DAG-NET: DUAL ATTENTION GRAPH NETWORK FOR EMOTION RECOGNITION IN THE WILD

Anonymous ICME Submission

ABSTRACT

Emotion Recognition in the Wild (ERW) is a challenging task due to the unconstrained scenes in the wild environment. Using contextual information for attention based mechanism to capture saliency of features can be a promising approach to improve recognition performance. However, the approaches reported in literatures are either Facial Emotion Recognition (FER) based or posture based without contexts. In this paper, we propose a new Dual Attention Graph Network (DAG-Net) to explore the usage of contextual cues to enhance the performance for ERW. The proposed DAG-Net consists of three parallel modules, specifically, (1) the body attention module to suppress the uncertainties of body gesture feature representation, (2) the global spatial attention module based on salient context to capture the global effective region for emotion recognition, and (3) the local spatial attention module based on Graph Convolutional Network to find the local discriminative regions with emotion cues. Quantitative evaluations have been carried out on EMOTIC, an in-the-wild emotion dataset. The results demonstrated that our DAG-Net outperforms the state-of-the-art methods.

Index Terms—Emotion Recognition, Attention based mechanism, Image partition

1. INTRODUCTION

Emotion Recognition in the Wild (ERW) aims to distinguish human emotional states in natural environment, such as happy, fear, surprise, and sad. ERW is an important research field of affective computing, which relies on the techniques developed in computer vision. The ERW has been broadly used in human-machine interaction [1], autonomous driving, healthcare, advertising recommendation, etc.

One of the key challenges in ERW is the unconstrained scenes in real life environment, which is as opposed to a controlled laboratory environment with predefined scenarios, e.g. indoor or outdoor environment. Because the individuals are not restricted, their emotional states are more natural and real, and thus are more complicated to analyse. With the establishment of increasingly natural scenes expression databases [2-4] and the advancement of deep learning, ERW has achieved great success. A number of works [5-7] have studied this problem, and mainly fixate on facial expression, pose gesture, voice, and electroencephalogram (EEG) signal. However, the performance suffers from the varying environment. On the one hand, face or body features tend to have problems caused by illumination, occlusion and orientation in natural scenes, thus degrade the performance to some extent. On the other hand, the same behavior (Facial expression, body gesture) may present different emotional states in different scenes, e.g., when we consider the postures, garments and surroundings for the common act of looking at computer at home and in the office, we may come to different emotional states.

To better tackle the above challenges, we propose a Dual Attention Graph Network (DAG-Net) to recognize human emotion. The main idea is to introduce attention mechanism to capture feature dependencies in body-part and global context dimensions. Meanwhile, graph conventional network is appended in DAG-Net to exploit the local contextual information. Specifically, we add two types of attention mechanisms to the backbone network, namely the body attention module and the spatial attention module.

As mentioned before, the characteristic of person exhibits uncertainties due to outside disturbance, which can influence the learning of the facial and body features. So, body attention module is first designed to suppress these uncertainties.

Existing works [8-10] have shown evidence that place or social situation affects one's emotional state. So the global spatial attention module is also included to capture the salient context and to reduce the influence of ambiguous region. In other words, spatial attention mechanism is designed to guide the model to focus on the regions of interests for ERW.

In order to further obtain the discriminative regions for the context, we introduce a local spatial attention module, where the image partition to "zoom-in" local details. Specifically, the image is partitioned into patches. For each patch, a Graph Convolutional Network (GCN) is designed to extract the connected local details. All patches are built

