Emotion Recognition for Multiple Context Awareness: Supplementary Material

Dingkang Yang[1,2,], Shuai Huang[1,2,], Shunli Wang[1,2], Yang Liu[1], Peng Zhai[1,2], Liuzhen Su[1,2], Mingcheng Li[1,2], and Lihua Zhang[1,3,2,4()]

[1] Academy for Engineering & Technology, Fudan University
[2] Engineering Research Center of AI and Robotics, Ministry of Education, China
[3] Jilin Provincial Key Laboratory of Intelligence Science & Engineering, China
[4] AI & Unmanned Systems Engineering Research Center of Jilin Province, China
lihuazhang@fudan.edu.cn

1 Datasets

1.1 HECO and Comparison with Related Datasets

Collection. HECO consists of images from Human-Object Interaction (HOI) datasets, film clips, and images from the Internet. The process of collecting is *strictly controlled and supervised* to ensure that HECO is free from potential negative impacts and to respect ethical behaviour generally. For the HOI datasets (V-COCO [3], HICO-DET [1]), which contain images of agents interacting with objects and scene contexts, we collect images with high resolution and clearly visible agent subjects. Moreover, we collect film clips from Web. After that, the film clips are divided to the shot with shot boundary detection, removing face-undetected shots and ambiguous shots. To avoid similarity and redundancy between samples, we manually select images with rich context information instead of processing the frame rate. Additionally, we collect images of agents in Google Engine by searching for healthy keywords (*e.g.*, *Warmth*, *Sports*, and *Partying*).

Annotation. We use Visual Object Tagging Tool (VOTT) interface to annotate emotions. The labelling process involves 3 psychologists and 10 graduate students. Psychologists utilize professional courses in cognitive psychology to train annotators rigorously. Only all annotators who pass the evaluation test are allowed to annotate. The annotation is performed blindly and independently, *i.e.*, the judgement of each annotator is not influenced by others. Figure 1 shows the eight discrete categories in the HECO. Furthermore, we annotate each image, including multiple agents with recognizable multimodal features, in an elaborate way that is rarely achieved in previous datasets. Figure 2 shows examples of agents with different levels of each one of VAD [7] dimensions. *Valence* (V) measures how positive or pleasant an emotion is. *Arousal* (A) measures the

Dinakang Yang and Shuai Huang : Equal contribution.

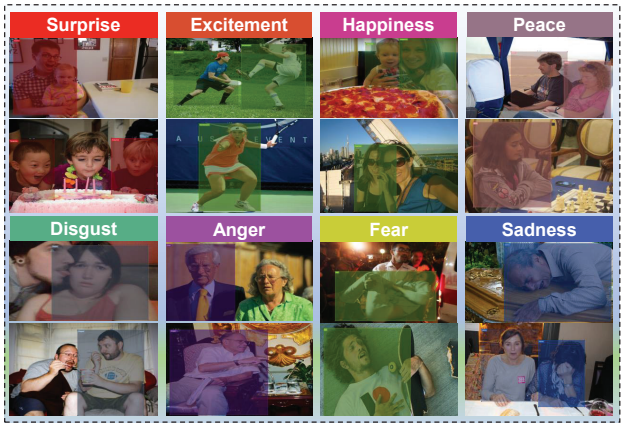

Fig. 1. Examples of the recognized agents with different discrete emotion categories are included in the HECO.

Fig. 2. Examples of the recognized agents with different scores of *Valence* (row 1), *Arousal* (row 2) and *Dominance* (row 3) are included in the HECO.

agitation level of the agent. *Dominance* (D) measures the control level of the situation by the agent. In Figure 3, we enforce numerical values to express relative percentages, showing each category's count and the distribution of continuous dimensions across different categories.

Moreover, following the emotion sociology theories [2, 10], we propose two novel label spaces : *Self-Assurance* (Sa) and *Catharsis* (Ca). Sa refers to the level of confidence in the agent's ability and judgement, *i.e.*, the agent conveys feelings of competence and adequacy, representing the degree to which the agent understands emotion at the cognitive level. Additionally, the emotion tendency of agents in social interaction occasions often depends on others or social status to update, producing an emotion catharsis called Ca. In emotion sociology, Ca reflects the influence of change in emotion from the agents on interaction and situation. Based on different expressiveness, Sa is divided into five categories,

