ISSN:1738-6764, eISSN:2093-7504

Multimodal Attention-Based Fusion Model for Context-Aware Emotion Recognition

Minh-Cong Vo [1] **and Guee-Sang Lee** [2,*]

[1] Dept of Artificial Intelligence Convergence, Chonnam National University; congvm.it@gmail.com
[2] Dept of Artificial Intelligence Convergence, Chonnam National University; gslee@jnu.ac.kr
* Correspondence

https://doi.org/10.5392/IJoC.2022.18.3.011

Manuscript Received 28 January 2022; Received 13 September 2022; Accepted 13 September 2022

Abstract: *Human Emotion Recognition is an exciting topic that has been attracting many researchers for a lengthy time. In recent years, there has been an increasing interest in exploiting contextual information on emotion recognition. Some previous explorations in psychology show that emotional perception is impacted by facial expressions, as well as contextual information from the scene, such as human activities, interactions, and body poses. Those explorations initialize a trend in computer vision in exploring the critical role of contexts, by considering them as modalities to infer predicted emotion along with facial expressions. However, the contextual information has not been fully exploited. The scene emotion created by the surrounding environment, can shape how people perceive emotion. Besides, additive fusion in multimodal training fashion is not practical, because the contributions of each modality are not equal to the final prediction. The purpose of this paper was to contribute to this growing area of research, by exploring the effectiveness of the emotional scene gist in the input image, to infer the emotional state of the primary target. The emotional scene gist includes emotion, emotional feelings, and actions or events that directly trigger emotional reactions in the input image. We also present an attention-based fusion network, to combine multimodal features based on their impacts on the target emotional state. We demonstrate the effectiveness of the method, through a significant improvement on the EMOTIC dataset.*

Keywords: Context-aware; Multimodal fusion; Attention-based fusion; Emotion recognition; Deep learning

1. Introduction

For many years, human emotion recognition has been an important topic and has attracted many researchers in the computer vision field. Researchers have been trying to model and apply this ability for different applications, such as robotics, entertainment, surveillance, e-commerce, games, human-computer interaction, and more. In the computer vision field, emotion recognition is encoding emotional information in the input image or video to predict the emotion. Many attempts have considered facial expressions as the main factors for emotional perception.

Consequently, the current performance is still limited in the wild settings because the same facial movement could have different meanings in different situations [1, 2]. Some previous studies in psychology [3-6] show that emotional perception is affected by facial expression and contextual information such as body pose, human activity, social interaction, and background context.

Inspired by these explorations in psychology, previous studies in computer vision have examined contextual information to improve performance. [7, 8] utilized multi-modal approaches to implicitly capture the contextual information from the scene fused with facial and body pose features. [9] assume that the entire set of people in an image share the social identity (emotion, etc.). [10] utilizes the context elements relationships for emotion recognition. However, the contextual information has not been fully exploited yet. The scene emotion created by the surrounding environment can shape how people perceive emotion [1, 2]. For example, people tend to feel happy when they are in a happy place (birthday party, etc.).

Another essential component is a combination/fusion mechanism on multiple input contexts. The common choices of fusion mechanism are sum or concatenation, which are considered the contribution of each modality to the final prediction. However, not all modalities are helpful and informative on the final output. For example, the emotional scene gist is crucial when the main agents stand lonely without showing any facial expressions.

This paper proposes a multi-modal and context-aware network to capture the emotional scene gist information to infer the emotion prediction. Besides, an intention-based fusion network is introduced to combine multi-modal features based on the impact of each modality on the target prediction.

The contribution of the paper is to explore the effectiveness of the emotional scene gist in the input image to infer the emotional state of the primary target. Also, an attention-based fusion network to combine multimodal features based on their impacts on the target emotional state has been newly introduced.

2. Related Works

This section may be divided by subheadings. It should provide a concise and precise description of the experimental results, their interpretation as well as the experimental conclusions that can be drawn.

2.1 Emotion Recognition in Psychological Research

The importance of contextual information for emotion recognition is well supported by previous studies in psychology. [1-3] argue that only using facial expression is not sufficient to detect the emotion of a person, since human emotion is heavily affected by different types of the context, including scene context, human activity, or social interaction. Researchers have also considered that emotional body language, human body pose and/or body motion [11-13] is another means to express emotions beside facial expression. [14, 15] shows that the presence or absence of another person affects the perceived emotional state of people. [6] suggest a psychological framework identifying three main contextual factors such as a person, situation, and culture affects our emotion perception.