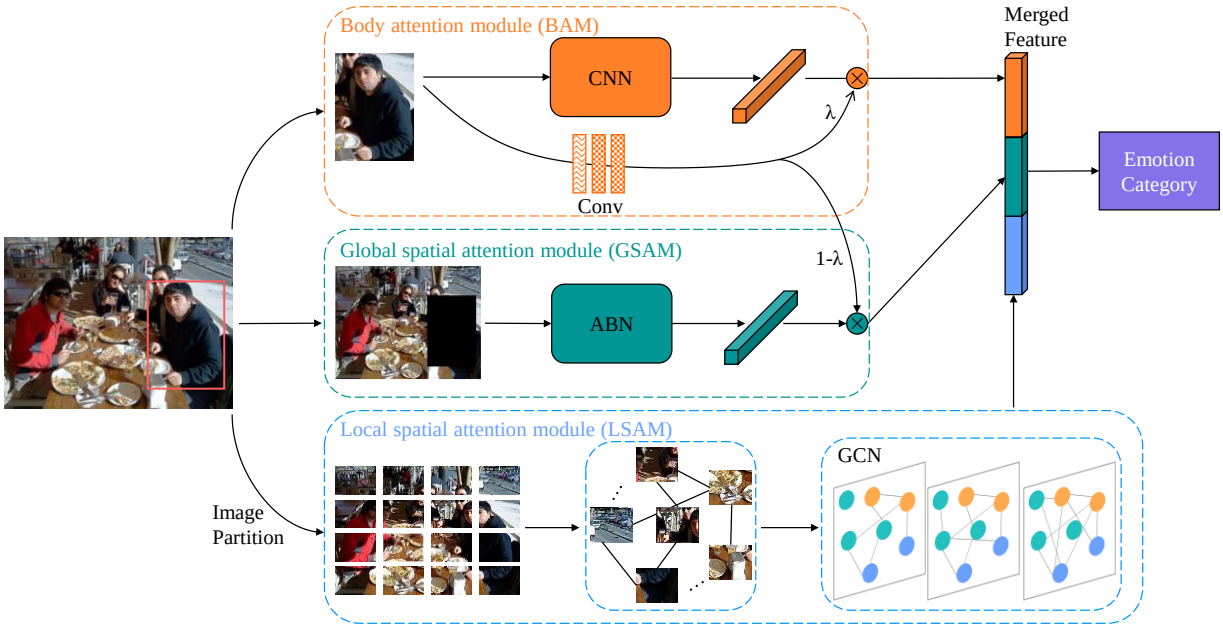

Fig. 1. The framework for proposed method.

as nodes of the graph, and the adjacent matrix is learned through the whole training process.

Our main contributions can be summarized as follows:

(1) We propose an end-to-end DAG-Net for ERW, which can suppress uncertainties and enhance useful cues automatically so as to boost features representation. (2) We combine image partition and GCN to obtain local details and find the semantic correlation. (3) Quantitative evaluations and extensive experimental results demonstrated the effectiveness of the proposed method.

2. RELATED WORK

Facial expression are the most straightforward work for emotion recognition [11, 12]. In [11], Facial Action Coding System (FACS) was utilized to encode the facial expression into a set of facial muscle movements. Zhang et al. [12] extracted SIFT features as input for convolutional neural networks (CNN) to learn the high-level semantic features related to expressions. More recently, Wang et al. [13] applied attention network to achieve better FER performance. In addition to FER, some studies have also used the body-part, which has more information such as posture, head movement, wearing, etc. Location of shoulders was used in [14] and body posture was employed to recognize 12 emotions under controlled laboratory in [6].

Apart from face and body, there are some work [10, 15, 16] using other visual cues to recognize emotions. In [15], global image and face individual are trained and fused them to obtain group emotion. To better investigate the emotions of individuals in natural scenes, Kosti et al. [10] construct EMOTIC dataset and propose a two-stream network, one stream is based on body information and the other focuses on contextual information, then fuse them to estimate emotion categories. Zhang et al. [16] further applied Faster R-CNN to detect context elements then fed them into the GCN as nodes to encode context, which outperforms the baseline in [10]. Generally speaking, the contextual information has been found to be a very important for Emotion Recognition in the natural scene, and the contribution is yet to be further improved. Unlike the above methods that use contextual information directly, we have proposed more sophisticated module to exploit both the global and local context and obtain a better representation of context features.

3. PROPOSED METHOD

3.1. Overview

We present our proposed model, DAG-Net, as shown in Fig. 1, the simple yet efficient framework consists of three streams: body stream, context stream and graph-based stream. Specifically, Body stream is used to extract individual features, context stream and graph-based stream is utilized to find emotion relevant regions.

