned above, a multimodal emotion recognition model for conversational videos based on reinforcement learning (RL) and domain knowledge (ERLDK) is proposed in this paper. Text, visual and audio are used in ERLDK as the multimodal input since they are the three most common expression modals in both videos and human daily life. Our model consists of three modules and the overview structure is depicted in Figure 1.

This work was supported in part by the National Natural Science Foundation of China under Grant 61871470, Grant U1801262 and Grant 61976179; in part by the Fundamental Research Funds for the Central Universities under Grant 3102019HTXM005 and Grant 3102017HQZZ003.

K. Zhang and Y. Li are with National Key Laboratory of Aerospace Flight Dynamics and School of Astronautics, Northwestern Polytechnical University, Xian 710072, P.R.China (e-mail: zhangke@nwpu.edu.cn; yuanqingli@mail.nwpu.edu.cn).

J. Wang is with the School of Astronautics, School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xian 710072, P.R.China (e-mail: jywang@nwpu.edu.cn).(*corresponding author: Jingyu Wang.*)

E. Cambria is with the School of Computer Science and Engineering, Nanyang Technological University, Singapore (e-mail:cambria@ntu.edu.sg).

X. Li is with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xian 710072, P.R.China (e-mail:xuelong_li@nwpu.edu.cn).

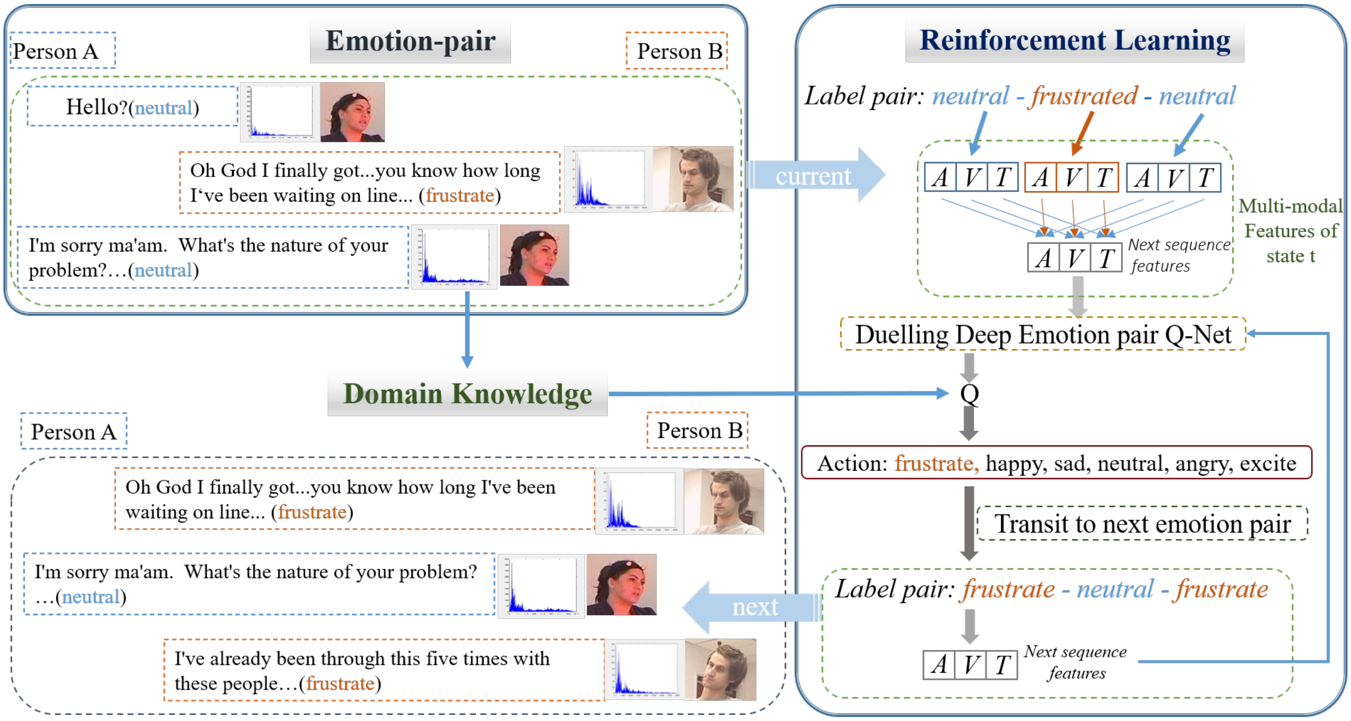

Emotion-pair
Person A
Person B
Hello?(neutral)
Oh God I finally got...you know how long I've been waiting on line... (frustrate)
I'm sorry ma'am. What's the nature of your problem?…(neutral)
current
Reinforcement Learning
Label pair: neutral - frustrated - neutral
A V T A V T A V T
A V T
Next sequence features
Multi-modal Features of state t
Duelling Deep Emotion pair Q-Net
Domain Knowledge
Q
Action: frustrate, happy, sad, neutral, angry, excite
Transit to next emotion pair
Person A
Person B
Oh God I finally got...you know how long I've been waiting on line... (frustrate)
I'm sorry ma'am. What's the nature of your problem? …(neutral)
I've already been through this five times with these people…(frustrate)
next
Label pair: frustrate - neutral - frustrate
A V T
Next sequence features

Fig. 1. The overview of the proposed ERLDK.

The emotion-pair is combined of multiple continuous sentences sampled from both sides of interlocutors with a fixed window length and the emotion of the coming next sentence from one of the interlocutors becomes the target waiting to be recognized. At the same time, the correlations between the emotion-pair and the target are summarized as the domain knowledge during this step. For real-time ERC, the sentences after the target utterance are not considered, because the following dialogues all happen in the future which cannot be used to help the recognition of the present. Secondly, based on the definition of the emotion-pair, the reinforcement learning (RL) module learns the transformation of the emotion states. In this module, different modalities are fused at feature level so that the fusion of unimodal between utterances comes before the fusion with other modalities.

To enhance the ability of ERC, the recognition results of the other states are taken into consideration while recognizing the emotion of the current state by utilizing the dueling deep-Q-network (DDQN) [12]. The structure of DDQN in this RL module is depicted in Figure 2 and the sample window is set to be three as an example. Followed by the RL modules, the third module uses domain knowledge of emotion-pair and the target to revise the recognition results from the RL module. However, as emotion is closely linked to each human's character, the actual situation is impossible to summarize in total as commonsense knowledge is not absolute [13].

Therefore, the first emotion-pair is utilized as the base domain knowledge of the current dialogue atmosphere to revise the following recognize outputs sentence by sentence. During this process, the revision effects of the domain knowledge will gradually decrease along with the accumulated recognize deviation brought by the RL module, but our method still achieves superiority over the baselines, especially on dialogues with the length less than 30 steps. The main contributions of this paper are listed as below:

- A new multimodal emotion recognition model for conversational videos based on reinforcement learning and domain knowledge (ERLDK) is proposed in this paper. To the best of our knowledge, this is the first time that RL and domain knowledge are combined for real-time ERC. The emotion of speakers are recognized on dialogue level as the conversations progress step by step. Text, visual and audio information are fused as multimodal inputs.
- Human daily emotion transformation habits are innovatively extracted as the domain knowledge of the emotion recognition for conversational videos in ERLDK.Three window sizes for extraction in dialogues are experimented that the size of three turns out to be the most appropriate.
- Experiments on two public datasets are carried out to examine the performance of ERLDK. The relationship between recognition ability and the changes of conversation lengths is also studied during the test. The results surpass the state-of-the-art baselines on most emotion categories.

The rest of this paper is structured as follows: Section II briefly reviews important previous work in multimodal emotion recognition in conversations. Section III describes our method in detail. Section IV presents the experimental setups. Section V illustrates results and discussions. Finally, Section VI provides concluding remarks and directions for future work.

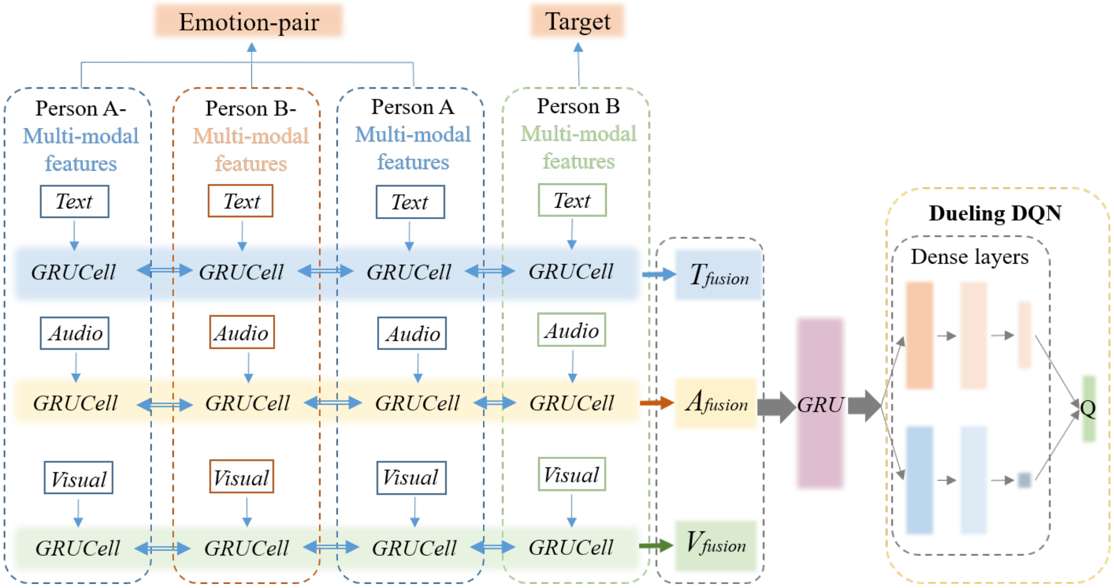


Fig. 2. The structure of the dueling DQN of RL module in ERLDK with sample size as three.

