Context-aware Cascade Attention-based RNN for Video Emotion Recognition

Man-Chin Sun
Emotibot Inc.
Taipei, Taiwan
manchinsun@emotibot.com

Shih-Huan Hsu
Emotibot Inc.
Taipei, Taiwan
cyrilhsu@emotibot.com

Min-Chun Yang
Emotibot Inc.
Taipei, Taiwan
kaiyang@emotibot.com

Jen-Hsien Chien
Emotibot Inc.
Taipei, Taiwan
kennychien@emotibot.com

***Abstract*—Emotion recognition can provide crucial information about the user in many applications when building human-computer interaction (HCI) systems. Most of current researches on visual emotion recognition are focusing on exploring facial features. However, context information including surrounding environment and human body can also provide extra clues to recognize emotion more accurately. Inspired by "sequence to sequence model" for neural machine translation, which models input and output sequences by an encoder and a decoder in recurrent neural network (RNN) architecture respectively, a novel architecture, "CACA-RNN", is proposed in this work. The proposed network consists of two RNNs in a cascaded architecture to process both context and facial information to perform video emotion classification. Results of the model were submitted to video emotion recognition sub-challenge in Multimodal Emotion Recognition Challenge (MEC2017). CACA-RNN outperforms the MEC2017 baseline (mAP of 21.7%): it achieved mAP of 45.51% on the testing set in the video only challenge.**

***Index Terms*—emotion recognition, video classification, action recognition, spatiotemporal model, human-computer interaction, HCI**

I. Introduction

Understanding human emotion has attracted a lot of attention recently. And it also plays an important role in many applications such as human-computer interaction, advertising, social media communication and cognitive science. However, emotion recognition is still a challenging task. It's very difficult to find an effective model for emotion and facial expressions, let alone combining of multimodal data of visual, vocal and even text to emotion recognition. In this work, a novel model is proposed to consider facial and context information concurrently, which leads to superior performance in emotion recognition.

One of the key problems of emotion recognition is emotion representation or emotion model. There are many researches about emotion representation, such as six discrete basic emotion classes proposed by Ekman [1], continuous dimensional models (e.g. valence and arousal in [2]), and facial action coding system (FACS) in [3], which describes facial movement in action units (AU). Combination of facial action units can be classified into different emotions.

Many emotion datasets and challenges have also been published using aforementioned emotion representation models such as AFF-Wild [4], which uses valence and arousal space to model facial expression. For the emotion classification representation, challenges such as Emotion Recognition Challenge in the Wild (EmotiW) [5] and Multimodal Emotion Recognition Challenge (MEC) [6], which is the challenge of this work attending, have been held for recent years. Furthermore, some challenges such as [5], [6] also includes vocal features for emotion recognition.

In psychological researches [7], [8], it has been discussed that other than facial expressions, contextual information such as body, pose and surrounding environment can also provide important clues for emotion perception. Evidence and experiments are also provided in [7], [8] to show that emotion perception can be influenced by context. Moreover, in some cases, context is even indispensable for emotion communication. Similar results are also proposed in computer vision literatures. Experiments in [9] show that when using both context and body information, performance of emotion recognition outperforms that of using only body image or only context image. A dataset "Emotions in Context Database" (EMOTIC) has also been published recently in [9].

Most of recent emotion recognition methods focus on exploring facial features based on deep neural network. Convolutional Neural Network (CNN) has been used to extract face features in some works [10], [11]. Some researches incorporate 3D Convolutional Networks (C3D) and Recurrent Neural Network (RNN) to model spatial and temporal clues of faces [12], [13]. Also some works [12]–[14] combine audio models to perform multimodal emotion recognition.

This work focuses on video emotion classification using sequence of both facial and context information. A Context-Aware Cascade Attention-based RNN is proposed here to leverage both facial and context features to perform video emotion recognition. To further evaluate the influence of context information, multiple RNN-based networks are designed to compare different methods of fusing face and context features.