Although the body stream and context stream have been employed in [10, 21], these methods encode body and contextual information by a backbone network, without discovering effective region information. The dual attention mechanism can better automatically inhibit or enhance feature representations. Hence it can use image

partition as another stream to identify the discriminative regions without prior knowledge.

3.2 Body attention module

The body attention module (BAM) endows with an emotional credibility weight for individual. Given a batch of body image $I_B = [x_1, x_2, \dots, x_N] \in \mathbb{R}^{3\times W\times H}$, where W and H are spatial dimension for x. The module let I_B takes as input and output the emotional credibility weight for each body image. Specifically, the body attention module is constructed from two convolution layers with 1×1 kernels, a global mean pooling and a sigmoid function. The definition of emotional credibility weight can be formulated as,

$$\lambda_i = \sigma(\mathrm{W}_a^{\mathrm{I}}\mathrm{x}_i) \quad (1)$$

where λ_i is the emotional credibility weight for the *i*-th body image, σ is a nonlinear function, and $\mathrm{W}_a^{\mathrm{I}}$ denotes the module parameters. After getting the attention weights, the feature vector ϑ_B can be obtained by multiplying the weight with features of each sample, which is formulated as,

$$\vartheta_B = \lambda \cdot \mathcal{F}(\mathrm{W}_B^{\mathrm{I}}; I_B) \quad (2)$$

where $\mathcal{F}$ is forward propagation and $\mathrm{W}_B^{\mathrm{I}}$ is the parameters of CNN feature extractor. Different from previous classical work [17] for attention mechanism, the body attention module is directly applied to the body image rather than to the feature map, which is more effective in determining the person's emotional credibility.

3.3 Global spatial attention module

Compared to the body stream, we use attention module to encode contextual information. Global spatial attention module (GSAM) is designed to train the attention map with image I_C. It is to highlight the attention map for visual interpretation that shows the regions of interest for emotion recognition. We introduce the state-of-the-art attention model, Attention Branch Network (ABN) [18], as spatial attention module.

$$g'_{CH}(I_C) = \mathrm{M}(I_C) \cdot g_C(I_C) \quad (3)$$

Equation 3 describes the Global spatial attention module mechanism, it is a dot-product between attention map $\mathrm{M}(I_C)$ and feature map $g_C(I_C)$ in particular channel *CH*. Note that image I_C has been pre-processed with masking, which is better for Global spatial attention module to focus on contextual information except body expression.

3.4 Local spatial attention module

In order to take full advantage of contextual cues, the local spatial attention module (LSAM) is designed to further exploit the contextual information. Previous studies [16] detect context elements to encode the scenes, which requires detection algorithms to find emotion relevant elements with prior knowledge. However, graph-based stream can easily extract the effective region for an image partition on the basis of region-based visual explanation. The research [19] has demonstrated that location information can enhance the feature presentation and improve performance on classification and detection tasks. Given an input image $I \in \mathbb{R}^{3\times W\times H}$, we uniformly partition it into $N\times N$ patches denoted by $R_{i,j}$, where *i* and *j* are the indices and $1 \le i,j \le N$. Each patch has $3\times\frac{W}{N}\times\frac{H}{N}$ dimensions.

Image partition destruct the visual appearance of image, which means noisy information also be introduced. To tackle this problem, we aggregate GCN to understand the relationship between each patch from the non-Euclidean data, which can prevent over-fitting the noise patches. $R_{i,j}$ is put into feature extractor to obtain 1024-dimension feature vector ϑ_{ij}, which as a node to construct graph. Specifically, the feature extractor is comprised of two convolution layers with 3×3 kernels generating 64 and 64 feature channels, following a mean pooling.

A GCN with nodes feature $\mathcal{V} = [\vartheta_{11}, \vartheta_{12}, \dots, \vartheta_{NN}] \in \mathbb{R}^{N\times N\times 1024}$ and an adjacency matrix $A \in \mathbb{R}^{N\times N}$ as inputs. For the *l*-th GCN layer H^l, it can be written as:

$$H^l = \sigma(AH^{l-1}W^{l-1}) \quad (4)$$

where $H^0 = \mathcal{V}$ and $W^{l-1} \in \mathbb{R}^{d'\times d}$ is the weighted matrix in layer *l*-1 with d' and d refers to the input and output feature dimension of (*l*-1)-th hidden layer. The adjacency matrix A is learned during the back-propagation process. Specifically, we use three GCN layers with 1024, 1024, and 512 output feature dimension, respectively. Each layer follows a Relu layer.