II. Related Work

ERC research has become popular and attractive in both the scientific community and the business world. ERC studies can be mainly divided into two categories as unimodal emotion recognition and multimodal emotion recognition according to the number of the input modalities.

A. Unimodal Emotion Recognition

Text is one of the most frequently unimodal used for ERC due to its advantages of expressing the conversation information clearly and continuously [14], [15]. Peng *et al.* [16] proposes a text based model which fuses the word-level and sentence-level features to learn emotional expression which focuses on humanmachine dialogue systems. Majumder *et al.* [17] proposes a Recurrent Neural Networks (RNN) based model named DialogueRNN that can not only analyze the contexts information but also keeps track of each individual party emotion state. Ghosal *et al.* [18] presents a Dialogue Graph Convolutional Network (DialogueGCN) model based on former DialogueRNN model. DialogueGCN solves the context propagation issues of DialogueRNN by leveraging the dependency of all the interlocutors. DialogueRNN and DialogueGCN are superior methods that conversation information are computed on sentence level and interlocutor level to capture the emotion habits of each person. Both of them can be utilized on multispeakers dialogue scenes and achieves competitive experiments results.

Some other unimodal approaches ERC include works that focus on audio [19], visual [20] and electroencephalogram (EEG) [21]. Francesca *et al.* [22] transcribes and analyzes audio conversations with verona coding definitions of emotional sequences to establish a quantitative relationship between asymmetrical variables. Bryan *et al.* [23] utilizes its proposed model based on audio features into health psychology field. Huang *et al.* [24] focuses on nonverbal sounds which naturally exists in our daily conversation. Verbal and nonverbal segments within an utterance are extracted by a Prosodic Phrase (PPh) auto-tagger and an attentive long short-term memory (LSTM)-based sequence-to-sequence model. Unlike audio signals in which tone-related information are paid mainly attention [25], visual pictures are more intuitive [26] and have better readability [27]. Tao *et al.* [28] proposes a two-stage module to find the low-dimensional tensor subspace and computing the spectral of the face tensors. Hofmann *et al.* [29] investigates the elicitation of smiling and laughter and the role of facial display regulation markers in positive emotions during conversations. Emotion-related features are encoded by Ryu *et al.* [30] that the robustness of edge patterns in the edge region are taken into consider with the smooth regions. Deep features and handcraft features of multiple views are innovatively combined in a simple but effective way by Tao *et al.* [31] for person re-identification. Different strategies are utilized for both deep features and handcraft features.Both audio and visual signals mentioned above are common and frequently used in daily life [32], however, these information are likely to be contradictory or deceptive when they are analyzed on their own [33]. On the contrary, EEG signals can make up for these ambiguities by recording the ongoing neuronal activities of the brain. Gupta *et al.* [34] proposes an effective method based on flexible analytic wavelet transform (FAWT) for recognition of emotion through the investigation of the channel specific nature of EEG. EEG performs better accuracy in many cases, however, it is not easy to obtain which makes it hard to be applied to daily life conveniently.

Unimodal ERC method achieves a lot of remarkable results and has its own application fields. However, using unimodal for ERC is insufficient and unstable in several cases, for example, faces can be blocked [35] and audio signals may mislead the results without the text [36]. Hence, the research focus of ERC is attempting for multimodal methods now.

B. Multimodal Emotion Recognition

The expression of human during conversations is the comprehensive result of multiple behavioral patterns [37], therefore, there are many benefits to use multimodal as the inputs [38]. RNN and RNN-like networks are mainstream algorithms due to their significant advantages in processing sequence information [39]. Zhang *et al.* [40] builds a quantum-like multimodal network (QMN), which uses quantum theory (QT) and a LSTM network to analyze sentiment in conversations. Text and image are used as the multimodal inputs. Some novel methods also choose to improve the performances by integrating RNN with other algorithms and other relative subjects knowledge [41]. Plaza *et al.* [42] recognizes emotions by integrating different affective lexical knowledge from Spanish social media with neural networks. Wang *et al.* [43] proposes a multimodal deep regression Bayesian network (MMDRBN) to compute the relationship between audio and visual modalities for emotion recognition and domain knowledge from videos are incorporated. However, the contextual information and the habitual of human are not included in these domain knowledges. Besides combing domain knowledge, the relationship between speakers are alao helpful during ERC. Xing *et al.* [11] proposes an Adapted Dynamic Memory Network (A-DMN) in which regards the self and cross-speaker influence as two mainly points to the emotion flows of conversation. This method uses audio, visual and text inputs and receives great results on public datasets. However, these multimodal ERC methods rely deeply on using whole dialogue information to improving accuracy that only a few of recent models pays attention on the real-time capability of ERC.

Hazarika *et al.* [44] introduces a conversational memory network (CMN) that comprises audio, visual and text features with gated recurrent units. The sample window of the dialogue is set to be eight that each speaker memories four sentences before recognition the next sentence. Lai *et al.* [45] proposes a different contextual window sizes based recurrent neural networks (DCWS-RNNs) model. Four different RNNs are designed to compute the contexts separately that two for utterances before the target and the other two for utterances after the target. These kind of methods show competitive results with good real-time recognition capability for it does not need the whole dialogue information as assistance, but they still have a certain delay that three utterances after the target are required.

C. Reinforcement Learning

To ensure the recognition accuracy as well as the real-time performance, RL is applied as the main module in this paper. Recognizing emotion using RL [46] and knowledge-base system [47] are not new things actually. Liu *et al.* [48] proposes a Reinforcement Online Learning (ROL) method for real-time emotion state prediction by using EEG. This method applies the ROL to least square (LS) and support vector regression (SVR) for emotion prediction. Li *et al.* [49] proposes a RL model for pre-selecting (RLPS) useful images for facial emotion recognition, which is made up of an image selector and a rough emotion classifier.

RL has its unique advantages on imitating the real-time conversion of the conversations [50] and the domain knowledge in conversations [51] is also meaningful because the emotion of speakers is relatively following regular emotion inertia [52]. However, few of them try to combine RL and domain knowledge in real-time ERC, even less using different modalities as inputs [53], [54].

III. Methodology

In this section, the ERLDK model for multimodal emotion recognition in video conversations is introduced.

A. The preprocess module

This module is the first module of the ERLDK model which is responsible for generalizing emotion-pair and corresponding domain knowledge. To recognize the emotion category of the target utterance inside a specific dialogue, the ability of properly combining contextual information is the key point. The concept of the emotion-pair is defined to represent the effective contextual information of the target utterance. The emotion-pair consists of several utterances before the target with a fixed sample length, in which the emotion persistence and contagious are potentially exposed. For example, when the sample length is set to be three, every three utterances are packed as an emotion-pair in the order of their appear sequences from the start to the end in a dialogue. During this step, both the multimodalities and the label of these three utterances are recoded in the new processed emotion-pair dataset. To find the best choice of this sample size, four different lengths are examined to make a balance between enough contexts and less redundancy. These four examined lengths are two utterances, three utterances, four utterances and 5 utterances before target respectively and the sampled corresponding emotion-pairs will construct the new trainsets and testsets.

The corresponding domain knowledge under each length is computed based on these new datasets under emotion-pair form. The labels of a emotion-pair with the original happening order is defined as its corresponding label-pair. The label-pair of a specific emotion-pair represent the current emotion atmosphere. On the basis of this atmosphere, not all emotion categories share the same probability of occurrence that some of the emotions are actually on the opposite side of each other. For example, with a label-pair as happy-excited-happy, it is impossible for the next target become sad or angry. On the contrary, happy and excited emotion enjoy very high probabilities and also neutral can happen in some cases. Such common regulations are summarized as the domain knowledge under different sample lengths to revise the recognition results of the next RL module at the final step. The widely accepted public dataset is applied as the hotbed to sum up the domain knowledge, because the emotion conversion rules are universal at most circumstances. The generalized domain knowledge will be transformed to apply on other dataset. The emotion categories are divided into six kinds: happy, sad, neutral, angry, excited and frustrated, each of which has different occurrence probability after giving the former label-pair.

Each probability of six emotions after all kinds of label-pair is computed by counting the number of occurrences and normalize the probabilities by a $softmax$ function. The domain knowledge with specific sample size is built. Each of the probability represents the correlation between an emotion and the corresponding label-pair. Let L represents the specific label-pair and e denotes one of the six emotion categories. $Num(e|L)$ represents the number of occurrences of the emotion e with the former label-pair as L. $Num(L)$ represents the number of occurrence of label-pair L in the trainset of the dataset. $P(e|L)$ represents the probability of occurrence of the emotion e under former label-pair as L. The $C(e|L)$ is the final correlation between emotion e and label-pair L which is recorded as domain knowledge. The calculation formulas are as below.

$$P(e|L) = \frac{Num(e|L)}{Num(L)} \tag{1}$$

$$C(e|L) = softmax(P(e|L)) \tag{2}$$

The size of sample window will influence the revise performance of this domain knowledge. Short label-pair will not be able to support enough domain knowledge for the target and conversely, label-pair with a too long size will greatly restrict the flexibility and generalization ability of the domain knowledge. Long label-pair will bring unnecessary constraints to the recognition results with less fault tolerance. When the RL module misidentified the emotion type of an utterance, the exclusivity between classes in emotions becomes greater that domain knowledge with too long label-pair will not only fail to revise, but will make its own claim and produce completely wrong results to all following recognition steps of the same dialogue. In general, the domain knowledge should exclusive the impossible emotion results and give minor corrective support instead of becoming the main factor affecting the final results. The domain knowledge of the four different sample sizes will be presented in following section with other experimental results.