II. Related Works

Many works proposed video recognition methods based on neural networks. C3D networks were utilized to learn temporal information from sequential images for action recognition [15], [16] and video emotion recognition [13]. Recurrent neural network, which has temporal recurrence of latent variables,

978-1-5386-5311-1/18/$31.00 ©2018 IEEE

was proposed in [17]. Long-Short Term Memory (LSTM) was proposed in [18] to handle long-range sequence learning.

"Sequence-to-sequence" model [19], which models sequence encoding and generation using LSTM RNNs, is one of the popular architectures of learning from sequential data. It has achieved the state-of-the-art results in neural machine translation. The sequence-to-sequence model consists of a RNN-based encoder and a decoder to learn temporal structures and generate output sequence. Furthermore, temporal attention mechanism has also been proposed in [20] as a soft-alignment method in machine translation to learn relevant temporal segments from encoder and decoder.

The encoder-decoder framework in sequence-to-sequence has also been used to describe videos with CNN and RNN to generate captions [25]. The temporal attention mechanism has also been incorporated in [26] to learn relevant temporal intervals from video sequences.

III. Proposed Method

In this section, the effect of context on improving emotion perception is first shown in Section. III-A. Then, inspired by Sequence-to-sequence model and attention mechanism for neural machine translation, a novel architecture for sequence classification will be proposed. Sequence-to-sequence model and attention mechanism will be briefly described in Section. III-B and III-C respectively. Then the proposed model will be introduced in detail in Section. III-D.

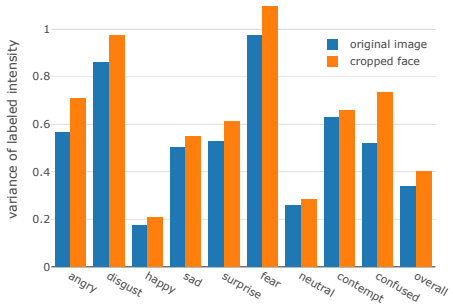


Fig. 1: Visualization of human emotion perception. The dataset described in Section. III-A is labeled **original images** are the label obtained providing the original images; **cropped face** are the labels when only lablers are provided with cropped face.

A. Effect of Context Information on Emotion Perception

To visualize the effect of context information to emotion perception for human, an experiment was designed to show disagreement of human label of the same dataset with and without context information. The dataset consists of 55504 images with bonding box of each face. Each image is labeled by intensity of 9 emotions: "angry", "disgust", "happy", "sad", "surprise", "fear", "neutral", "contempt", "confused" and intensity is in score from 0 to 5 by each labeler. Each image is labeled by at least 5 labelers. Then the disagreement of each emotion is calculated by averaging the variance of labeled intensity of each image.

In Fig. 1, it is shown that when providing the labelers with original images of both face and context, more consensus can be observed in the obtained labels. Furthermore, emotion class "happy" and "neutral" have the least disagreement, which means happy and neutral can be recognized easily, while other emotions of higher disagreement shows more confusion in labelers.

B. Sequence-to-Sequence Framework

Sequence-to-sequence, which is an encoder-decoder framework, was proposed in [19] to perform machine translation. The encoder reads a sequence of input vectors over input time $1 \dots T$ as $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_T)$, into a "context vector" at encoder $\mathbf{c}_{e,T}$. The context vector here represents the stored state and information of the encoder. When the encoder is a RNN, the hidden states are represented as $\mathbf{h}_{e,t} = f(\mathbf{x}_{e,t}, \mathbf{h}_{e,t-1})$ and $\mathbf{c}_{e,T} = g(\mathbf{h}_{e,1}, \mathbf{h}_{e,2}, \dots, \mathbf{h}_{e,T})$, where $f(\cdot)$ and $g(\cdot)$ are nonlinear functions.