4. EXPERIMENTS

4.1 Dataset

We evaluate the performance of our proposed DAG-Net on EMOTIC dataset [10], which contains 34,320 persons labeled in unconstrained environment. The annotation has 26 categories of emotions, each person has at least one emotion label. The split of training set and test set is 80:20.

Fig. 2. Ground truth (green box) and prediction results (blue box) on images randomly selected.

4.2 Experimental Settings

The proposed DAG-Net was implemented using Pytorch. We set learning rate to be 0.0001, batch size to be 32 and N set to be 4. We resize I_B and I_C to 224×224. For data augmentation, set color jitter to 0.4 and apply random horizontal flip. We use ResNet-50 and ResNet-18 as backbone networks to extract feature vectors in body stream and context stream, respectively. Following [20], we used Kullback-Leibler divergence for back propagation.

4.3 Experimental Results

We use AP (Average Precision) as evaluation metric for multi-label classification. The proposed work is first compared with three state-of-art methods. Then, we conduct ablation studies to evaluate the main components of proposed method. The results of the following state-of-art work are included for comparison:

1. Kosti et al. [20] designed a dual branch network, one branch encode body information and another focus on the contextual information, two branches fused with Fully Connected layer.
2. Zhang et al. [16] used Region Proposal Network (RPN) to extract context elements, then fed into GCN to learn the affective relationship.
3. Bendjoudi et al. [21] built three modules for emotion classification and devised Multi-label Focal Loss (MFL) to deal with imbalanced data.

4.3.1 Performance

The results are present in Table 1. We can see that the mean Average Precision (mAP) of ours is higher than the other methods. Specifically, compared with [20] and [21], where contextual information are also included in the module ,our DAG-Net achieved about 2.58% and 1.63% mAP improvement individually. Clearly, these results illustrate the superiority of our proposed method can better exploit

Table 1. The comparison on Average Precision and mean Average Precision on EMOTIC

Category	In [20]	In [21]	In [16]	Ours
Affection	27.85	31.92	**46.89**	34.03
Anger	9.49	13.94	10.87	**15.55**
Annoyance	14.06	17.42	11.23	**19.1**
Anticipation	58.64	57.73	**62.64**	56.46
Aversion	7.48	8.18	5.93	**10.69**
Confidence	**78.35**	75.29	72.49	75.42
Disapproval	14.97	14.88	11.28	**18.32**
Disconnection	21.32	28.32	26.91	**30.53**
Disquietment	16.89	19.72	16.94	**21.52**
Doubt/Confusion	**29.63**	23.11	18.68	21.98
Embarrassment	3.18	**2.84**	1.94	2.55
Engagement	87.53	85.83	**88.56**	86.82
Esteem	**17.73**	16.72	13.33	16.11
Excitement	**77.16**	70.43	71.89	71.19
Fatigue	9.7	14.43	13.26	**14.76**
Fear	**14.14**	8.27	4.21	9.18
Happiness	58.26	76.61	73.26	**78.14**
Pain	8.94	9.38	6.52	**12.64**
Peace	21.56	24.31	**32.85**	24.27
Pleasure	45.46	46.89	**57.46**	48.49
Sadness	19.66	23.94	25.42	**32.96**
Sensitivity	9.28	6.28	5.99	**9.87**
Suffering	18.84	26.24	23.39	**34.01**
Surprise	**18.81**	10.07	9.02	12.23
Sympathy	14.71	13.98	**17.53**	14.19
Yearning	8.34	9.71	**10.55**	7.93
mAP	27.38	28.33	28.42	**29.96**

the context information. As for context elements based method from [16], our DAG-Net achieved absolute 1.54% improvement, which can further validate that combining global and local contextual information with attention mechanism can better improve performance.

