Supplementary for VEATIC: Video-based Emotion and Affect Tracking in Context Dataset

Zhihang Ren*[1], Jefferson Ortega*[1], Yifan Wang*[1], Zhimin Chen[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,wyf020803,zhimin,dwhitney}@berkeley.edu,
[2]yunhui.guo@utdallas.edu,[3]stellayu@umich.edu

1. More About Stimuli

All videos used in the VEATIC dataset were selected from an online video-sharing website (YouTube). The VEATIC dataset contains 124 video clips, 104 clips from Hollywood movies, 15 clips from home videos, and 5 clips from documentaries or reality TV shows. Specifically, we classify Documentary videos as any videos that show candid social interactions but have some form of video editing, while home videos refer to videos that show candid social interactions without any video editing. All Videos in the dataset had a frame rate of 25 frames per second and ranged in resolution with the lowest being 202 x 360 and the highest being 1920 x 1080.

Except for the overview of video frames in Figure 2, we show more samples in Figure 1. Moreover, unlike previously published datasets where most frames contain the main character [2, 1, 3], VEATIC not only has frames containing the selected character but also there are lots of frames containing unselected characters and pure backgrounds (Figure 2). Therefore, VEATIC is more similar to our daily life scenarios, and the algorithms trained on it will be more promising for daily applications.

2. Annotation Details

In total, we had 192 participants who annotated the videos in the VEATIC dataset. Eighty-four participants annotated video IDs 0-82. One hundred and eight participants annotated video IDs 83-123 prior to the planning of the VEATIC dataset. In particular, Fifty-one participants annotated video IDs 83-94, twenty-five participants annotated video IDs 95-97, and 32 participants annotated video IDs 98-123.

Another novelty of the VEATIC dataset is that it contains videos with interacting characters and ratings for separate characters in the same video. These videos are those with video IDs 98-123. For each consecutive video pair, the video frames are exactly the same, but the continuous emotion ratings are annotated based on different selected characters. Figure 3 shows an example. In this study, we first propose this annotation process and it can provide future algorithms a way to test whether models learn the emotion of the selected characters given the interactions between characters and the exact same context information. A good emotion recognition algorithm should deal with this complicated situation.

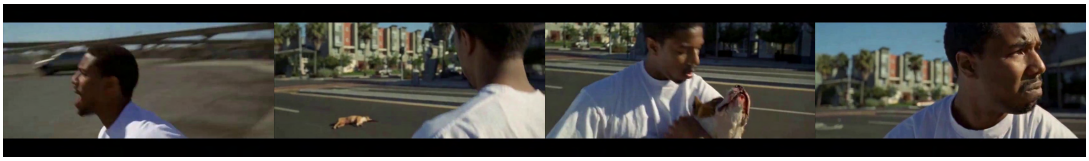
3. Outlier Processing

We assessed whether there were any noisy annotators in our dataset by computing each individual annotator's agreement with the consensus. This was done by calculating the Pearson correlation between each annotator and the leave-one-out consensus (aggregate of responses except for the current annotator) for each video. Only one observer in our dataset had a correlation smaller than .2 with the leave-one-out consensus rating across videos. We chose .2 as a threshold because it is often used as an indicator of a weak correlation in psychological research. Importantly, if we compare the correlations between the consensus of each video and a consensus that removes the one bad annotator, we get a very high correlation (r = 0.999) indicating that leaving the one bad subject does not significantly influence the consensus response in our dataset. Thus, we decided to keep the annotator in the dataset in order to not remove any important alternative annotations to the videos.

4. Subject Agreement Across Videos

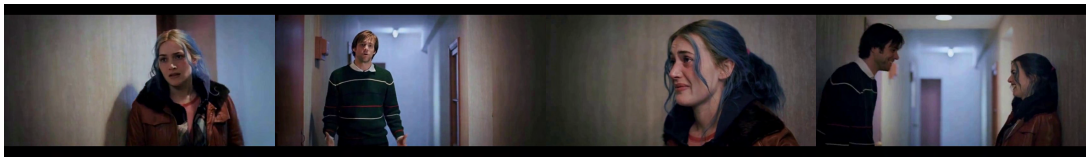
A benefit of the VEATIC dataset is that it has multiple annotators for each video with the minimum number of annotators for any given video being 25 and the maximum being 73. Emotion perception is subjective and observers

Video 2, *"Fruitvale Station"*, 2013, Hollywood Movie



Video 3, *"Patch Adams"*, 1998, Hollywood Movie

Video 10, *"Eternal Sunshine of the Spotless Mind"*, 2004, Hollywood Movie

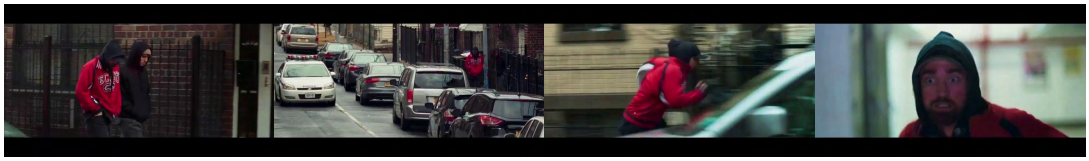


Video 18, *"Fargo"*, 1996, Hollywood Movie

Video 19, *"American Psycho"*, 2000, Hollywood Movie

Video 21, *"Crash"*, 2005, Hollywood Movie

Video 27, *"Good Time"*, 2017, Hollywood Movie



Video 48, *"Good Will Hunting"*, 1997, Hollywood Movie

Figure 1. More sample video frames in VEATIC. The video clips in VEATIC contain various backgrounds, lighting conditions, character interactions, etc., making it a comprehensive dataset for not only emotion recognition tasks but also other video understanding tasks.

