BEAT: Berkeley Emotion and Affect Tracking Dataset

Zhihang Ren[1], Jefferson Ortega[1], Yunhui Guo[2], Stella X. Yu[1,3], David Whitney[1]
[1]University of California, Berkeley, [2]University of Texas at Dallas,
[3]University of Michigan, Ann Arbor
[1]{peter.zhren,jefferson_ortega,dwhitney}@berkeley.edu,
[2]yunhui.guo@utdallas.edu,[3]stellayu@umich.edu

Abstract

Recognizing the emotions of other humans is critical for our lives. Along with the rapid development of robotics, it is also crucial to enable machine recognition of human emotion. Many previous studies have focused on designing automatic emotion perception algorithms to understand the emotions of human characters. Limited by dataset curation procedures and small numbers of annotators, these algorithms heavily rely on facial expressions and fail to accurately reveal various emotional states. In this work, we build the first large video-based Emotion and Affect Tracking Dataset (BEAT) that contains not only facial expressions but also rich contextual information. BEAT has 124 *videos involving Hollywood movie cuts, documentaries, and homemade videos, and is annotated with continuous arousal and valence ratings as well as* 11 *categorical emotional states. We recruited* 245 *annotators, which guarantees the robustness of our annotations. The emotional annotations of BEAT span a wide range of arousal and valence values and contain various emotion categories. BEAT will be of great benefit to psychology studies on understanding human emotion perception mechanisms and the computer vision community to develop social-aware intelligent machines that are able to perceive human emotions.*

1. Introduction

Emotion perception is a ubiquitous need in our social lives. With the rapid development of robotics, intelligent machines should have the ability to understand human emotions in the human-populated world. Previous studies [7, 8, 12–16] have already attempted to design automated emotion perception algorithms. However, the dataset curation procedures were limited in various ways [1, 5, 6, 11, 15], including using lab-controlled (unnatural) backgrounds, short duration videos, and heavy focus on facial expressions, etc. Because of this, the algorithms are brittle: they fail to deal with more natural emotion perception tasks, where there is long-term temporal information as well as rich contextual information in addition to facial expressions [2–4]. In addition, most datasets [5, 10, 15] only contain a single type of annotation, either continuous annotations of arousal and valence or categorical emotion states. In this work, we build the first large video-based Emotion and Affect Tracking Dataset (BEAT) that contains not only facial expressions but also rich contextual information. BEAT has 124 videos involving Hollywood movie cuts, documentaries, and homemade videos, and is annotated with continuous arousal and valence ratings as well as 11 categorical emotional states.

2. Data Collection

We recruited 245 participants to annotate the videos in the BEAT dataset. All participants provided signed consent in accordance with the guidelines and regulations of the UC Berkeley Institutional Review Board and all experimental procedures were approved.

Participants watched and rated a total of 124 videos in the dataset. Annotators were instructed to track the valence and arousal or the emotion of the target character in the video by continuously moving their mouse pointer in real-time within the 2D valence-arousal grid or the emotion rating circle, respectively. Figure 1 illustrates the annotation procedures.

Before participants started the annotations, they were shown an image with the target character circled (Figure 1a) which informs the participants which character they will track when the video begins. In order to avoid annotator fatigue, all 124 video clips were split into two sessions. Participants rated the video clips in two sessions separately.

3. Visualization

Figure 2 shows the sample of continuous arousal and valence ratings as well as the categorical emotional ratings. The responses contain individual differences, highlighting the importance of collecting a large number of annotators, and revealing the pitfalls of previous work that employed