2.2 Context-aware Emotional Recognition

Previous research studies are based on deep learning networks to exploit contextual information for emotion recognition. [7] and [8] introduce the two novel networks having similar architectures. Both of them are constructed in the two-stream block fashion followed by a fusion network. [8] captures facial features while [7] utilizes body features in the first stream and the other focuses on exploiting contextual information from the whole input image. [10] considers all context elements in the image and uses a Graph Convolution Network to encode context features. All previous methods assume that all of the people in the image share the same emotional state and some social identities.

2.3 Attention-based Fusion Techniques

The attention mechanism has been studied and used in convolutional neural networks for a wide range of tasks, e.g. machine translation [16], image caption generation [17], object detection [18], scene segmentation [19], etc. These mechanisms include early fusion [20], late fusion [21], or incorporate these two into a hybrid fusion. In emotion recognition, additive combinations [22,23] are the common choices for either early or late fusion. However, in the real-world, not every modality is equally reliable for every data point due to sensor noise, occlusions, etc. The contributions of each modality are not equal because of noises, variances of data features, etc. Recent works have also examined variations on more sophisticated data-driven [24], hierarchical [25,26] proposed multiplicative combination methods motivated by the idea that the contributions of each modality are different among data points.

3. Proposed Method

3.1 Emotional Scene Gist Extraction and dataset

As a human, we can perceive the emotion while watching a scene. Emotion created in the scene can be seen as an emotional feature which is beneficial to infer emotional state of main agent in the input image. To capture this feature, we pretrained a Convolution Neural Network (CNN) M_e on StockEmotion [27] dataset. StockEmotion is a large-scale emotion dataset including about 1.2 million images collected from Adobe Stock with 690 emotional classes. These emotional classes consist of four main types: emotions (disappointed, nervous, frustrated, etc.), feelings (unfortunate, severe, tranquil, etc.), actions (quarrel, threat, yell pray, etc.), and events (Christmas, Halloween, wedding funeral, nightmare, etc.). The model M_e is constructed and trained based on a proposal from [27]. For the input image I, we obtain emotional scene feature x_e by passing I through M_e.

3.2 Body Features Extraction

The relation between body poses and emotion perception has been studied over the last decades [12, 13][28,29]. Recent studies in [30,31] suggest that body poses/gestures are beneficial to infer emotional state. To extract body-related features x_b, the visible part of the body I is cropped and then passed through the body feature extraction module. These features contain important cues like head and pose or body appearance. This module is a CNN pre-trained with ImageNet dataset [32], which contains the category person.

3.3 Place Feature Extraction

Place category and place attribute are highly correlated with emotion expression. For example, people usually show Anticipation, Excitement and Confidence and less frequently show Sadness or Annoyance while playing sports. These observations show that some common senses knowledge patterns between places and emotions could be potentially extracted from the data. To capture place category and attribute from the scene, we trained a CNN M_p on PLACES dataset [33]. PLACES dataset is a huge dataset consisting of 10 million scene photographs and annotated with 476 scene semantic categories and attributes. The architect and training procedure is followed by a proposal in [33]. Place feature x_p is extracted by passing the input image I through M_p.

3.4 Multimodal Attention-based Fusion

Combining all extracted features by can be beneficial to infer emotion. The common choices of fusion mechanism are sum or concatenation, which treats the contribution of each modality to the final prediction equally. However, this way can make the model confuse to recognize which information is necessary. Given the extracted feature vector for each module, we propose an attention-based fusion module to automatically fuse multimodal input by estimating the impact of these features corresponding to the final emotion. The importance of those features is computed as followed:

$$\mathrm{X} = \left(\mathrm{x_b} \oplus \mathrm{x_p} \oplus \mathrm{x_e}\right) \tag{1}$$

$$\mathrm{W} = \mathrm{softmax}\left(\mathcal{F}(\mathrm{X})\right) = \left[\mathrm{w_b}; \mathrm{w_e}; \mathrm{w_p}\right] \tag{2}$$

$$\mathrm{x_a} = \mathrm{w_b} * \mathrm{x_b} + \mathrm{w_e} * \mathrm{x_e} + \mathrm{w_p} * \mathrm{x_p} \tag{3}$$

Where $\oplus$ is a concatenation operator. $\mathcal{F}(.)$ is a simple Neural Network with three outputs shown in Figure 2. $\mathrm{W} \in \mathrm{R}^{3\mathrm{x}1}$ is a normalized context weights presenting for three modules body, place, and the emotional scene gist. A fusion vector $\mathrm{x_a}$ is combination of three inputs based on the weight matrix W.