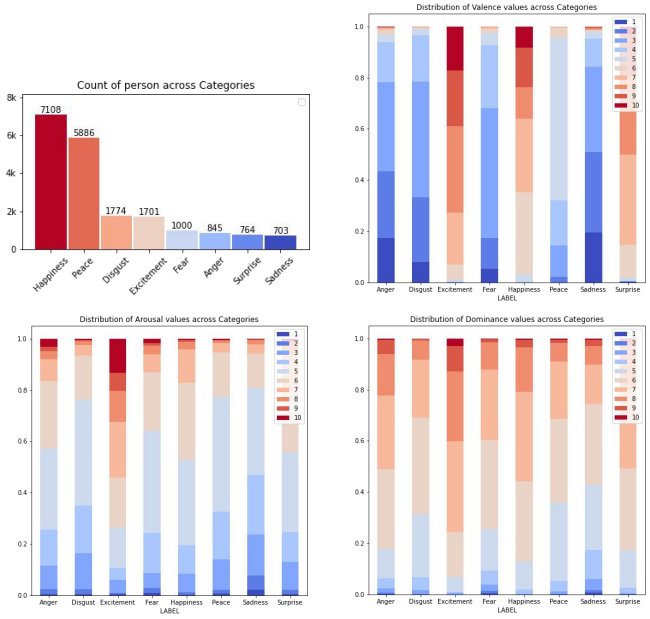

Fig. 3. Count and per each continuous dimension's distribution of the scores across the different categories.

including *Extremely-Sa*, *Slightly-Sa*, *Neutral*, *ExtremelyN-Sa*, and *SlightlyN-Sa*, where N denotes as not (*e.g.*, *ExtremelyN-Sa* is the opposite of *Extremely-Sa*). Ca is divided into four categories, including *Highly-Ca*, *Mildly-Ca*, *Slightly-Ca*, and *Neutral*. The range of these labels depends on the VAD values.The scope function of Sa is defined as follows:

$$Sa = \frac{1}{1 + e^{-(\omega_1 V + \omega_2 A + \omega_3 D)}}. \tag{1}$$

The scope function of Ca is formulated as follows:

$$Ca = log\left(1 + e^{\omega_4 V^2 + \omega_5 A + \omega_6 D}\right), \tag{2}$$

where $\omega_i, i \in [1, 6]$ denote weight coefficients. As shown in Figure 4, each distribution of them approximately obeys the Normal Distribution $N(\mu, \delta^2)$. The result shows that the HECO is closer to a real-world representation. The model trained on the HECO has a more robust generalization.

Analysis. Existing context-aware emotion recognition datasets have several shortcomings and issues. As shown in Figure 5, the EMOTIC [5] contains a large sample of irregular annotations, *i.e.*, two or more agents are annotated in the same bounding box, which may cause the ambiguity of emotion expression

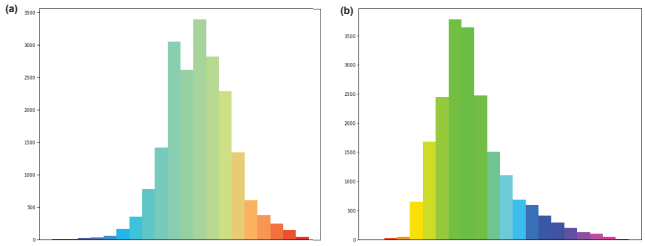

Fig. 4. (a) Distribution map of Sa. (b) Distribution map of Ca.

Fig. 5. We show some typical examples from EMOTIC [5], CAER-S [6], and HECO datasets. Due to intrinsic defects, there are several irregularly annotated samples in the EMOTIC and lots of similar samples in the CAER-S.

from the recognized agent. Furthermore, we observe many similar samples in the CAER-S [6], which could easily lead to overfitting the model. In comparison, the superiority of the HECO is demonstrated in *three aspects*. (1) The annotation of the images is strictly supervised, *i.e.*, ensuring that each bounding box contains a single agent subject in most cases. (2) The diversity and dissimilarity of the images are ensured by benefiting from different collection pathways and careful manual selection via professionally trained annotators. (3) An elaborate processing pattern ensures that the HECO is free from offensive content. Further, we compare the HECO with existing datasets including the EMOTIC [5], CAER-S

Table 1. Comparison with existing context-aware emotion recognition datasets. Specifically, *Agents Percentage* represents the percentage of image samples that contain multiple agents. *Sample Diversity* represents the coverage degree of agents in the samples from different stages.