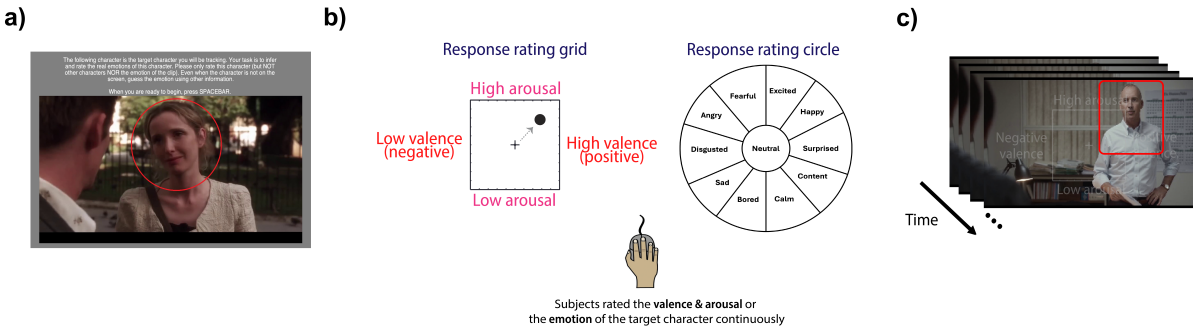

Figure 1. User interface used for video annotation. a) Participants were first shown the target character and were reminded of the task instructions before the start of each video. b) The overlaid valence and arousal grid / emotional states wheel that was present while observers annotated the videos. c) Observers were instructed to continuously rate the emotion of the target character in the video in real-time.

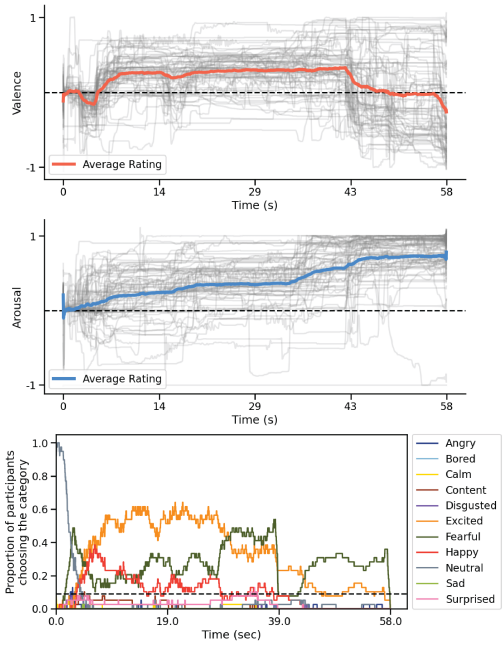

Figure 2. Example valence and arousal ratings and categorical emotion state ratings for a single video (video 78). Transparent gray lines indicate individual subject ratings and the red/blue line is the average rating across participants. For the categorical ratings, we show the proportion of participants choosing the specific emotion category. The final categorical rating is based on popularity voting.

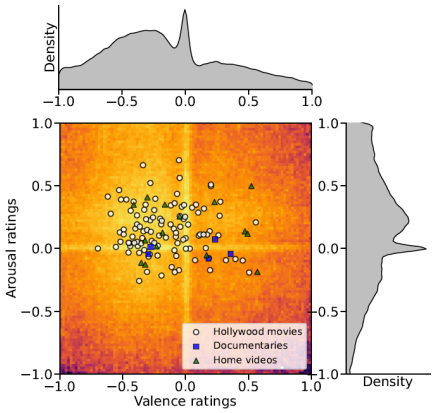

Figure 3. Distribution of valence and arousal ratings across participants. Individual white dots represent the average valence and arousal of the continuous ratings for each video clip for Hollywood movies. Blue squares and green triangles represent the average valence and arousal for documentaries and home videos, respectively.

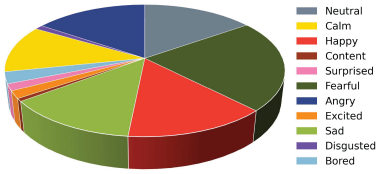

Figure 4. Distribution of 11 categorical emotion states across participants.

very few annotators [5, 9, 10, 15]. Because we recruited a large number annotators, the consensus among participants is clear in the data.