3.5 Multimodal Loss Function

To deal with multi-label problem with an inherent class imbalance issue, we leverage the discrete category loss function proposed from [7] to calculate the loss of each modality and the fusion network. The loss is defined as follows:

$$\mathrm{L}(\mathrm{y}, \hat{\mathrm{y}}) = \sum_{\mathrm{i}=1}^{26} \mathrm{w_i}(\hat{y_\imath} - \mathrm{y_i})^2 \tag{4}$$

$$w_i = \frac{1}{ln(c+p_i)} \tag{5}$$

Where y_i is the ground-truth label and $\hat{y_\imath}$ is the prediction for the i-th category. The parameter w_i is the weight assigning to each category, where p_i is the probability of the i-th category and c is a parameter to control the range of valid values for w_i. The total loss is computed by summing all of the losses from each modality as follows:

$$L_{total} = \alpha L_b(y,\widehat{y_b}) + \beta L_e(y,\widehat{y_e}) + \gamma L_p(y,\widehat{y_p}) + \delta L_a(y,\widehat{y_a}) \tag{6}$$

where α, β, γ, δ are scalar coefficients controlling the importance of each loss.

4. Experimental Result

4.1 EMOTIC Dataset

EMOTIC dataset [7] is a huge dataset created for context-aware emotion recognition. All of the images were collected from MSCOCO [34] and ADE20K [35] along with images downloaded from web searches. The dataset consists of 23,571 images, with 34,320 people annotated for 26 discrete emotional classes and continuous Valence, Arousal, and Dominance dimensions. The emotion categories have a wide range of emotional states, including Peace, Happiness, Esteem, Anticipation, Engagement, Confidence, Affection, Pleasure, Excitement, Surprise, Embarrassment, Sympathy, Doubt/Confusion, Disconnection, Fatigue, Yearning, Disapproval, Annoyance, Aversion, Anger, Sensitivity, Sadness, Disquietment, Pain, Fear, and Suffering. In this paper, we focus on the emotion classification problem, so only the 26 discrete categories are used in the experiments. The dataset is divided into three subsets: train, validation, and test set. For a fair comparison with other researches, we use the standard train, validation, and test splits from the original paper. Some examples of this dataset are shown in Figure. 1.

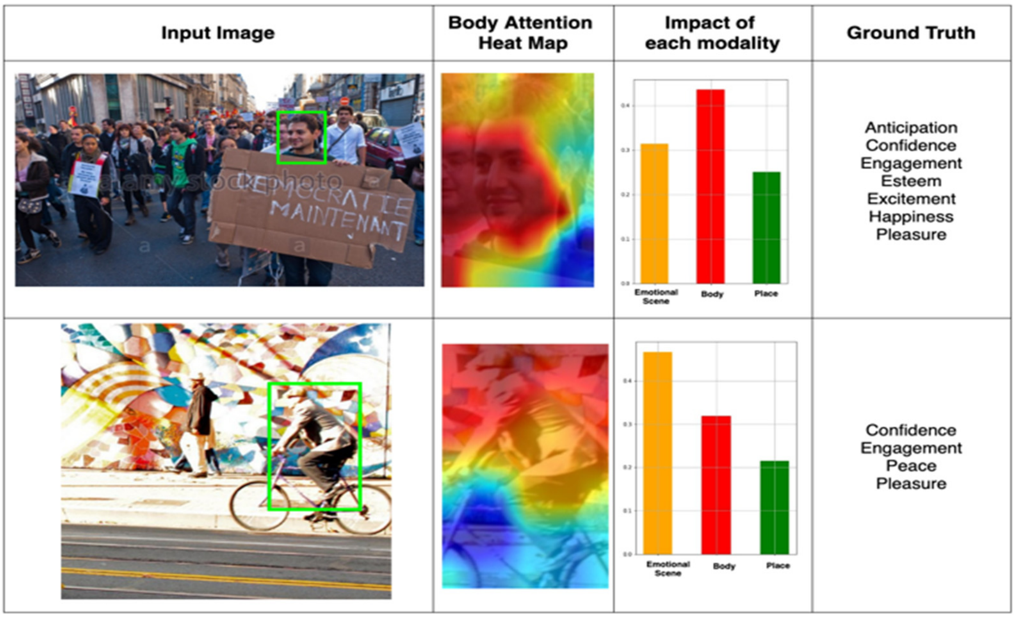

Figure 1. Qualitative Results: Illustration of attention block on randomly selected images. In the first row, given the location of the main agent, the body attention heat map clearly masks a happy face and the contribution of body information is significantly higher than the others. In the second row, the face and body pose of the main agent is not clearly shown emotion. However, the semantic context of the scene is quite peaceful, which contains the emotional information to predict the main agent's feelings.