The decoder then generates one prediction at output time i, $\mathbf{y}_i$, from context vector $\mathbf{c}_{e,T}$ and all the previous predicted output. With a RNN decoder, its conditional probability models as

$$p(\mathbf{y}_i \,|\, \mathbf{y}_1, \dots, \mathbf{y}_{i-1}, \mathbf{c}_{e,T}) = h(\mathbf{y}_{i-1}, \mathbf{c}_{e,T}, \mathbf{h}_{d,i}), \qquad (1)$$

where $\mathbf{h}_{d,i}$ is the hidden states of the decoder RNN at output time i, $\mathbf{h}_{d,i} = f(\mathbf{h}_{d,i-1}, \mathbf{y}_{i-1}, \mathbf{c}_{e,T})$, and $f(\cdot)$, $h(\cdot)$ are nonlinear functions.

C. Attention Mechanism

Attention mechanism has been proposed to improve the performance of English-to-French machine translation in [20] as a sequence-to-sequence model. The mechanism makes decoder not only depend the context vector from encoder but allow the decoder model to learn relevant parts over source sequence to predict target.

The soft dot global attention in [22] is adapted here. An alignment vector $\mathbf{a}_i = (a_{i,1}, \dots, a_{i,T})$ is a sequence over input time, where $a_{i,t}$ is derived by encoder's hidden states $\mathbf{h}_{e,t}$ at input time t and current decoder's hidden state $\mathbf{h}_{d,i}$ at output time i, which is

$$a_{i,t} = \frac{\exp(\text{score}(\mathbf{h}_{d,i}, \mathbf{h}_{e,t}))}{\sum_{t=1}^{T} \exp(\text{score}(\mathbf{h}_{d,i}, \mathbf{h}_{e,t}))}. \qquad (2)$$

The score function implemented here is dot operation from $\mathbf{h}_{d,i}$ and $\mathbf{h}_{e,t}$, where $\text{score}(\mathbf{h}_{d,i}, \mathbf{h}_{e,t}) = \mathbf{h}_{d,i}^T \mathbf{h}_{e,t}$ and noted that the hidden size of encoder and decoder should be equal. Then the context vector $\mathbf{c}_i$ at output time i is derived as weighted sum of all source $\mathbf{h}_e$ and $a_{i,t}$, which is computed as

$$\mathbf{c}_i = \sum_{t=1}^{T_x} a_{i,t} \cdot \mathbf{h}_{e,t}. \qquad (3)$$

To summarize (2) and (3), the context vector with attention mechanism is a function of hidden states of both encoder and decoder. Based on both decoder and encoder's hidden states, the decoder can pay attention to relevant input sequence of encoder and generate aligned output sequence.

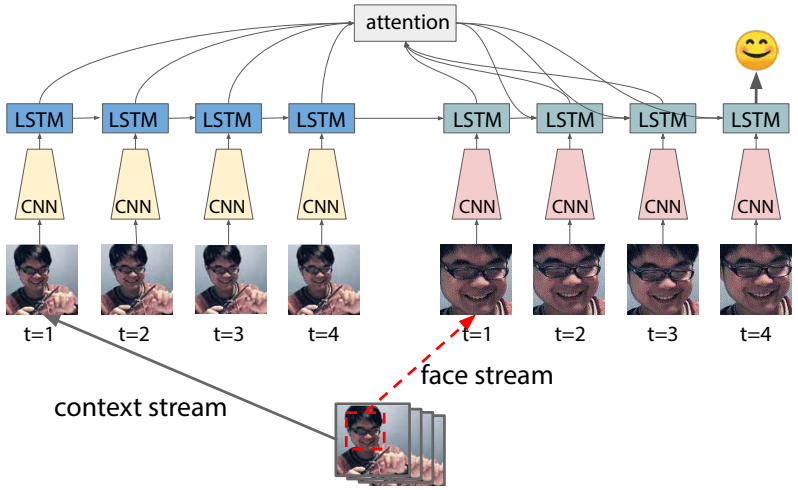


Fig. 2: Context-Aware Cascade Attention-based RNN (CACA-RNN). A video clip is preprocessed into a face stream (sequence of cropped faces) and a context stream (sequence of original frames). Then, the extracted features of the two sequences are fed into two LSTMs: the context RNN learns contextual temporal information and the face RNN learns facial information. The two LSTMs are in a cascaded architecture with attention mechanism. The context RNN stores context clue in LSTM cells and initiates the first state of face RNN at the first time step. The face RNN then learns from face features and processed context information from attention mechanism. Then the output of face RNN of the last frame of clip is considered to be the prediction of the clip.