B. The reinforcement learning module

Emotion in conversations is inter-related that happens one after another step by step. This is similar to the sequence state transitions in RL and the action chosen according to the current emotion-state. The reward function is influenced by both of the current emotion-state and the chosen action, which represents the context in ERC and the recognition results of the target respectively. This reinforcement learning module is the core module to recognize the target emotion with features of the corresponding emotion-pair. Dueling deep-Q-network (DDQN) is used as the learning algorithm in RL module and separate process approaches are utilized on trainset and testset. The flow chart of the DDQN and the structure of the reinforcement learning module are depicted in Figure 2.In Figure 2, the multi-modalities of the emotion-state $s(t)$ are the input of the $Q(s(t), a(t))$ at time step t. Text $T(t)$, visual $V(t)$ and audio $A(t)$ are separately fused first before cross modal fusion. Bidirectional gated recurrent unit cell($\overleftrightarrow{GRUCell}$), a RNN based network, is utilized to capture the contextual relationships inside of each unimodal. The fused unimodal features, $T_f(t)$,$V_f(t)$ and $A_f(t)$, are cascaded as $F_{usion}(T_f(t), V_f(t), A_f(t))$ for cross modal fusion through a bidirectional multi-layer gated recurrent unit RNN ($\overleftrightarrow{GRU}$). Three dense layers $Dense$ are connected to the output of the $\overleftrightarrow{GRU}$ layer for utilizing the dueling mechanism to optimize the convergence speed.The specific calculation process is as follows.

At time step t, the features of an emotion-pair $E_{pair}(t)$ and the corresponding target $T_{arget}(t)$ are packed in pairs as an new integrated emotion-state $s(t), s(t) = [E_{pair}(t), T_{arget}(t)]$ of the input of DDQN, regardless of the original conversation they belong to and all these kind of states construct the whole RL environment $S, s \in S$. During training, the Q net of DDQN is trained to output the right probabilities $q_{eval}(s(t), a(t))$ of chosen action from the six alternative emotions, which is represented as $a(t) \in A, A_{ction} = [0, 1, 2, 3, 4, 5]$. The numbers in A_{ction} represent the happy, sad, neutral, angry, excited and frustrated respectively. The recognition result of the target utterance q_{action} is achieved as below.

$$q_{eval}(s(t), a(t)) = Q(s(t), a(t)) \tag{3}$$

$$q_{action} = argmax(Log\_Softmax(q_{eval}(s(t), a(t)))) \tag{4}$$

where $Q(s(t), a(t))$ is the Q net of DDQN.

The detailed calculation formula of $Q(s(t), a(t))$ are as below.

$$T_f(t) = \overleftrightarrow{GRUCell}(T(t)) \tag{5}$$

$$V_f(t) = \overleftrightarrow{GRUCell}(V(t)) \tag{6}$$

$$A_f(t) = \overleftrightarrow{GRUCell}(A(t)) \tag{7}$$

$$F_{usion}(t) = \overleftrightarrow{GRU}(T_f(t), V_f(t), A_f(t)) \tag{8}$$

$$V, A = Dense(F_{usion}(t)) \tag{9}$$

$$q_{eval}(s(t), a(t)) = Dueling(V + (A - \frac{1}{|A|}\sum_a A)) \tag{10}$$

where V and A represent the original $q_{eval}(s(t), a(t))$ from the $Q(s(t), a(t))$ and the average of advantages of each action on $s(t)$. $Dueling$ represents the update of the original $q_{eval}(s(t), a(t))$ by dueling mechanism. The reward function R is computed as below. r is the value of reward and the $label(t)$ means the right emotion label of the $T_{arget}(t)$.

$$R = \begin{cases} q_{action} = label(t), R = r \\ else, R = -r \end{cases} \tag{11}$$

The features of the next state $s(t+1)$ at next time step of current time t is input to the target Q net, which is represented as Q' to compute the loss function $Loss(t)$ as below.

$$q_{eval}(s(t+1), a(t+1)) = Q'(s(t+1), a(t+1)) \tag{12}$$

$$q_{expect}(s(t{+}1), a(t{+}1)) = R{+}\gamma \max_{a(t+1)} q_{eval}(s(t{+}1), a(t{+}1)) \tag{13}$$

$$Loss(t){=}\mathbb{E}[q_{expect}(s(t{+}1), a(t{+}1)){-}q_{eval}(s(t{+}1), a(t{+}1))] \tag{14}$$

The $Loss(t)$ is back propagated to optimize the Q net and the target Q net will updated using the parameters of the Q net after fixed step.

C. The domain knowledge revise module

The domain knowledge revise module is applied during the test step. Unlike the random drawing states from the trainset with replacement strategy, the test is undertaken on the dialogue level. The emotion-states from the same dialogue are tested state by state to examine the recognition performance. The initialization state of the RL environment is one of the very beginning emotion-pair of a dialogue randomly sampled from the video clips of testset. The corresponding emotion label-pair of this emotion-pair will be given as the clue of the original emotion atmosphere. The third revise module use this clue to revise the recognized output from the RL module which recognizes the emotion states of the conversation step by step by just using the multimodal features until reaching the final signal and begins recognizing the other dialogue by same steps. The recognition results of a same dialogue on each step are recorded. The three recognized emotion types before the target utterances are used as the index to search for the corresponding $C(e|L)$ in domain knowledge. According to Equation (1) and Equation (2), $C(e|L)$ contains six probability of occurrence of the six alternative emotion types based on the former three recognized emotions which is in the same format of the RL module output. In order not to lose generality, we just simply add these two sets of probability values together to accomplish the revise instead of training extra layers on specific dataset for better improvement.

IV. Experimental Setup

In this section, the experimental setups and steps of our method are showed, including the datasets, baselines, metrics, training task and the process design of experiments.

A. Datasets

Our model is evaluated on two widely accepted public datasets, IEMOCAP [55] and MELD [56]. These two datasets are multimodal datasets recording textual, visual and audio information of each video conversation on utterance level. Both of two datasets have been randomly partitioned into training set and testing set when they are released and share no same speakers.

IEMOCAP dataset contains conversational videos between two interlocutors that ten unique speakers are included. Only the first eight speakers belong to the trainset. Each video contains a single dyadic dialogue and is labeled on utterance level with one of six emotion labels, which are happy, sad, neutral, angry, excited, and frustrated. Text, audio and visual information of each utterance are included.

MELD is a multimodal emotion and sentiment classification dataset which is annotated on utterance level as one of the seven emotion classes: anger, disgust, sadness, joy, surprise, fear and neutral. MELD is a multi-party conversational videos dataset that contains text, audio and visual modalities for more than 1400 dialogues and 13000 utterances from the Friends TV series.

In the pre-process module, these original utterance-level items are packed into the form of emotion-states, which include the current emotion-pair and target and the emotion-pair and target of next time step for DDQN to compute the $q_{t+1}\_expect(s(t+1), a(t+1))$. To prevent overfitting, in trainset, either current emotion-pair and target or the next step emotion-pair and target are unique that will not exits in any other emotion-state twice. In testset, the $q_{t+1}\_expect(s(t+1), a(t+1))$ for DDQN is not needed that the emotion-states just include the current emotion-pair and target. The test are carried out step by step on dialogue level. As a result, the generalized emotion-states for trainset and testset are less than the number of items in original datasets. The proportion between train and test is about $3:1$. The Table I shows the distributional statistics of the datasets.

TABLE I
Datasets split.

datasets	dialogues		emotion-states	
	train	test	train	test
IEMOCAP	120	31	2758	1530
MELD	1153	280	5333	1770

B. Baselines

The experimental results are discussed with the following recent baselines.

c-LSTM+Att [57]: To represent the efficient context information for the target utterance, utterances around the target are input to attention based bidiredectional LSTM at each time spot.

TFN [58]: This multimodal method fuses information of both each unimodal and cross modals only from present target object.

MFN [59]: Multi-view learning is utilized in this model that self-view and cross-view of multimodalities are fused. This method also only use information only from present target object.

CMN [44]: This method samples the sentences from two interlocutors before the current target as its history. The sample size is set to be eight and two separate memory networks are utilized.

ICON [60]: ICON is similar to CMN that separate memory networks are used for speakers. The ICON adds a extra memory network between interlocutors besides self-speakers and the sample size is set to be the overall dialogues.

BiDialogueRNN+Att [17]: BiDialogueRNN+Att uses three GRUs to capture the target emotion, the context information and the changes of the overall dialogue. This method also needs the whole conversation features that has no capable of real-time performance.