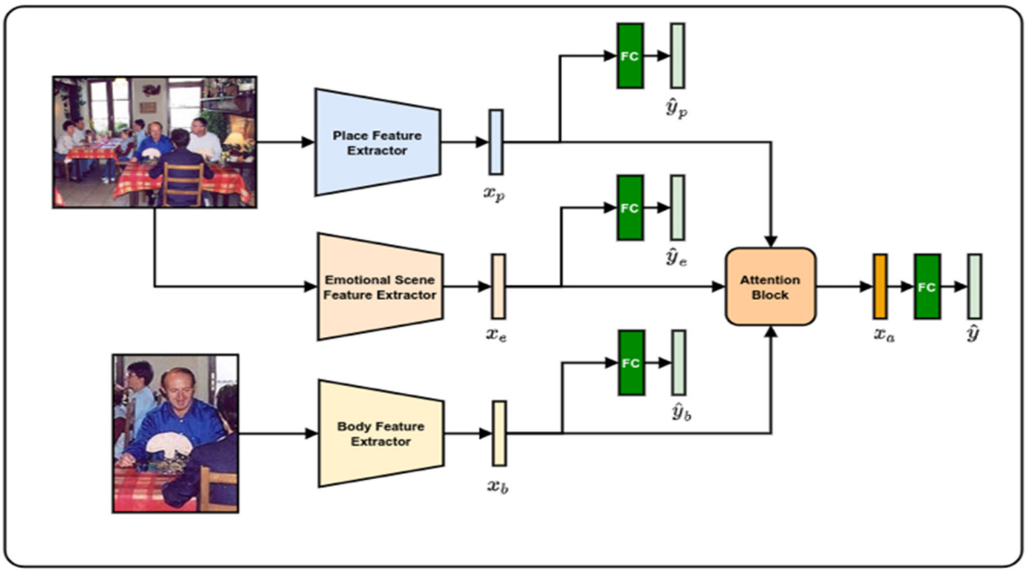

Figure 2. An overview of our pipeline. We use three contextual interpretations: face and body cues, place attributions, and emotional scene gist to predict the final emotion.

4.2 Implementation

Our method is implemented by using Pytorch [36] framework. For the place module and emotional scene gist extractors, we pretrain two ResNet50 [37] networks on PLACES dataset and StockEmotion dataset, respectively. For the body module, we use ResNet50 with ImageNet weights initialized while the fusion network is trained from the scratch. The body, place, and emotional scene input features are denoted by B, P, and E, respectively. To optimize the model, the Stochastic Gradient Descent optimizer is used in this experiment with a batch size of 256. The model is trained in 60 epochs with a learning rate decreasing from 0.01 to 0.0001 gradually following the Multistep learning rate strategy. For more generalization, we apply some augmentations: horizontal flip, grid distortion, shift-scale, and rotation. The model has experimented on a desktop PC with AMD Ryzen 7 2700X is equipped with a NVIDIA GTX 2080Ti GPU processor.

4.3 Results in EMOTIC Dataset

In this section, we show the results of various experiments for evaluating our model. We also show in detail the ablation studies to examine and evaluate each module in our model. Finally, we conclude with quantitative and qualitative results of the context interpretations.

4.3.1 Ablation Experiments

To highlight the importance of the emotional scene gist, we combine three kinds of interpretations and then remove them one by one. We notice that the body information is permanently retained because it contains all information of the main agents. We also demonstrate the effectiveness of the attention-based fusion by comparing this fusion with additive fusion(concatenation). The mean Average Precision(mAP) is used for evaluating the emotion recognition performance. The results of ablation experiments have been shown in Table 1.