D. Context-Aware Cascade Attention-based RNN

Context-Aware Cascade Attention-based RNN (CACA-RNN) is proposed to leverage both face and context features to perform emotion recognition. The CACA-RNN consists of two LSTMs in a hierarchical cascade architecture with attention mechanism (Fig. 2).

Unlike encoder-decoder framework, which the decoder is a generative model, there are two encoder-like RNNs in CACA-RNN, "context RNN" and "face RNN", which they read face feature and context feature respectively. The two RNNs are cascaded with attention mechanism in between. The attention mechanism of the CACA-RNN can locate relevant context information from the context RNN when processing the face sequence in the face RNN. After the context RNN reads whole context features, the face RNN encodes face features and its LSTM context vector is derived from attention operation of face and context RNN's hidden states. Then the face RNN outputs the final prediction class after forwarding whole sequence.

The context RNN reads the context feature sequence from time step 1 to T as $\mathbf{X}_{\text{context}} = (\mathbf{x}_{\text{context}_1}, \ldots, \mathbf{x}_{\text{context}_T})$ into a LSTM context vector $\mathbf{c}_{\text{context}_T}$, where its hidden states at time t being denoted as $\mathbf{h}_{\text{context}_t} = f(\mathbf{x}_{\text{context}_t}, \mathbf{h}_{\text{context}_{t-1}})$. Similarily, the face RNN reads face feature sequences $\mathbf{X}_{\text{face}} = (\mathbf{x}_{\text{face}_1}, \ldots, \mathbf{x}_{\text{face}_T})$ and its conditional probability is modeled as

$$p(\mathbf{y}_i \mid \mathbf{x}_{\text{face}_1}, \ldots, \mathbf{x}_{\text{face}_i}, \mathbf{c}_{\text{face}_i}) = h(\mathbf{x}_{\text{face}_i}, \mathbf{c}_{\text{face}_i}, \mathbf{h}_{\text{face}_i}), \quad (4)$$

where $\mathbf{h}_{\text{face}_i}$ is the hidden state of the face RNN, denoted as $\mathbf{h}_{\text{face}_i} = f(\mathbf{h}_{\text{face}_{i-1}}, \mathbf{x}_{\text{face}_i}, \mathbf{c}_{\text{face}_i})$. And the LSTM context vector $\mathbf{c}_{\text{face}_i}$ is derived from Eq. (2) and Eq. (3), which is

$$\mathbf{c}_{\text{face}_i} = \sum_{t=1}^{T} \frac{\exp(\text{score}(\mathbf{h}_{\text{face}_i}, \mathbf{h}_{\text{context}_t}))}{\sum_{t=1}^{T} \exp(\text{score}(\mathbf{h}_{\text{face}_i}, \mathbf{h}_{\text{context}_t}))} \cdot \mathbf{h}_{\text{context}_t}. \quad (5)$$

IV. Comparison Architectures

To investigate performance of RNN-based models fusing face and context features on emotion recognition, various architectures of combining both features or using only one of the features are introduced as the following:

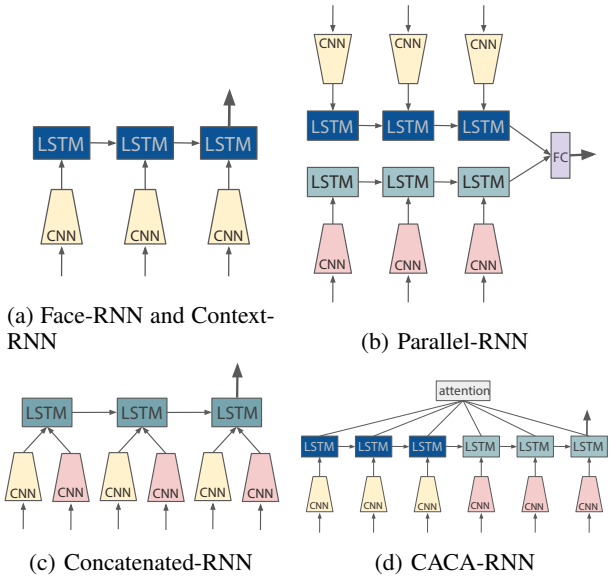
Context-RNN performs emotion classification using video frames as input, depicted in Fig. 3a. The model is similar to a model in [23] for action recognition, where a LSTM processes sequential features from CNN and learns to predict human action. The Context-RNN reads context features which are extracted from a CNN feature extractor, and last prediction is taken as output.