C. Evaluation metrics

To evaluate the experimental performances, two classical parameters, the accuracy and macro-average F-score are calculated. F-score is computed as below.

$$F_\beta = (1+\beta^2) \cdot \frac{precision \cdot recall}{(\beta^2 \cdot precision) + recall} \tag{15}$$

where β represents the weight between precision and recall. In this paper, β is set to be 1 which means precision and recall are regarded to share same weight. To evaluate the significance of our experimental results, paired T-test is conducted to calculate the P-value between our method and baselines on two datasets. The paired T-test is computed as below.

$$t = \frac{\overline{d} - d_0}{S_d/\sqrt{n}} \tag{16}$$

$$df = n - 1 \tag{17}$$

where $\overline{d}$ represents the sample mean of differences, d_0 represents the hypothesized population mean difference, S_d is the standard deviation of differences, df is the degree of freedom. P-value is calculated with the significance level α set to be 0.05.

D. Experimental parameter setting

The dimension of pre-processed text, audio and visual are 100, 100 and 512 respectively. The number of layer for both $\overleftrightarrow{GRUCell}$ and $\overleftrightarrow{GRU}$ are set to be one with the number of hidden layer as 512 and drop probably as 0.3. The γ parameter for DDQN is 0.9 and the updated frequency of the target Q net is 100 with the learning rate as 0.00015 and the weight decay of the optimizer is set to be 0.00001.

V. Results and Discussion

A. Results

Firstly, the domain knowledge from the first module with four different sample window sizes from two utterances to five utterances are presented as form of heat map as Figure 3, Figure 4, Figure 5 , Figure6 and Figure 7. Figure 3, Figure 4 and Figure 5 represented the domain knowledge of the sample window size as two, three and four respectively. The domain knowledge of window size as five are split into two figures to present which are Figure6 and Figure 7. The abscissa of the heat map is all the possible candidate emotion types. The ordinate of the heat map represents the current recognized emotion-pair. The virous heat colors means the different probability of occurrence of the coming candidate emotion types after the different known emotion-pair. The lighter color represents the higher probability of occurrence of the target emotion based on the current emotion-pair and vice versa. In Figure 3, which the size of sample window is 2, the base emotion-pair cases are too little to give enough assistances for revise. On the contrary, there are too many probabilities of occurrence as near to 1.0 that the instructions from domain knowledge are too strong to have enough tolerance and generalize ability.

Secondly, to test the performances of different sizes of sample window, experiments on IEMOCAP and MELD are carried out with total dialogues in testsets and the results are listed in Table II compared with the other results of baselines.

From the experimental results from Table II, ERLDK with sample window size of three and four achieves best performances. However, ERLDK with sample window size of four has a much lower generalization ability than sample window size as three. As the result, the size of three is chosen as the sample window for the following testing.

To examine the real-time capabilities of our model step by step during recognizing the emotion of every utterance in dialogues with the original order, experiments on dialogue-level are conducted. The specific F1-scores of every step of the dialogues in test set are listed in Table III. In Table III, our ERLDK method has the real-time recognition ability, so the F1-scores are changing as the length of the dialogues increase. On the contrary, all the baselines has no real-time ability, so the F1-scores are all average results that summerized through the overall test set. The experimental results with sample window size as three on IEMOCAP are presented in Figure 8. In Figure 8, results of ERLDK and all baselines are depicted. The curves of ERLDK are varied as the steps of the dialogues and the curves of the baselines are straight dottted lines that keep in same on whole dialogue level.

B. Discussion

In ERC model, the span of contexts sampled for a specific utterance have a great influence on the final recognition ability. This is also the reason that many mainstream methods explore to improve their accuracies by capturing all related information from the total dialogue. However, this is not practical in many application scenarios without the real-time recognition ability. In our model, to find the best sample span, four sample window size are chosen to integrate the needed contexts. The emotions of these contexts provide reference information for the next appearing target which are then summed up as the domain knowledge. The domain knowledge under different sample lengths is captured which are presented in Figure 3 to Figure 7 by showing the correlation between emotion-pairs and the next six alternative emotion types through heat maps. Compared to the few emotion-pairs in Figure 3, the other four figures obviously make a more detailed distinction of emotional clues for the potential follow-ups. To find the best window size and the corresponding domain knowledge, we conduct experiments with four kinds of window sizes and domain knowledge on two datasets. The experimental results are listed in Table II. In Table II, our method with four sample window sizes are compared to other baselines on two datasets by using the recognition results on total length of dialogues. Experimental statistics on each classification of IEMOCAP are listed in detailed for sufficient public data for reference. From this table, the accuracy and F1-score of four sample sizes which are showed in the last four rows surpass most baselines. Particularly, our ERLDK with sample size as three achieves the best F1-score performance on happy, excited, frustrated and total average recognitions. ERLDK with sample size as three achieves the best F1-score performance on sad recognition. Especially, ERLDK model surpasses the sate-of-the-art accuracy and F1-score for about 10% on happy emotion which is a big margin. Although the best results are not obtained in all six categories, the recognition efficiency is more balanced than the baselines instead of being completely biased to a certain emotion type.

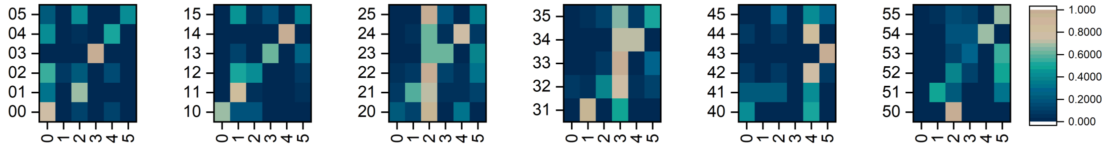


Fig. 3. The correlation between emotion and emtion-pire with sample window as two.

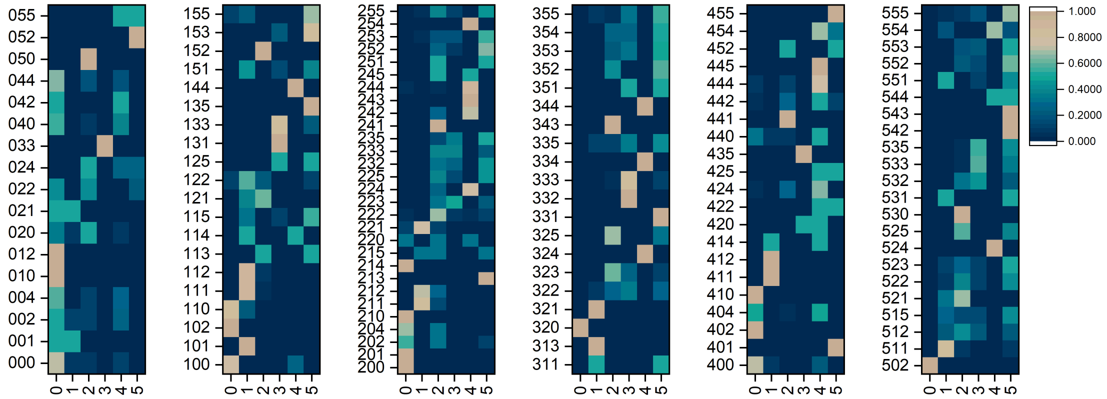


Fig. 4. The correlation between emotion and emtion-pire with sample window as three.

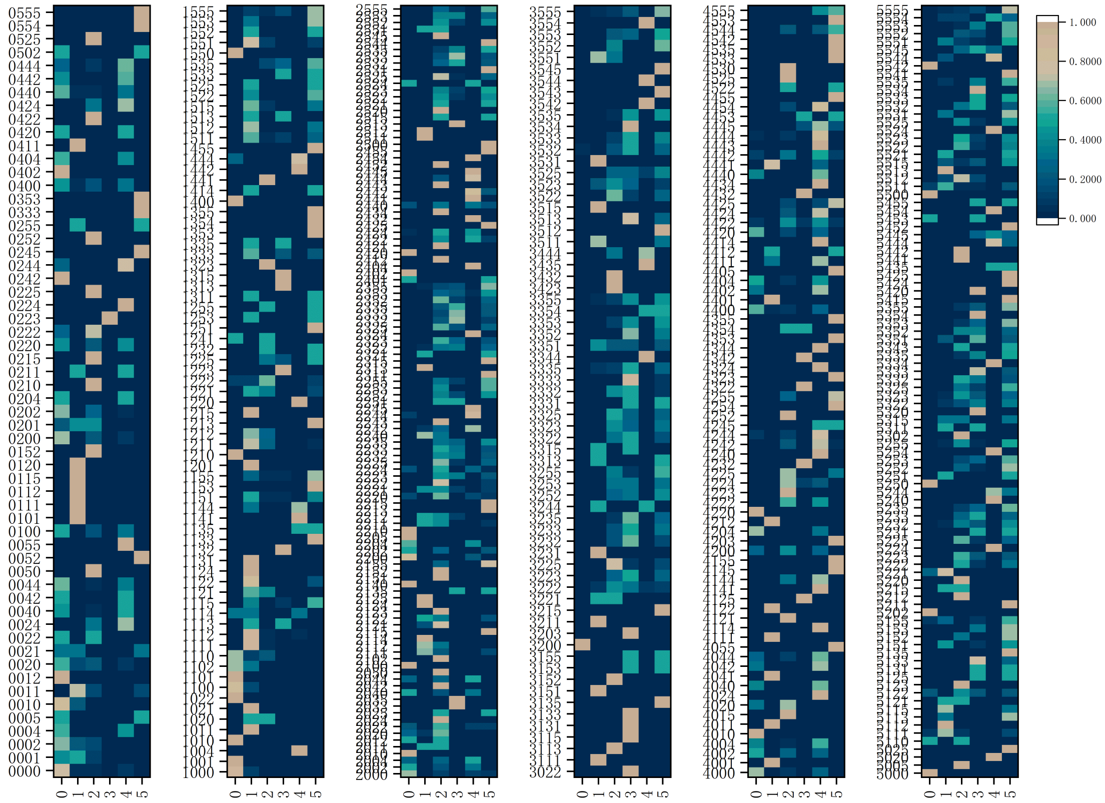


Fig. 5. The correlation between emotion and emtion-pire with sample window as four.