Table 1. Performances on EMOTIC Dataset

Labels	[7]	[10]	[8]	Our
Affection	27.85	46.89	19.90	40.61
Anger	9.49	10.87	11.5	**27.32**
Annoyance	14.06	11.23	16.4	**22.9**

Anticipation	58.64	62.64	53.05	57.59
Aversion	7.48	5.93	16.2	9.29
Confidence	78.35	72.49	32.34	**79.14**
Disapproval	14.97	11.28	16.04	**20.50**
Disconnection	21.32	26.91	22.80	**28.40**
Disquietment	16.89	16.94	17.19	**20.39**
Doubt/Confusion	29.63	18.68	28.98	21.03
Embarrassment	3.18	1.94	15.68	2.96
Engagement	87.53	88.56	46.58	87.88
Esteem	17.73	13.33	19.26	17.54
Excitement	77.16	71.89	35.26	72.19
Fatigue	9.70	13.26	13.04	**18.16**
Fear	14.14	4.21	10.41	10.04
Happiness	58.26	73.26	49.36	**75.08**
Pain	8.94	6.52	10.36	**12.19**
Peace	21.56	32.85	16.72	28.08
Pleasure	45.46	57.46	19.47	49.59
Sadness	19.66	25.42	11.45	**38.50**
Sensitivity	9.28	5.99	10.34	**15.68**
Suffering	18.84	23.39	11.68	**40.95**
Surprise	18.81	9.02	10.92	12.02
Sympathy	14.71	17.53	17.125	17.46
Yearning	8.34	10.55	9.79	9.79
mAP	27.38	28.42	20.84	**32.13**

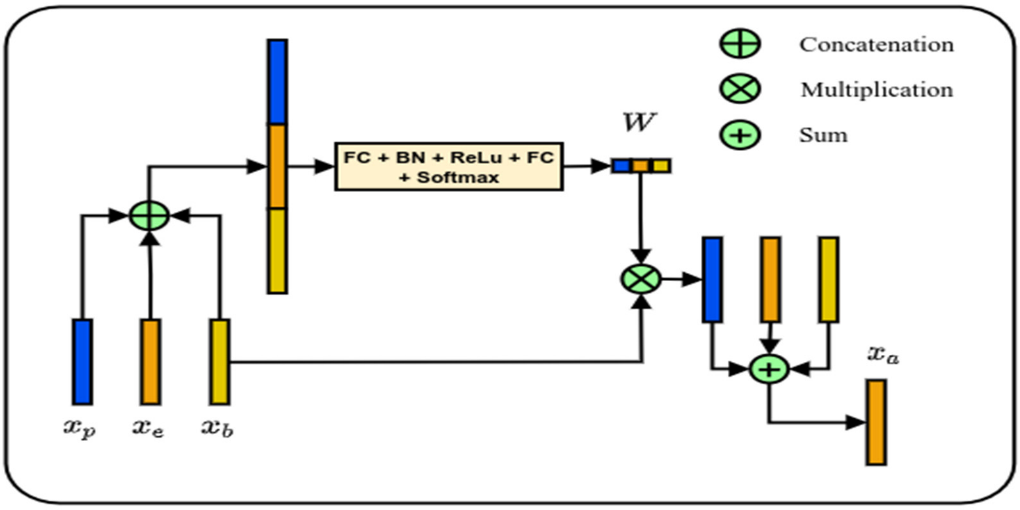

Figure 3. Illustration of attention fusion block

To evaluate the contribution of each modality, we run our model through the test set of the EMOTIC dataset and collect the impact weights from the attention-based fusion block. Our model's predictions are mainly based on the emotional scene gist because most of the faces, bodies, or places in the image are not very clear. In some cases where the image only contains the main agent, the contribution of body information is significantly higher than in other contexts. This explains the appearance of the tail distribution shown in Figure 4.