Face-RNN, on the contrary to the Context-RNN, is only using face feature sequence to predict emotion. The architecture is depicted in Fig. 3a.

Parallel-RNN, illustrated in Fig. 3b, consists of two RNNs processing face and context features individually. Then the hidden states of the two RNNs are fused for final prediction at latest time step.

Concatenated-RNN contains a RNN taking concatenated face and context features as input. Then its output at the last time step is taken to be the prediction of the video clip. (Fig. 3c)

CACA-RNN A, Context-Aware Cascade Attention-based RNN, consists of two RNN in a cascade architecture with



(a) Face-RNN and Context-RNN

(b) Parallel-RNN

(c) Concatenated-RNN

(d) CACA-RNN

Fig. 3: Comparison of RNN-based emotion recognition architectures using face and/or context streams as input. Both streams are processed by a CNN feature extractor into corresponding feature stream as LSTM's input. (a) A single LSTM performs emotion prediction from a feature stream. (b) Two LSTMs reading two streams individually. Hidden states of the two LSTMs are fused together by fully-connected network for final prediction. (c) Two feature streams are concatenated together as input of single LSTM. (d) A cascaded RNN architecture of two LSTMs processing two feature streams: the left LSTM reads one feature stream and store a context vector in its hidden states. And the right LSTM reads the other feature stream with its hidden states is initiated by the left LSTM's context vector. The right LSTM's hidden states are generated from its input and attention from the left LSTM.

attention mechanism (Fig. 2). The left RNN reads context stream into a context vector. And the right RNN takes attention-processed context vector and face stream as input, then predicts output at the latest time.

CACA-RNN B has the same structure as CACA-RNN A (Fig. 2) except for its left RNN reads face stream and the right RNN reads context stream.

V. Experiments and Evaluation

A. Datasets

Chinese Natural Audio-Visual Emotion Database (CHEAVD) 2.0 is used by Multimodal Emotion Recognition Challenge (MEC) 2017 challenge. CHEAVD 2.0 includes 4917 training clips, 707 validation clips and 1406 testing clips from Chinese movies and TV programs. Each clip is labeled with one emotion category according to both video and audio content in 8 classes: "happy", "sad", "angry", "surprise", "disgust", "worried", "anxious", and "neutral".

To enlarge the training set, the submitted model was trained with additional private dataset. The dataset includes 4562 video clips ("happy": 409, "sad": 951, "angry": 212, "surprise": 357, "disgust": 270, "worried": 58, "anxious": 88 , and "neutral": 359) and training data clips are trimmed from the videos.

B. Feature Extraction

For each video clip, two streams are generated in preprocessing stage: "face stream" and "context stream". Face stream contains sequence of each detected faces from each frame, cropped and scaled to 128×128. With each detected face, corresponding context stream contains the original frame, center-cropped and scaled to 224×224.

Then, two pre-trained CNNs are used to extract features from the face and context streams as the input of CACA-RNN's face RNN and context RNN. The face feature extractor is a classifier trained with a private dataset for image emotion classification without last layer. And the context feature extractor is a pre-trained model from Squeezenet [27] with ImageNet classification [28], which the output of the last convolution layer is used as context feature.

C. Training Details

Videos are downsampled to 5 fps in both training and inference phases. In training phase, each video is sampled from a random initial frame for data augmentation. In evaluation phase, the initial frame is fixed to be the first frame of video clip.