As for the difference performances between the four sampling sizes, ERLDK with sample size as three and four perform better that the other two sizes on almost all metrics. Longer sample size does not receive positive effects not only on the dataset it extracted from but also worse than the smallest sample size of the contexts. It shows that it is not more instructions suggesting better performances. This is because the lack of the fault tolerance ability that leads to an imbalance

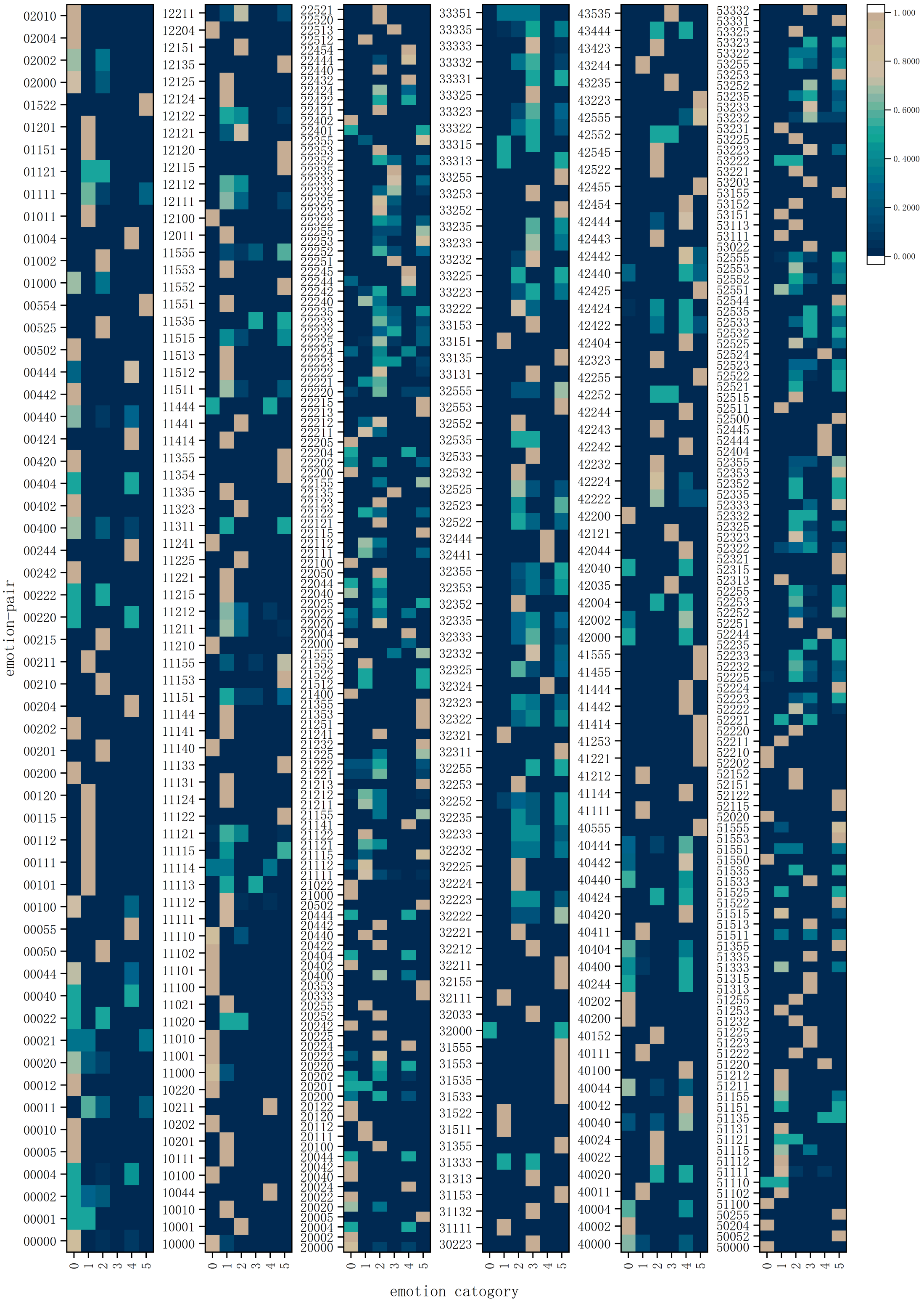


Fig. 6. The correlation between emotion and emtion-pire with sample window as five.

between the recognition ability of RL module and the revise ability of the domain knowledge. This imbalance brings about poor generalization ability that result in the lowest accuracy and F1-score while transforming to another dataset.

To test the real-time performance of ERLDK, we conduct dialogue level experiments on IEMOCAP dataset and the results are showed in Figure 8 and Table III. Figure 8 shows the recognition performance changes with the conversation progresses and the Table III lists all ground truth of each recognition steps. In Figure 8, the length of steps in X axis

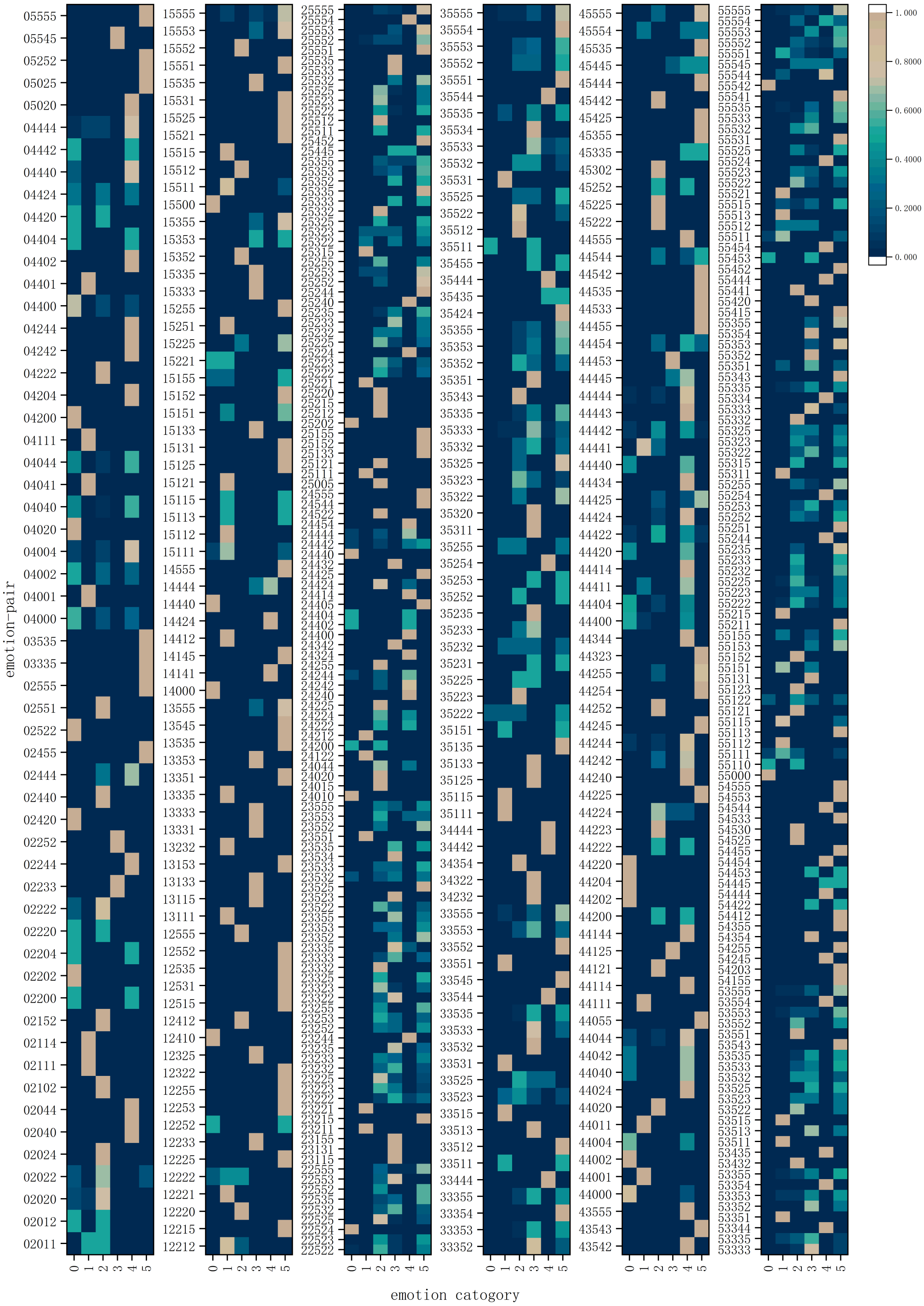


Fig. 7. The correlation between emotion and emtion-pire with sample window as five.

mean the number of utterances which accept recognition in occurring order after the very first given emotion-pair of each dialogue. The accuracy and F1-score in Y axis mean the test results of each length of step. Figure 8(a) and 8(b) are weighted average accuracy and F1-score on all emotion classifications. The other subfigures are F1-score on six emotion classification respectively. The black thick solid curves are results of ERLDK and the green thick solid curves are the results of only RL module along with the steps of the dialogues without the domain knowledge revises for ablation study.