It is worth noting that the performance of DAG-Net on some APs of emotion category are lower than the above methods, such as "Doubt/Confusion" and "Embarrassment", which may be caused by two reasons. One is the complexity of the scenes. Fig.2 shows some sample results, and it can be noticed that with the increasing complexity of scenes, the recognition performance degrades. Dataset category imbalance can be the other reason, e.g. "Embarrassment" is only about 1% [10].

4.3.2 Ablation Studies

In this section, we conduct two ablation studies, including the validity of three components, and effects of different values of N with GCN (w/ GCN) or without GCN (w/o GCN).

Effects of different components. To verify the individual contributions of the three modules of BAM, GSAM and LSAM, we remove each module one by one to conduct ablation experiments. The results of experiments are shown in Table 2. It can be seen that the combination of all three gives the best performance, and even the ones with only one module has better performance than methods in literatures, where GSAM component obtained the best performance boost.

Table 2. Ablation Experiments on EMOTIC Dataset

Methods	mAP
Kosti et al. [20]	27.38
Bendjoudi et al. [21]	28.33
Zhang et al. [16]	28.42
BAM	**28.99**
GSAM	**29.23**
LSAM	**29.09**
BAM + GSAM	**29.35**
BAM + GSAM + LSAM(w/o GCN)	**29.39**
DAG-Net (BAM + GSAM + LSAM)	**29.96**

Effects of different values of N for image partition. To search the appropriate value of N, we tried to change the values of N in a set of {2, 3, …, 9}, Fig. 3 shows the effects of different values of N. It is worth noting that, there is no result for N=1 due to it is a complete image, where no GCN is required. As shown in Fig.3, The appropriate value of N is 4 for EMOTIC dataset due to it strikes a better balance between memory footprint and performance.

As discussed in Section 3, we use GCN to solve the problem of visual appearance corruption caused by image partition. In order to further demonstrate the contribution of the GCN in the proposed module, we implemented another strategy without GCN for the evaluation of GCN's contribution. Specifically, for each patch, a fully connected layer is used to obtain d dimensional feature vector, then summing over $N \times N \times d$ feature vector to get the final d dimensional feature vector as the representation of the stream. As shown in Fig. 3, method with GCN performs better than the one without it. When N = 9, non GCN-based method drops significantly, but GCN-based method still tends to stable, which implied that GCN is beneficial for image partition in our model.

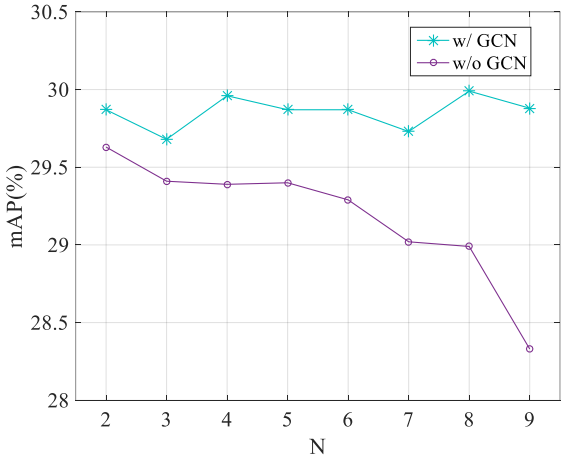

Fig. 3. The influence of partition number N and GCN.

5. CONCLUSION

In this paper, we presented a new DAG-Net for emotion recognition in the wild which consists of three attention modules for better emotion recognition performance in natural environment. The ablation experiment has demonstrated the effectiveness of each module being added to our model. Specifically, the BAM module can suppress body's uncertainties in natural scenes, while the GSAM focusing on global contextual information and the LASM on image partition with GCN further exploiting the discriminative region for emotion recognition. The experimental results conducted on EMOTIC dataset have shown that outperforms the state-of-the-art and validates the advantages of the proposed method. In the future, we consider incorporating action recognition and more visual cues of context to better understand the character's emotions in natural scenes.

6. REFERENCES

[1] R. Hortensius, F. Hekele, and E. S. Cross, "The Perception of Emotion in Artificial Agents," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 10, no. 4, pp. 852-864, 2018.