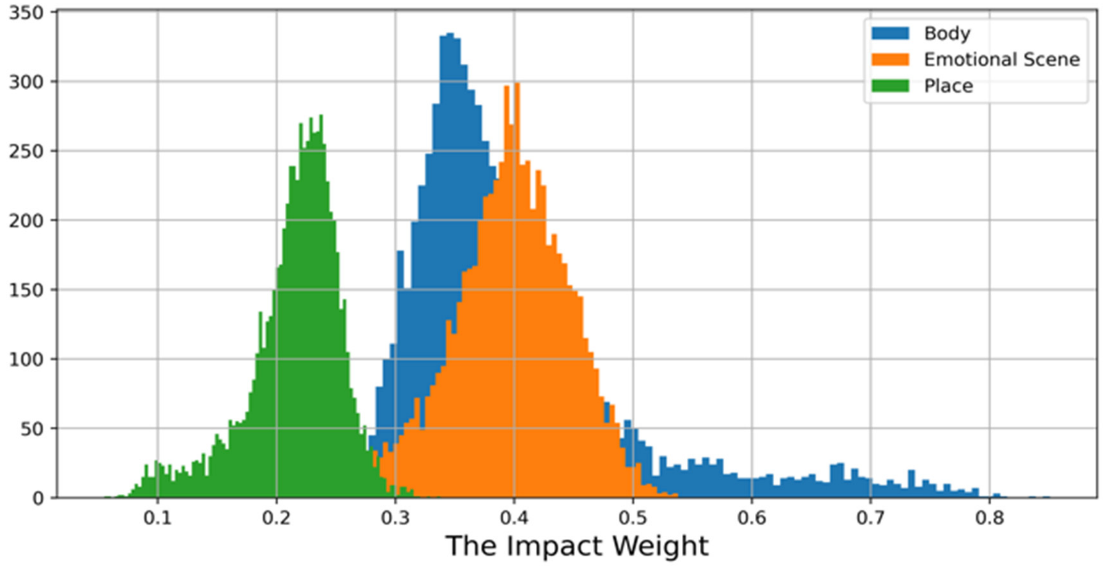

Figure 4. The distribution of impact weights of body, emotional scene gist, and place features on the EMOTIC test set. The importance of the emotional scene gist and body features highly outweigh the one of the place features. Notice that the sum of weights on three modalities equals 1 for a single image.

4.3.2 Qualitative and Quantitative Results

We show the qualitative results for two samples in Figure 3. To better understand what the model learned, we use GradCAM [38] to generate a heat map that indicates the influence of each area in the image corresponding to the final prediction. In the first row, the smiling face clearly expresses the happiness of the main agent so the contribution of face and body information is significantly higher than the others. In the second row, the face and body pose of the main agent is not clearly shown emotion. However, the semantic context of the scene is quite peaceful, which contains the emotional information to predict the main agent's feelings. We also summarize the evaluation of mean Average Precisions for all the methods conducting on the EMOTIC dataset in Table 1. Performance measures of different labels appear differently because of the variance in the size of the dataset and distribution characteristics of the data itself.

Table 2. Ablation Studies: We perform ablation experiments to understand how much of each modality and fusion mechanism benefits on EMOTIC dataset. In the experiments with additive fusion mechanism, we use concatenation operation to fuse multimodal features**.**

Labels	**Additive Fusion**			**Attention Fusion**
	B+P	**B+E**	**B+E+P**	**B+E+P**
Affection	27.02	37.33	40.42	40.61
Anger	10.79	23.88	26.25	27.32
Annoyance	15.89	16.89	22.65	22.9
Anticipation	56.74	57.33	57.69	57.59
Aversion	7.58	6.40	8.87	9.29

Confidence	76.31	75.21	78.54	79.14
Disapproval	13.81	17.03	20.15	20.50
Disconnection	24.8	24.08	27.72	28.40
Disquietment	15.77	19.74	21.52	20.39
Doubt/Confusion	29.63	18.68	28.98	21.03
Embarrassment	2.22	2.07	2.73	2.96
Engagement	86.02	86.01	87.78	87.88
Esteem	14.86	15.51	16.81	17.54
Excitement	69.42	68.45	71.24	72.19
Fatigue	9.87	17.83	18.45	18.16
Fear	5.99	8.37	9.57	10.04
Happiness	66.11	72.44	75.55	75.08
Pain	8.34	9.86	12.33	12.19
Peace	22.77	25.99	26.84	28.08
Pleasure	41.81	45.84	49.66	49.59
Sadness	18.32	32.86	38.74	38.50
Sensitivity	6.64	7.24	15.76	15.68
Suffering	20.62	31.51	41.05	40.95
Surprise	8.59	13.56	9.40	12.02
Sympathy	13.08	14.50	16.59	17.46
Yearning	7.53	8.80	9.49	9.79
mAP	25.67	29.30	31.83	32.13

5. Conclusion

In summary, this paper presents a multimodal attention-based fusion network for context-aware emotion recognition. Our model explores the emotional scene gist combining with the body (pose/gait) and places attributions. Our multimodal attention-based fusion allows us to decide the context feature should be more focused on for making a prediction based on their impacts on the target emotional state. The extensive experiments have demonstrated significant improvement in the performance of the current human emotion recognition system. Our results show that the emotional scene gist is a potential factor to automatically recognize human emotion in the wild and motivate further research in this direction

Acknowledgments: This research was supported by the Bio & Medical Technology Development Program of the National Research Foundation (NRF) & funded by the Korean government (MSIT) (NRF-2019M3E5D1A02067961) and by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2018R1D1A3B05049058 & NRF-2020R1A4A1019191). The corresponding author is Guee-Sang Lee.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

[1] 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. doi: https://doi.org/10.1177/0963721411422522.