TABLE II

COMPARED WITH THE BASELINES METHODS ON IEMOCAP AND MELD DATASETS WITH TOTAL DIALOGUES IN TESTSETS. ACC. = ACCURACY; BOLD FONT DENOTES THE BEST PERFORMANCES. P-VALUE IS LESS THAN 0.05 FOR PAIR T-TEST. ERLDK($size$) REPRESENTS ERLDK MODELS USING FOUR DIFFERENT SAMPLE WINDOW SIZES. $size \in [2, 3, 4, 5]$

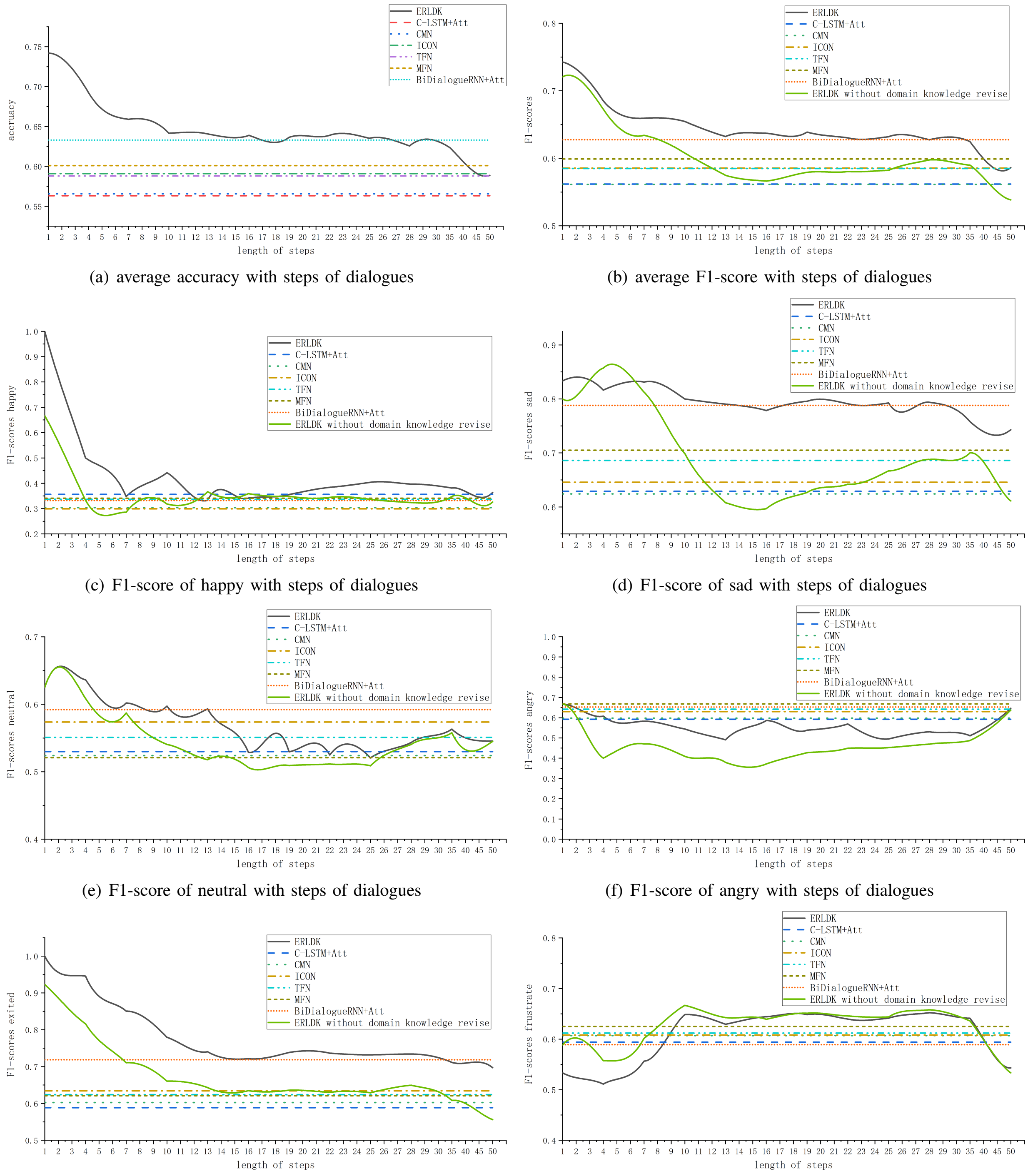
Methods	IEMOCAP														MELD
	Happy		Sad		Neutral		Angry		Excited		Frustrated		Average		Average
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	F1
C-LSTM+Att	30.56	35.63	56.73	62.90	57.55	53.00	59.41	59.24	52.84	58.85	65.88	59.41	56.32	56.19	56.70
CMN	20.1	28.0	62.9	68.3	56.0	57.4	58.8	60.4	68.2	66.7	74.3	63.2	60.7	59.8	-
ICON	22.22	29.91	58.78	64.57	62.76	57.38	64.71	63.04	58.86	63.42	67.19	60.81	59.09	58.54	-
TFN	23.2	33.7	58.0	68.6	56.6	55.1	69.1	64.2	63.1	62.4	65.5	61.2	58.8	58.5	-
MFN	24.0	34.1	65.6	70.5	55.5	52.1	72.3	**66.8**	64.3	62.1	67.9	62.5	60.1	59.9	-
BiDialogueRNN+Att	25.69	33.18	75.10	78.80	58.59	**59.21**	64.71	65.28	80.27	71.86	61.15	61.73	63.40	62.75	57.03
ERLDK ($size$=2)	39.23	42.50	78.03	75.65	45.71	50.39	56.12	61.54	67.06	64.65	61.94	59.36	59.89	60.67	56.80
ERLDK ($size$=3)	44.87	**47.30**	77.67	79.19	57.85	56.42	59.34	60.54	77.95	**74.44**	67.36	63.85	62.19	**63.90**	**59.72**
ERLDK ($size$=4)	45.69	42.61	78.32	**79.95**	55.36	55.94	62.45	63.57	75.84	71.28	65.40	**64.31**	60.72	62.27	55.34
ERLDK ($size$=5)	40.12	45.83	75.67	76.28	58.38	57.62	61.28	60.37	76.86	73.59	65.13	62.01	59.98	60.54	54.44

TABLE III

THE SPECIFIC F1-SCORES OF EVERY STEP OF THE DIALOGUES IN TEST SET. BOLD FONT DENOTES THE BEST PERFORMANCES. P-VALUE IS LESS THAN 0.05 FOR PAIR T-TEST.