For the submitted result, Adam [29] was used with learning rate 10^{-4} with mini-batch size 32. The weights of feature extractor CNNs are fixed while training. There is a pooling layer after the context CNN to down sample feature size to size 25088 as the input of context RNN. Also, a fully connected layer is added after the face CNN to encode face features with vector of size 128. Additionally, the face RNN is a two-layer LSTM with 128 hidden states. The context RNN is also a LSTM with 128 hidden states and attention mechanism is on face RNN.

Model	#params	mAP (%)	ACC (%)
Face-RNN	1.58M	38.87	52.72
Context-RNN	1.19M	28.74	37.68
Parallel-RNN	0.86M	40.44	54.30
Concatenated-RNN	1.38M	39.78	54.15
CACA-RNN A	0.79M	**40.73**	53.01
CACA-RNN B	0.79M	39.91	**54.44**

TABLE I: Experimental results on MEC2017 validation set.

D. Quantitative results

To explore performance of the architectures described in Section. IV, the validation results are summarized in Table. I, with *only* CHEAVD 2.0 was used for training and validation in the experiments.

To make comparison fair, the models are adjusted such that number of parameters are close. For each experiment, training was repeated five times with different random seeds and median performance metrics are selected among them. The

feature sizes of face feature and context feature are encoded to 128 by fully-connected layers. Face-RNN, Context-RNN and Concatenated-RNN has a two-layer LSTM with 256 hidden nodes. In Parallel-RNN, there are two two-layer LSTMs of 128 hidden nodes. And the first LSTM of CACA-RNN also has two layers with 128 hidden states, the second RNN is a LSTM with 128 hidden nodes.

Notably, the models fusing both context and face clues outperform the models using only one of the features. It shows that the context information is helpful but not enough to perform emotion recognition without cropped face.

For models fusing face and context features, the results show that Parallel-RNN outperforms Concatenated-RNN. Learning from multiple different feature sequences can be more effective for networks with multiple branches than a single branch model. The performance of CACA-RNN A and CACA-RNN B is higher than the others in evaluation in mAP (mean average precision) and accuracy. It shows that cascade attention-based RNN can improve the performance of handling two kinds of features.

VI. Results on MEC 2017

Five submissions are allowed to submit to video-based emotion recognition sub-challenge in MEC2017. Evaluation on mean average precision (mAP) and accuracy are both considered in MEC, two (of five) evaluation results of submissions are shown in Table. II: (1) "CACA-RNN", described in previous sections, the one with highest mAP, (2) "CACA-RNN+2D-3D-CNN", the one with highest best accuracy. Their confusion matrices are shown in Fig. 5.

CACA-RNN+2D-3D-CNN is an ensemble network of CACA-RNN and a 2D-3D-CNN network using linear weighted output from each model as the output prediction. The 2D-3D-CNN network is a spatiotemporal convolutional network, depicted in Fig. 4, learning temporal information from high level face feature maps from the same face feature extractor CNN in CACA-RNN.

Method	Validation Set		Testing Set	
	mAP (%)	ACC (%)	mAP (%)	ACC (%)
CACA-RNN	41.53	51.34	**45.51**	47.30
CACA-RNN+ 2D-3D-CNN	52.16	56.58	44.74	**52.99**
baseline [6]	34.1	36.5	21.7	35.3

TABLE II: Submitted Results to MEC2017.

VII. Conclusion

In this work, context information is first shown to be helpful in emotion perception. Then, different architectures using facial and/or context information for video emotion recognition was implemented and evaluation results were compared. It shows that the models using both information can achieve the highest accuracy.

Among them, a novel architecture, CACA-RNN, was proposed with a cascaded LSTM attention-based architecture to leverage both face and context information from video.