Dataset	Setting	Discrete Categories	Continuous Dimensions	Annotated Agents	Agents Percentage	Sample Diversity
EMOTIC[5]	Web	✓	✓	$\sim 34k$	50%	**High**
CAER-S[6]	TV Show	✓	×	$\sim 70k$	38%	Middle
CAER[6]	TV Show	✓	×	$\sim 13k$	26%	Middle
GroupWalk[8]	Real Settings	✓	×	$\sim 3k$	46%	Low
HECO	Web, Films	✓	✓	$\sim 20k$	**52%**	**High**

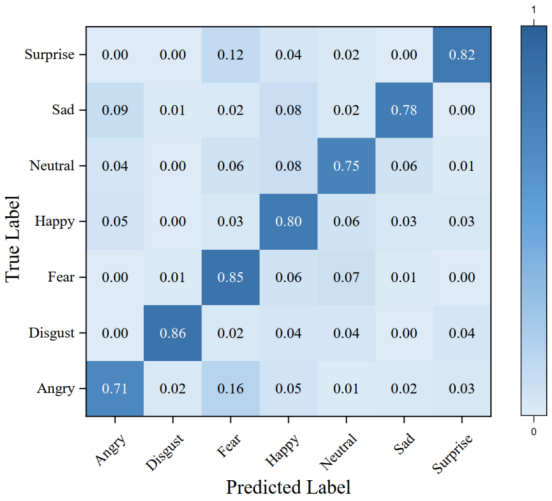

Fig. 6. Confusion matrix for classification accuracy on the CAER-S dataset.

Table 2. Comparison results on the CAER-S dataset.

Method	Accuracy
Fine-tuned VGGNet[9]	64.85
Fine-tuned ResNet[4]	68.46
CAER-Net-S[6]	73.51
Ours	**79.57**

[6], CAER [6], and GroupWalk [8] in Table 1. Note that the CAER-S is a subset of static images created by filtering video clips in the CAER. Through the statistics of the samples of the HECO, we show that the sum of image samples containing multiple agents in the HECO is around 52%, which is the most, *i.e.*, the percentages of individual images containing one, two, and three or more agents annotated are approximately 20%, 28%, and 52%, respectively. The percentages of the samples from children, teenagers, adults, and elderly are approximately 15.3%, 26.7%, 40.5%, and 17.5%, respectively. Meanwhile, we calculate the percentages of the samples from real-world and films is approximately 64.8% and 35.2%. Note that considering samples from *reality* and *acting* regarding emotion is beneficial to improve the generalization of the models trained on the HECO.

2 Comparison Results on the CAER-S Dataset

To further verify the generalization of our method, we perform multi-class classification using cross-entropy loss to conduct comparative experiments on the CAER-S dataset. Our experimental setups follow the previous method [6] exactly and use the standard dataset partitions of 7:1:2. The baseline methods are

to fine-tune the standard CNN networks on the CAER-S, such as VGGNet [9] and ResNet [4]. In Table 2, our method achieves a 6% improvement in accuracy over the previous methods. Moreover, the confusion matrix for the seven emotion categories is shown in Figure 6.

References

1. Chao, Y.W., Liu, Y., Liu, X., Zeng, H., Deng, J.: Learning to detect human-object interactions. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 381–389. IEEE Computer Society (2018)
2. Gordon, S.L.: The sociology of sentiments and emotion. In: Social psychology, pp. 562–592. Routledge (2017)
3. Gupta, S., Malik, J.: Visual semantic role labeling. arXiv preprint arXiv:1505.04474 (2015)
4. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
5. Kosti, R., Alvarez, J.M., Recasens, A., Lapedriza, A.: Context based emotion recognition using emotic dataset. IEEE transactions on pattern analysis and machine intelligence **42**(11), 2755–2766 (2019)
6. Lee, J., Kim, S., Kim, S., Park, J., Sohn, K.: Context-aware emotion recognition networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10143–10152 (2019)
7. Mehrabian, A.: Basic dimensions for a general psychological theory: Implications for personality, social, environmental, and developmental studies, vol. 2. Oelgeschlager, Gunn & Hain Cambridge, MA (1980)
8. Mittal, T., Guhan, P., Bhattacharya, U., Chandra, R., Bera, A., Manocha, D.: Emoticon: Context-aware multimodal emotion recognition using frege's principle. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14234–14243 (2020)
9. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
10. Stets, J.E.: Current emotion research in sociology: Advances in the discipline. Emotion Review **4**(3), 326–334 (2012)