Step	Emotion														
	Happy		Sad		Neutral		Angry		Excited		Frustrated		Average		
	ERLDK	SOTA	ERLDK	SOTA	ERLDK	SOTA	ERLDK	SOTA	ERLDK	SOTA	ERLDK	SOTA	ERLDK	ERLDK(no domain)	SOTA
1	1.00	35.63	83.33	78.80	62.50	59.21	66.67	66.8	1.00	71.86	53.33	61.73	74.27	72.06	62.75
2	80.00	35.63	84.62	78.80	68.57	59.21	66.671	66.8	92.86	71.86	51.85	61.73	73.72	73.17	62.75
3	66.67	35.63	84.21	78.80	64.00	59.21	59.38	66.8	95.24	71.86	52.38	61.73	71.61	70.21	62.75
4	50.00	35.63	81.63	78.80	63.64	59.21	60.77	66.8	94.55	71.86	51.12	61.73	68.49	67.27	62.75
5	46.15	35.63	82.76	78.80	60.76	59.21	55.11	66.8	86.15	71.86	52.63	61.73	66.15	64.23	62.75
6	47.06	35.63	83.58	78.80	58.06	59.21	58.00	66.8	88.31	71.86	51.61	61.73	65.82	62.84	62.75
7	34.78	35.63	83.12	78.80	60.19	59.21	58.28	66.8	85.06	71.86	55.65	61.73	65.93	63.40	62.75
8	54.55	35.63	83.72	78.80	59.19	59.21	53.33	66.8	83.67	71.86	54.05	61.73	66.05	62.99	62.75
9	40.56	35.63	81.72	78.80	57.60	59.21	55.71	66.8	81.82	71.86	62.50	61.73	66.02	62.05	62.75
10	44.15	35.63	80.01	78.80	59.74	57.62	54.44	60.37	77.97	73.59	64.84	62.01	65.45	60.76	60.54
11	41.11	35.63	79.63	78.80	56.96	59.21	52.11	66.8	76.69	71.86	65.10	61.73	64.60	59.59	62.75
12	31.37	35.63	79.31	78.80	58.32	59.21	50.91	66.8	73.33	71.86	63.89	61.73	63.81	58.54	62.75
13	33.33	35.63	79.03	78.80	59.32	59.21	49.13	66.8	74.03	71.86	62.93	61.73	63.23	57.48	62.75
14	43.33	35.63	78.79	78.80	55.72	59.21	56.04	66.8	71.86	71.86	63.49	61.73	64.07	56.86	62.75
15	32.10	35.63	78.57	78.80	57.46	59.21	54.62	66.8	71.86	71.86	64.26	61.73	63.73	56.76	62.75
16	34.15	35.63	77.85	78.80	52.91	59.21	58.60	66.8	72.13	71.86	64.43	61.73	63.71	56.63	62.75
17	34.94	35.63	78.75	78.80	52.17	59.21	58.71	66.8	71.79	71.86	64.60	61.73	63.55	56.83	62.75
18	34.09	35.63	79.29	78.80	59.50	59.21	51.18	66.8	72.91	71.86	65.51	61.73	62.64	57.63	62.75
19	35.56	35.63	79.56	78.80	52.96	59.21	53.84	66.8	73.83	71.86	64.85	61.73	63.88	57.91	62.75
20	36.56	35.63	80.41	78.80	53.23	57.62	54.44	60.37	74.44	73.59	65.28	62.01	63.33	58.19	60.54
21	37.89	35.63	79.62	78.80	55.99	59.21	54.74	66.8	74.46	71.86	64.38	61.73	63.30	57.76	62.75
22	38.38	35.63	79.11	78.80	52.48	59.21	56.91	66.8	73.64	71.86	63.96	61.73	63.03	58.04	62.75
23	38.83	35.63	78.48	78.80	55.17	59.21	52.73	66.8	73.39	71.86	63.41	61.73	62.52	57.93	62.75
24	39.62	35.63	78.88	78.80	54.15	59.21	48.94	66.8	73.15	71.86	63.87	61.73	62.95	58.14	62.75
25	40.37	35.63	79.25	78.80	52.06	59.21	49.48	66.8	73.21	71.86	64.14	61.73	63.19	58.22	62.75
26	41.07	35.63	75.22	78.80	53.43	59.21	51.16	66.8	73.19	71.86	64.98	61.73	63.93	59.08	62.75
27	40.02	35.63	79.72	78.80	53.33	59.21	52.55	66.8	73.43	71.86	64.86	61.73	63.24	59.31	62.75
28	39.66	35.63	79.37	78.80	54.34	59.21	53.06	66.8	73.40	71.86	65.22	61.73	62.74	59.76	62.75
29	39.67	35.63	79.04	78.80	55.71	59.21	52.00	66.8	73.38	71.86	65.11	61.73	63.20	59.93	62.75
30	39.02	35.63	78.73	78.80	54.93	57.62	53.94	60.8	72.73	73.59	64.17	62.01	63.45	59.16	60.54
35	38.10	35.63	75.80	78.80	56.32	59.21	51.09	66.8	71.14	71.86	64.12	61.73	62.46	58.98	62.75
40	38.89	35.63	73.53	78.80	54.44	59.21	54.22	66.8	70.03	71.86	60.66	61.73	59.79	57.57	62.75
45	30.77	35.63	72.26	78.80	54.56	59.21	58.14	66.8	72.76	71.86	54.04	61.73	57.03	54.29	62.75
50	36.36	35.63	74.29	78.80	54.56	59.21	64.67	66.8	69.71	71.86	54.31	61.73	58.61	53.84	62.75

It can be seen clearly that the results of only RL module perform less than the complete ERLDK from 2.21% to 7.08% which shows the effectiveness of the domain knowledge in improving the recognizing performance. The lines in other colors are the results of the baselines. The baselines have no real-time ability to recognize the emotion of the utterances as their appearances, so the results of baselines keep unchanged on whole dialogue level which is the biggest difference between ERLDK and other baselines. To show the performances in more detail, we list all specific F1-score values of each steps on average and six emotion types in Table III for reference. In Table III, SOTA means the state-of-the-art results which has already showed in Table II on each emotion types and the average. From these figures, most of them receive the highest recognition F1-score in the first step except the frustrated emotion because of the revise from the domain knowledge. This advantage becomes less conspicuous during around the fifteenth step and enters a plateau. ERLDK surpasses the baselines on most emotion types and the average from 1.21% to 6.64% at the plateau and surpasses the baselines more on the first fifteen steps. The frustrated emotion in Figure 8(h) is still the opposite exception that the recognition performance is not good enough until the plateau. We analysis this exception that frustrated emotion is easy to be confused by both recognition algorithms and the domain knowledge which lead to this negative influences. After the step of around thirty to thirty-five, the recognition performance has a significant decline.



(a) average accuracy with steps of dialogues

(b) average F1-score with steps of dialogues

(c) F1-score of happy with steps of dialogues

(d) F1-score of sad with steps of dialogues

(e) F1-score of neutral with steps of dialogues

(f) F1-score of angry with steps of dialogues

(g) F1-score of excited with steps of dialogues

(h) F1-score of frustrated with steps of dialogues

Fig. 8. The comparation of accuracy and F1-score with baselines.

This is because most dialogues in testsets do not have this long conversation length. Only four dialogues contain more than fifty length of steps for testing. This makes the uneven distribution of emotion types that influence the results. From the ablation study which shows in Figure 8 and Table III, domain knowledge can greatly improve the performance in the early stages and help our method stabilize at a better value than all the baselines in a relative long-term.

VI. Conclusion and Future Work

In this paper, we proposed a multimodal emotion recognition model based on reinforcement learning and domain knowledge in video conversations. In particular, the reinforcement learning module is utilized to perform real-time recognition of conversation occurrence, while domain knowledge leverages emotion-pairs to revise recognition results. To the best of our knowledge, this is the first time that these two elements are combined for real-time ERC.

We achieved state-of-the-art results on average and most of specific emotion categories. As future work, we plan to continue optimize the RL module and the extraction of domain knowledge to improve recognition ability on all emotion classifications.

REFERENCES

[1] Y. Hu, M. Lu, C. Xie, and X. Lu, "Driver drowsiness recognition via 3d conditional gan and two-level attention bi-lstm," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2019.

[2] B. Peng, J. Lei, H. Fu, C. Zhang, T.-S. Chua, and X. Li, "Unsupervised video action clustering via motion-scene interaction constraint," *IEEE Transactions on Circuits and Systems for Video Technology*, 2018.

[3] X. Sun, C. Zhang, and L. Li, "Dynamic emotion modelling and anomaly detection in conversation based on emotional transition tensor," *Information Fusion*, vol. 46, pp. 11–22, 2019.

[4] Z. Ji, K. Xiong, Y. Pang, and X. Li, "Video summarization with attention-based encoder-decoder networks," *IEEE Transactions on Circuits and Systems for Video Technology*, 2019.

[5] Y. Huang, X. Cao, Q. Wang, B. Zhang, X. Zhen, and X. Li, "Long-short-term features for dynamic scene classification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 4, pp. 1038–1047, 2019.

[6] J. Gao, Q. Wang, and X. Li, "Pcc net: Perspective crowd counting via spatial convolutional network," *IEEE Transactions on Circuits and Systems for Video Technology*, 2019.

[7] D. Li, Y. Li, and S. Wang, "Interactive double states emotion cell model for textual dialogue emotion prediction," *Knowledge-Based Systems*, vol. 189, p. 105084, 2020.

[8] Y. Ou, Z. Chen, and F. Wu, "Multimodal local-global attention network for affective video content analysis," *IEEE Transactions on Circuits and Systems for Video Technology*, 2020.

[9] S.-L. Yeh, Y.-S. Lin, and C.-C. Lee, "An interaction-aware attention network for speech emotion recognition in spoken dialogs," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6685–6689.

[10] J. Bhaskar, K. Sruthi, and P. Nedungadi, "Hybrid approach for emotion classification of audio conversation based on text and speech mining," *Procedia Computer Science*, vol. 46, pp. 635–643, 2015.

[11] S. Xing, S. Mai, and H. Hu, "Adapted dynamic memory network for emotion recognition in conversation," *IEEE Transactions on Affective Computing*, pp. 1–1, 2020.

[12] X. Cheng, J. Lu, B. Yuan, and J. Zhou, "Identity-preserving face hallucination via deep reinforcement learning," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2019.

[13] S. Wang, L. Hao, and Q. Ji, "Knowledge-augmented multimodal deep regression bayesian networks for emotion video tagging," *IEEE Transactions on Multimedia*, vol. 22, no. 4, pp. 1084–1097, 2020.

[14] D. Hazarika, S. Poria, R. Zimmermann, and R. Mihalcea, "Conversational transfer learning for emotion recognition," *Information Fusion*, 2020.

[15] X. Li, M. Chen, F. Nie, and Q. Wang, "Locality adaptive discriminant analysis." in *Proceedings of International Joint Conference on Artificial Intelligence(IJCAI)*, 2017, pp. 2201–2207.

[16] D. Peng, M. Zhou, C. Liu, and J. Ai, "Human-machine dialogue modelling with the fusion of word-and sentence-level emotions," *Knowledge-Based Systems*, vol. 192, p. 105319, 2020.

[17] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, and E. Cambria, "Dialoguernn: An attentive rnn for emotion detection in conversations," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 6818–6825.

[18] D. Ghosal, N. Majumder, S. Poria, N. Chhaya, and A. Gelbukh, "Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation," *arXiv preprint arXiv:1908.11540*, 2019.

[19] F. Alam, M. Danieli, and G. Riccardi, "Annotating and modeling empathy in spoken conversations," *Computer Speech & Language*, vol. 50, pp. 40–61, 2018.

[20] X. Xiang and T. D. Tran, "Linear disentangled representation learning for facial actions," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 12, pp. 3539–3544, 2017.

[21] W. Zheng, J. Zhu, and B. Lu, "Identifying stable patterns over time for emotion recognition from eeg," *IEEE Transactions on Affective Computing*, vol. 10, no. 3, pp. 417–429, 2019.