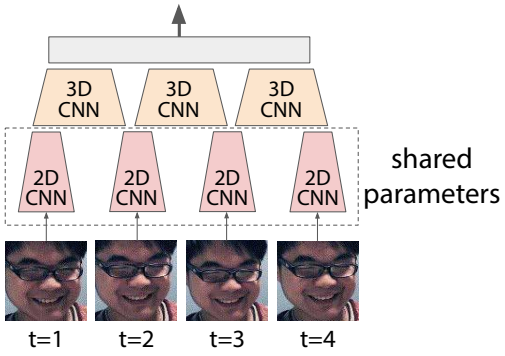


Fig. 4: A 2D-3D-CNN model for emotion recognition. The video stream is preprocessed into face stream of cropped faces. Each face of face stream is first encoded by a pre-trained CNN face feature extractor to high level feature maps. Then a 3D convolution neural network learns spatiaotemporal kernel from the feature map stream. There is an adaptive pooling before classifier to handle various length of input.

CACA-RNN has the best performance in the compared models. CACA-RNN consists of two LSTMs, context RNN and face RNN, processing context and face features respectively, and attention mechanism in the face RNN enables it to learn relationship to the context RNN and fuse information from two sequences. Compared to a multi-branch RNN (Parallel-RNN) and a single RNN handling concatenated features (Concatenated-RNN), CACA-RNN outperforms on evaluation MEC2017 validation dataset. The experiments also shows that context information improves for video emotion recognition. The models with additional context features perform better than the models using only face features.

CACA-RNN may be further extended to fusing more kinds of inputs for multi-modal emotion recognition or be extended to other video-based tasks such as action recognition.

References

[1] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion," Journal of personality and social psychology, vol. 17, no. 2, p. 124, 1971.

[2] J. A. Russell, "A circumplex model of affect," Journal of Personality and Social Psychology, 39, 1980, pp1161-1178.

[3] P. Ekman and W. V. Friesen, "Facial action coding system," 1977.

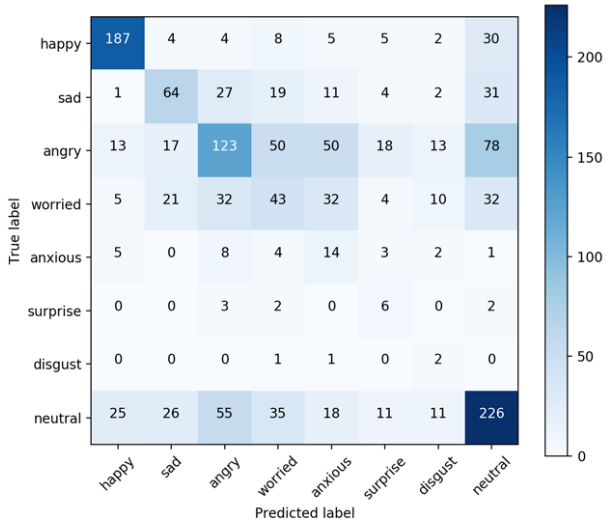
[4] S. Zafeiriou, M. Nicolao, I. Kotsia, F. Benitez-Quiroz, and G. Zhao, "Aff-wild: Valence and arousal in-the-wild challenge," In IEEE CVPR Workshop, 2017.

[5] A. Dhall, R. Goecke, S. Ghosh, J. Joshi, J. Hoey and T. Gedeon, "From individual to group-level emotion recognition: EmotiW 5.0," ACM ICMI 2017.

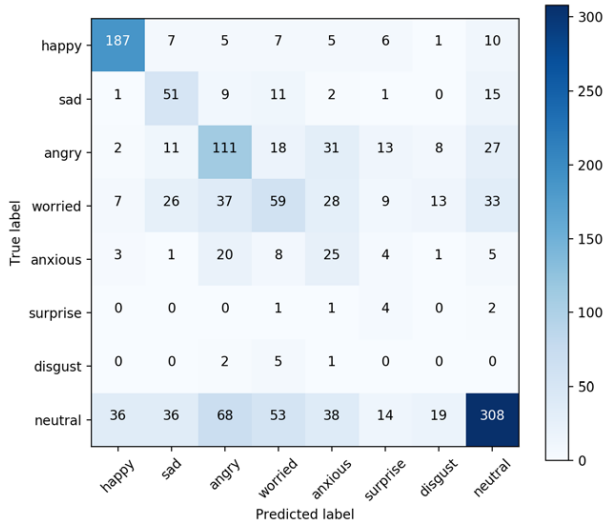
[6] Y. Li, J. Tao, B. Schuller, S. Shan, D. Jiang, and J. Jia, MEC 2017: the multimodal emotion recognition challenge, In the first Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia 2018), May, 2018.