[2] L. F. Barrett, “How emotions are made: The secret life of the brain,” Houghton Mifflin Harcourt, 2017. doi: https://psycnet.apa.org/doi/10.1037/teo0000098.

[3] A. M. Martinez, “Context may reveal how you feel,” Proceedings of the National Academy of Sciences, vol. 116, no. 15, pp. 7169-7171, 2019. doi: https://doi.org/10.1073/pnas.1902661116.

[4] B. Mesquita and M. Boiger, “Emotions in context: A sociodynamic model of emotions,” Emotion Review, vol. 6, no. 4, pp. 298-302, 2014. doi: https://doi.org/10.1177/1754073914534480.

[5] A. Aldao, “The future of emotion regulation research: Capturing context,” Perspectives on Psychological Science, vol. 8, no. 2, pp. 155-172, 2013. doi: https://doi.org/10.1177/1745691612459518.

[6] K. H. Greenaway, E. K. Kalolerinos, and L. A. Williams, “Context is everything (in emotion research),” Social and Personality Psychology Compass, vol. 12, no. 6, p. e12393, 2018. doi: https://doi.org/10.1111/spc3.12393.

[7] 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, 2019. doi: https://doi.org/10.1109/TPAMI.2019.2916866.

[8] J. Y. Lee, S. R. Kim, S. O. Kim, J. G. Park, and K. H. Sohn, “Context-aware emotion recognition networks,” *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019. doi: https://doi.org/10.1109/ICCV.2019.01024.

[9] K. Wang, X. Zeng, J. Yang, and D. Meng, “Cascade attention networks for group emotion recognition with face, body and image cues,” *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, 2018. doi: https://doi.org/10.1145/3242969.3264991.

[10] M. Zhang, Y. Liang, and H. Ma, “Context-aware affective graph reasoning for emotion recognition,” *2019 IEEE International Conference on Multimedia and Expo (ICME) IEEE*, 2019. doi: https://doi.org/10.1109/ICME.2019.00034.

[11] De Gelder, Beatrice, “Towards the neurobiology of emotional body language,” Nature Reviews Neuroscience 7.3 (2006): 242-249, doi: https://doi.org/10.1038/nrn1872.

[12] Grezes, Julie, Swann Pichon, and Beatrice De Gelder, “Perceiving fear in dynamic body expressions,” Neuroimage 35.2 (2007): 959-967, doi: https://doi.org/10.1016/j.neuroimage.2006.11.030.

[13] Peelen, Marius V., and Paul E. Downing, “The neural basis of visual body perception,” Nature reviews neuroscience 8.8 (2007): 636-648, doi: https://doi.org/10.1038/nrn2195.

[14] Yamamoto, Kyoko, and Naoto Suzuki, “The effects of social interaction and personal relationships on facial expressions,” Journal of Nonverbal Behavior 30.4 (2006): 167-179, doi: https://doi.org/10.1007/s10919-006-0015-1.

[15] E. Jakobs, A. S. Manstead, and A. H. Fisxher, “Social context effects on facial activity in a negative emotional setting,” Emotion, vol. 1, no. 1, p. 51, 2001. doi: https://doi.org/10.1037/1528-3542.1.1.51.

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomea, L. Kaiser, and L. Polosukhin, “Attention is all you need,” 2017. arXiv preprint arXiv:1706.03762

[17] K. Xu, J. Ba, R. Kiros, K. H. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio, “Show, attend and tell: Neural image caption generation with visual attention,” *International conference on machine learning*, PMLR, 2015. doi: https://dl.acm.org/doi/10.5555/3045118.3045336.

[18] H. Hu, J. Gu, Z. Zhang, J. Dai, and Y. Wei, “Relation networks for object detection,” *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. doi: https://doi.ieeecomputersociety.org/10.1109/CVPR.2018.00378.

[19] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, “Dual attention network for scene segmentation,” *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. doi: https://doi.org/10.1109/CVPR.2019.00326.

[20] K. Sikka, K. Dykstra, S. Suchitra, and L. Gwen, “Multiple kernel learning for emotion recognition in the wild,” *Proceedings of the 15th ACM on International conference on multimodal interaction*, 2013. doi: https://dl.acm.org/doi/10.1145/2522848.2531741.