We also analyzed the dataset characteristics. Figure 3 shows the wide distribution of the continuous arousal and valence annotations. Figure 4 shows the various emotional states of our annotated data.

4. Future Works

The extensive BEAT dataset will be a rich resource for psychologists to investigate the mechanisms of human emotion perception. Furthermore, as a large video dataset for emotion and affect tracking, BEAT will also benefit the computer vision community in automatic emotion recognition, video understanding, scene understanding, video summarizing, and other applications.

References

[1] Pablo Barros, Nikhil Churamani, Egor Lakomkin, Henrique Siqueira, Alexander Sutherland, and Stefan Wermter. The omg-emotion behavior dataset. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE, 2018. 1

[2] Zhimin Chen and David Whitney. Tracking the affective state of unseen persons. *Proceedings of the National Academy of Sciences*, 116(15):7559–7564, 2019. 1

[3] Zhimin Chen and David Whitney. Inferential affective tracking reveals the remarkable speed of context-based emotion perception. *Cognition*, 208:104549, 2021.

[4] Zhimin Chen and David Whitney. Inferential emotion tracking (iet) reveals the critical role of context in emotion recognition. *Emotion*, 22(6):1185, 2022. 1

[5] Abhinav Dhall, Roland Goecke, Simon Lucey, Tom Gedeon, et al. Collecting large, richly annotated facial-expression databases from movies. *IEEE multimedia*, 19(3):34, 2012. 1, 2

[6] Ellen Douglas-Cowie, Roddy Cowie, Cate Cox, Noam Amir, and Dirk Heylen. The sensitive artificial listener: an induction technique for generating emotionally coloured conversation. In *LREC workshop on corpora for research on emotion and affect*, pages 1–4. ELRA Paris, 2008. 1

[7] Dimitrios Kollias and Stefanos Zafeiriou. Aff-wild2: Extending the aff-wild database for affect recognition. *arXiv preprint arXiv:1811.07770*, 2018. 1

[8] Dimitrios Kollias and Stefanos Zafeiriou. Expression, affect, action unit recognition: Aff-wild2, multi-task learning and arcface. *arXiv preprint arXiv:1910.04855*, 2019. 1

[9] Ronak Kosti, Jose M Alvarez, Adria Recasens, and Agata Lapedriza. Context based emotion recognition using emotic dataset. *IEEE transactions on pattern analysis and machine intelligence*, 42(11):2755–2766, 2019. 2

[10] Jiyoung Lee, Seungryong Kim, Sunok Kim, Jungin Park, and Kwanghoon Sohn. Context-aware emotion recognition networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10143–10152, 2019. 1, 2

[11] Fabien Ringeval, Andreas Sonderegger, Juergen Sauer, and Denis Lalanne. Introducing the recola multimodal corpus of remote collaborative and affective interactions. In *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, pages 1–8. IEEE, 2013. 1

[12] Jiahui She, Yibo Hu, Hailin Shi, Jun Wang, Qiu Shen, and Tao Mei. Dive into ambiguity: Latent distribution mining and pairwise uncertainty estimation for facial expression recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6248–6257, 2021. 1

[13] Kai Wang, Xiaojiang Peng, Jianfei Yang, Shijian Lu, and Yu Qiao. Suppressing uncertainties for large-scale facial expression recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6897–6906, 2020.

[14] Fanglei Xue, Zichang Tan, Yu Zhu, Zhongsong Ma, and Guodong Guo. Coarse-to-fine cascaded networks with smooth predicting for video facial expression recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2412–2418, 2022.

[15] Stefanos Zafeiriou, Dimitrios Kollias, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, and Irene Kotsia. Aff-wild: valence and arousal'in-the-wild'challenge. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 34–41, 2017. 1, 2

[16] Yuan-Hang Zhang, Rulin Huang, Jiabei Zeng, and Shiguang Shan. M 3 f: Multi-modal continuous valence-arousal estimation in the wild. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 632–636. IEEE, 2020. 1