[7] U. Hess and S. Hareli, "The influence of context on emotion recognition in humans," 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), 2015, pp. 1-6.

[8] L. F. Barrett, B. Mesquita and M. Gendron, "Context in emotion perception," Current Directions in Psychological Science, 2011, pp.286290.



(a) CACA-RNN



(b) CACA-RNN+2D-3D-CNN

Fig. 5: Confusion matrices of MEC2017 testing set. (a) The model has the highest mAP among five submissions in MEC2017. (b) The model has the highest accuracy among five submissions in MEC2017.

[9] R. Kosti, J. M. Alvarez, A. Recasens, and A. Lapedriza, "Emotion recognition in context," In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[10] M. Liu, R. Wang, S. Li, S. Shan, Z. Huang and X. Chen, "Combining multiple kernel methods on Riemannian Manifold for emotion recognition in the wild," ACM ICMI, 2014.

[11] A. Yao, D. Cai, P. Hu, S. Wang, L. Sha, and Y. Chen, "HoloNet: Towards robust emotion recognition in the wild," Proceedings of the 18th ACM International Conference on Multimodal Interaction, 2016, pp472478.

[12] S. Pini, Stefano, O. Ben-Ahmed, M. Cornia, L. Baraldi, R. Cucchiara and B. Huet, "Modeling multimodal cues in a deep learning-based framework for emotion recognition in the wild," 10.1145/3136755.3143006, 2017.

[13] Y. Fan, X. Lu, D. Li, and Y. Liu, "Video-based emotion recognition using CNN-RNN and C3D hybrid networks," in Proceedings of the 18th ACM International Conference on Multimodal Interaction. ACM, 2016, pp. 445450.

[14] P. Hu, D. Cai, S. Wang, A. Yao, and Y. Chen, "Learning supervised scoring ensemble for emotion recognition in the wild," Proceedings of the 19th ACM International Conference on Multimodal Interaction (ICMI), 2017, pp553-560.

[15] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and F. F. Li, "Large-scale video classication with convolutional neural networks," In Proc. CVPR, 2014.

[16] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," arXiv:1412.0767, 2014.

[17] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," DTIC Document, Tech. Rep., 1985.

[18] S. Hochreiter and J. Schmidhuber, "Long short-term memory," in Neural Computation, MIT Press, 1997.

[19] I. Sutskever, O. Vinyals, and Q. Le, "Sequence to sequence learning with neural networks," In Advances in Neural Information Processing Systems (NIPS), 2014.

[20] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," In ICLR, 2015.

[21] C. F. Benitez-Quiroz, R. Srinivasan, and A. M. Martinez, "Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild," In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[22] M. T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," In EMNLP, 2015.

[23] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," arXiv preprint arXiv:1411.4389, 2014.

[24] J. Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond short snippets: Deep networks for video classification," In CVPR, 2015.

[25] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," In Advances in Neural Information Processing Systems (NIPS), 2014.

[26] L. Yao, A. Torabi, K. Cho, N. Ballas, C. Pal, H. Larochelle, and A. Courville, "Describing videos by exploiting temporal structure," arXiv:1502.08029v4, 2015.

[27] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "Squeezenet: Alexnet-level accuracy with 50x fewer parameters and 0.5 mb model size," arXiv preprint arXiv:1602.07360, 2016.

[28] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, Alexander C. Berg, and F. F. Li, "Imagenet large scale visual recognition challenge. International Journal of Computer Vision," 115(3):211252, 2015.

[29] Kingma, Diederik P. and Ba, Jimmy, "Adam: A method for stochastic optimization," arXiv:1412.6980, 2014.