[21] H. Gunes and P. Massimo, “Bi-modal emotion recognition from expressive face and body gestures,” Journal of Network and Computer Applications, vol. 30, no. 4, pp. 1334-1345, 2007. doi: https://doi.org/10.1016/j.jnca.2006.09.007.

[22] S. H. Yoon, S. H. Byun, S. Dey, and K. M. Jung, “Speech emotion recognition using multi-hop attention mechanism,” *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE*, 2019. doi: https://doi.org/10.1109/ICASSP.2019.8683483.

[23] Y. L. Kim, H. L. Lee, and Emily Mower Provost, “Deep learning for robust feature generation in audiovisual emotion recognition,” *2013 IEEE international conference on acoustics, speech and signal processing. IEEE*, 2013. doi: https://doi.org/10.1109/ICASSP.2013.6638346.

[24] C. W. Lee, K. Y. Song, J. H. Jeong, and W. Y. Choi, “Convolutional attention networks for multimodal emotion recognition from speech and text data,” ACL, 2018., doi: http://dx.doi.org/10.18653/v1/W18-3304.

[25] S. Li, D. Weihong, and D. JunPing, “Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild,” *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017. doi: https://doi.org/10.1109/CVPR.2017.277.

[26] K. Liu, Y. Li, N. Xu, and P. Natarajan, “Learn to combine modalities in multimodal deep learning,” 2018. arXiv preprint arXiv:1805.11730.

[27] W. Zijun, Z. Jianming, L. Zhe, J. Y. Lee, B. niranjan, M. Hoai, and D. Samaras, “Learning visual emotion representations from web data,” *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. doi: https://doi.org/10.1109/CVPR42600.2020.01312.

[28] B. Gelder, “Towards the neurobiology of emotional body language,” Nature Reviews Neuroscience, vol. 7, no. 3, pp. 242-249, 2006. doi: https://doi.org/10.1038/nrn1872.

[29] K. M. Meeren, C. R. J. Corne, and B. Gelder, “Rapid perceptual integration of facial expression and emotional body language,” Proceedings of the National Academy of Sciences, vol. 102, no. 45, pp. 16518-16523, 2005. doi: https://doi.org/10.1073%2Fpnas.0507650102.

[30] A. Kleinsmith, N. Bianchi-Berthouze, and A. Steed, “Automatic recognition of non-acted affective postures,” IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), vol. 41, no.4, pp. 1027-1038, 2011. doi: https://doi.org/10.1109/TSMCB.2010.2103557.

[31] S. Konrad, L. Gool, and B. Gelder, “Recognizing emotions expressed by body pose: A biologically inspired neural model,” Neural networks vol. 21, no. 9, pp. 1238-1246, 2008. doi: https://doi.org/10.1016/j.neunet.2008.05.003.

[32] J. Deng, W. Dong, R. Socher, L. Li, L. Kai, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” *2009 IEEE conference on computer vision and pattern recognition*, IEEE, 2009, doi: https://doi.org/10.1109/CVPR.2009.5206848.

[33] B. Zhou, A. Khosla, A. Lapedriza, A. Torralba, and A. Oliva, “Places: An image database for deep scene understanding,” 2016. arXiv preprint arXiv:1610.02055

[34] L. Tsung-Yi, M. Maichael, B. Serge, B. Lubomir, G. Ross, H. James, P. Pietro, R. Deva, C. Z. Lawrence, and D. Piotr, “Microsoft coco: Common objects in context,” European conference on computer vision. Springer, Cham, 2014, doi: https://doi.org/10.1007/978-3-319-10602-1_48.

[35] B. Zhou, Z. Hang, P. Xavier, X. Tete, F. Sanja, B. Adela, and T. Antonio, “Semantic understanding of scenes through the ade20k dataset,” International Journal of Computer Vision, vol. 127, no. 3, pp. 302-321, 2019. doi: https://doi.org/10.1007/s11263-018-1140-0.

[36] Paszke, Adam, et al, “Automatic differentiation in pytorch,” NIPS(2017)

[37] H. Kaiming, Z. Xiangyu, R. Shaoqing, and S. Jian, “Deep residual learning for image recognition,” *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, doi: https://doi.org/10.1109/CVPR.2016.90.

[38] Selvaraju, R. Ramprasaath, M. Cogswell, D. Abhishek, V. Ramakrishna, D. Parikh, and B. Dhruv, “Grad-cam: Visual explanations from deep networks via gradient-based localization,” *Proceedings of the IEEE international conference on computer vision*, 2017, doi: https://doi.org/10.1109/ICCV.2017.74.