[2] J. Kossaifi, G. Tzirniropoulos, S. Todorovic, and M. Pantic, "AFEW-VA Database for Valence and Arousal Estimation In-the-wild," *Image and Vision Computing*, vol. 65, pp. 23-36, 2017.

[3] 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 *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5562-5570.

[4] Y. Li, J. Zeng, S. Shan, and X. Chen, "Occlusion Aware Facial Expression Recognition Using CNN With Attention Mechanism," *IEEE Transactions on Image Processing*, vol. 28, no. 5, pp. 2439-2450, 2019.

[5] Y. Luo, J. Ye, R. B. Adams Jr., J. Li, M. G. Newman, and J. Z. Wang, "ARBEE: Towards Automated Recognition of Bodily Expression of Emotion in The Wild," *International Journal of Computer Vision*, vol. 128, no. 1, pp. 1-25, 2020.

[6] N. Dael, M. Mortillaro, and K. R. Scherer, "Emotion Expression in Body Action and Posture," *Emotion*, vol. 12, no. 5, pp. 1085-1101, 2012.

[7] Y. Peng, R. Tang, W. Kong, and F. Nie, "A Factorized Extreme Learning Machine and Its Applications in EEG-Based Emotion Recognition," in *Neural Information Processing*, 2020, pp. 11-20.

[8] L. F. Barrett, B. Mesquita, and M. Gendron, "Context in Emotion Perception," *Current Directions in Psychological Science*, vol. 20, no. 5, pp. 286-290, 2011.

[9] J. Lee, S. Kim, S. Kim, J. Park, and K. Sohn, "Context-Aware Emotion Recognition Networks," in *IEEE International Conference on Computer Vision*, 2019, pp. 10142-10151.

[10] R. Kosti, J. M. Alvarez, A. Recasens, and A. Lapedriza, "Emotion Recognition in Context," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1960-1968.

[11] E. Friesen and P. Ekman, "Facial Action Coding System: A Technique for The Measurement of Facial Movement," *Palo Alto*, vol. 3, 1978.

[12] T. Zhang, W. Zheng, Z. Cui, Y. Zong, J. Yan, and K. Yan, "A Deep Neural Network-Driven Feature Learning Method for Multi-view Facial Expression Recognition," *IEEE Transactions on Multimedia*, vol. 18, no. 12, pp. 2528-2536, 2016.

[13] K. Wang, X. Peng, J. Yang, D. Meng, and Y. Qiao, "Region Attention Networks for Pose and Occlusion Robust Facial Expression Recognition," *IEEE Transactions on Image Processing*, vol. 29, pp. 4057-4069, 2020.

[14] M. A. Nicolaou, H. Gunes, and M. Pantic, "Continuous Prediction of Spontaneous Affect from Multiple Cues and Modalities in Valence-Arousal Space," *IEEE Transactions on Affective Computing*, vol. 2, no. 2, pp. 92-105, 2011.

[15] L. Tan, K. Zhang, K. Wang, X. Zeng, X. Peng, and Y. Qiao, "Group Emotion Recognition with Individual Facial Emotion CNNs and Global Image Based CNNs," in *ACM International Conference on Multimodal Interaction*, 2017, pp. 549-552.

[16] M. Zhang, Y. Liang, and H. Ma, "CONTEXT-AWARE AFFECTIVE GRAPH REASONING FOR EMOTION RECOGNITION," in *IEEE International Conference on Multimedia and Expo*, 2019, pp. 151-156.

[17] J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation Networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132-7141.

[18] H. Fukui, T. Hirakawa, T. Yamashita, and H. Fujiyoshi, "Attention Branch Network: Learning of Attention Mechanism for Visual Explanation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10705-10714.

[19] M. Noroozi and P. Favaro, "Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles," in *European Conference on Computer Vision*, 2016, pp. 69-84.

[20] R. Kosti, J. M. Alvarez, A. Recasens, and A. Lapedriza, "Context Based Emotion Recognition Using EMOTIC Dataset," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 11, pp. 2755-2766, 2020.

[21] I. Bendjoudi, F. Vanderhaegen, D. Hamad, and F. Dornaika, "Multi-Label, Multi-Task CNN Approach for Context-Based Emotion Recognition," *Information Fusion*, 2020.