[22] F. Dicé, P. Dolce, A. Maiello, and M. F. Freda, "Exploring emotions in dialog between health provider, parent and child. an observational study in pediatric primary care," *Pratiques Psychologiques*, vol. 26, no. 1, pp. 69–84, 2020.

[23] B. A. Sisk, A. B. Friedrich, J. DuBois, and J. W. Mack, "Emotional communication in advanced pediatric cancer conversations," *Journal of pain and symptom management*, vol. 59, no. 4, pp. 808–817, 2020.

[24] K.-Y. Huang, C.-H. Wu, Q.-B. Hong, M.-H. Su, and Y.-H. Chen, "Speech emotion recognition using deep neural network considering verbal and nonverbal speech sounds," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5866–5870.

[25] L. Stappen, A. Baird, E. Cambria, and B. W.Schuller, "Sentiment analysis and topic recognition in video transcriptions," *IEEE Intelligent Systems*, vol. 36, no. 2, 2021.

[26] C. Zhao, X. Li, and Y. Dong, "Learning blur invariant binary descriptor for face recognition," *Neurocomputing*, vol. 404, pp. 34 – 40, 2020.

[27] S. Zhang, R. Ji, J. Hu, X. Lu, and X. Li, "Face sketch synthesis by multidomain adversarial learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 5, pp. 1419–1428, 2019.

[28] D. Tao, Y. Guo, Y. Li, and X. Gao, "Tensor rank preserving discriminant analysis for facial recognition," *IEEE Transactions on Image Processing*, vol. 27, no. 1, pp. 325–334, 2018.

[29] J. Hofmann, T. Platt, and W. Ruch, "Laughter and smiling in 16 positive emotions," *IEEE Transactions on Affective Computing*, vol. 8, no. 4, pp. 495–507, 2017.

[30] B. Ryu, A. R. Rivera, J. Kim, and O. Chae, "Local directional ternary pattern for facial expression recognition," *IEEE Transactions on Image Processing*, vol. 26, no. 12, pp. 6006–6018, 2017.

[31] D. Tao, Y. Guo, B. Yu, J. Pang, and Z. Yu, "Deep multi-view feature learning for person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 10, pp. 2657–2666, 2018.

[32] D. Gong, Z. Li, W. Huang, X. Li, and D. Tao, "Heterogeneous face recognition: A common encoding feature discriminant approach," *IEEE Transactions on Image Processing*, vol. 26, no. 5, pp. 2079–2089, 2017.

[33] X. Li, M. Chen, F. Nie, and Q. Wang, "A multiview-based parameter free framework for group detection," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, vol. 31, no. 1, 2017, pp. 4147–4153.

[34] V. Gupta, M. D. Chopda, and R. B. Pachori, "Cross-subject emotion recognition using flexible analytic wavelet transform from eeg signals," *IEEE Sensors Journal*, vol. 19, no. 6, pp. 2266–2274, 2019.

[35] M. Hu, Y. Yang, F. Shen, L. Zhang, H. T. Shen, and X. Li, "Robust web image annotation via exploring multi-facet and structural knowledge," *IEEE Transactions on Image Processing*, vol. 26, no. 10, pp. 4871–4884, 2017.

[36] S. Zhang, S. Zhang, T. Huang, W. Gao, and Q. Tian, "Learning affective features with a hybrid deep model for audio–visual emotion recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 10, pp. 3030–3043, 2017.

[37] P. Singh, N. Pisipati, P. R. Krishna, and M. V. Prasad, "Social signal processing for evaluating conversations using emotion analysis and sentiment detection," in *2019 Second International Conference on Advanced Computational and Communication Paradigms (ICACCP)*. IEEE, 2019, pp. 1–5.

[38] K.-Y. Huang, C.-H. Wu, Q.-B. Hong, M.-H. Su, and Y.-R. Zeng, "Speech emotion recognition using convolutional neural network with audio word-based embedding," in *2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. IEEE, 2018, pp. 265–269.

[39] S.-L. Yeh, Y.-S. Lin, and C.-C. Lee, "A dialogical emotion decoder for speech motion recognition in spoken dialog," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6479–6483.

[40] Y. Zhang, D. Song, X. Li, P. Zhang, P. Wang, L. Rong, G. Yu, and B. Wang, "A quantum-like multimodal network framework for modeling interaction dynamics in multiparty conversational sentiment analysis," *Information Fusion*, 2020.

[41] X. Huang, W. Wu, H. Qiao, and Y. Ji, "Brain-inspired motion learning in recurrent neural network with emotion modulation," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 10, no. 4, pp. 1153–1164, 2018.

[42] F. M. Plaza-del Arco, M. T. Martín-Valdivia, L. A. Ureña-López, and R. Mitkov, "Improved emotion recognition in spanish social media through incorporation of lexical knowledge," *Future Generation Computer Systems*, vol. 110, pp. 1000–1008, 2020.

[43] S. Wang, L. Hao, and Q. Ji, "Knowledge-augmented multimodal deep regression bayesian networks for emotion video tagging," *IEEE Transactions on Multimedia*, vol. 22, no. 4, pp. 1084–1097, 2019.

[44] D. Hazarika, S. Poria, A. Zadeh, E. Cambria, L.-P. Morency, and R. Zimmermann, "Conversational memory network for emotion recognition in dyadic dialogue videos," in *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*, vol. 2018. NIH Public Access, 2018, p. 2122.

[45] H. Lai, H. Chen, and S. Wu, "Different contextual window sizes based rnns for multimodal emotion detection in interactive conversations," *IEEE Access*, 2020.

[46] I. Kansizoglou, L. Bampis, and A. Gasteratos, "An active learning paradigm for online audio-visual emotion recognition," *IEEE Transactions on Affective Computing*, 2019.

[47] Q. Wang and Y. Hao, "Alstm: An attention-based long short-term memory framework for knowledge base reasoning," *Neurocomputing*, 2020.

[48] W. Liu, L. Zhang, D. Tao, and J. Cheng, "Reinforcement online learning for emotion prediction by using physiological signals," *Pattern Recognition Letters*, vol. 107, pp. 123–130, 2018.

[49] H. Li and H. Xu, "Deep reinforcement learning for robust emotional classification in facial expression recognition," *Knowledge-Based Systems*, p. 106172, 2020.

[50] X. Huang, W. Wu, and H. Qiao, "Connecting model-based and model-free control with emotion modulation in learning systems," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2019.

[51] Q. Wang, Y. Hao, and J. Cao, "Adrl: An attention-based deep reinforcement learning framework for knowledge graph reasoning," *Knowledge-Based Systems*, p. 105910, 2020.

[52] L. Chen, M. Wu, M. Zhou, J. She, F. Dong, and K. Hirota, "Information-driven multirobot behavior adaptation to emotional intention in human–robot interaction," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 10, no. 3, pp. 647–658, 2018.

[53] L. M. Hunnikin and S. H. van Goozen, "How can we use knowledge about the neurobiology of emotion recognition in practice?" *Journal of Criminal Justice*, vol. 65, 2019.

[54] X. Sun, J. Li, X. Wei, C. Li, and J. Tao, "Emotional editing constraint conversation content generation based on reinforcement learning," *Information Fusion*, vol. 56, pp. 70–80, 2020.

[55] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, p. 335, 2008.

[56] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "Meld: A multimodal multi-party dataset for emotion recognition in conversations," *In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 527–536, 2019.

[57] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, and L.-P. Morency, "Context-dependent sentiment analysis in user-generated videos," in *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)*, 2017, pp. 873–883.

[58] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 1103–1114, 2017.

[59] A. Zadeh, P. P. Liang, N. Mazumder, S. Poria, E. Cambria, and L.-P. Morency, "Memory fusion network for multi-view sequential learning," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018, pp. 5634–5641.

[60] D. Hazarika, S. Poria, R. Mihalcea, E. Cambria, and R. Zimmermann, "Icon: Interactive conversational memory network for multimodal emotion detection," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 2594–2604.

Ke Zhang received his Ph.D. degree from Northwestern Polytechnical University, Xi'an, China, in 1998. He is the associate dean of School of Astronautics, and he works as a Full Professor with National Key Laboratory of Aerospace Flight Dynamics in Northwestern Polytechnical University, Xi'an, China.His current research interests include image and signal processing, intelligent perception.

Yuanqing Li received her B.E degree from Northwestern Polytechnical University, Xi'an, China, in 2015. She worked at the 3rd Institute of the China North Industries Group Corporation from 2015 to 2017. She is currently a Ph.D student in School of Astronautics and National Key Laboratory of Aerospace Flight Dynamics, Northwestern Polytechnical University. Her research interests include multimodal emotion recognition and sentiment analysis.

Jingyu Wang (Member, IEEE) received the Ph.D. degree in signal, image and automatic from the Université Paris-Est, Paris, France, in 2015. He is currently an Associate Professor with the School of Astronautics, School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. His research interests include image processing, computer vision and intelligent perception.

Erik Cambria (Fellow, IEEE) is an associate professor at Nanyang Technological University, Singapore. His main research interests are AI and affective computing. He received the Ph.D. degree in computing science and mathematics through a joint programme between the University of Stirling, Stirling, U.K., and MIT Media Lab, Cambridge, MA, USA.

Xuelong Li (Fellow, IEEE) Xuelong Li (M'02-SM'07-F'12) is a full professor with School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an 710072, P.R. China.