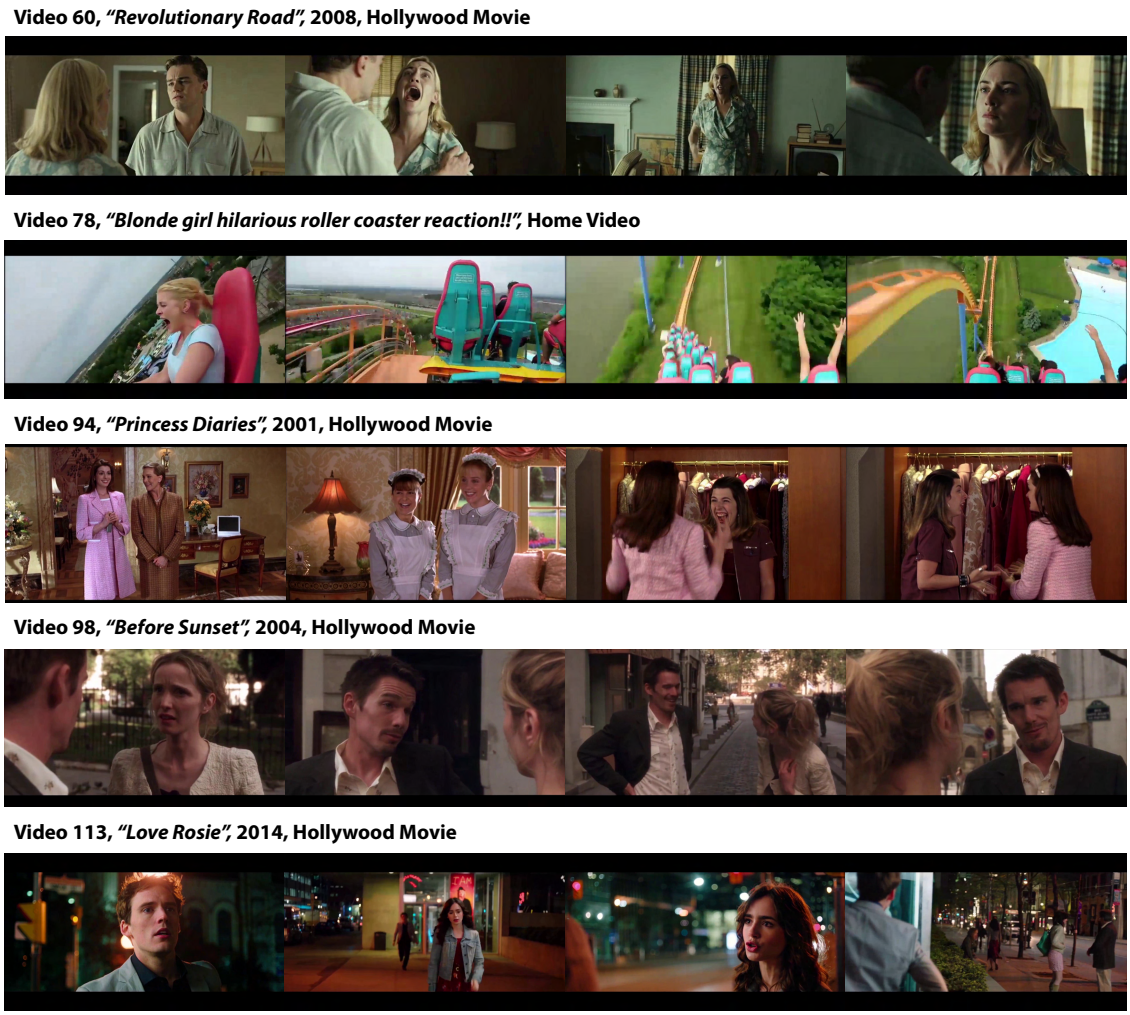


Figure 2. Sample video frames of unselected characters and pure background in VEATIC. The first sample frame in each row shows the selected character. The rest sample frames are either unselected characters or pure backgrounds.

judgments can vary across multiple people. Many of the previously published emotion datasets have a very low number of annotators, often having only single digit $(n < 10)$ number of annotators. Having so few number of annotators is problematic because of the increased variance across observers. To show this, we calculated how the average rating for each video in our dataset varied if we randomly sampled, with replacement, five or all annotators. We repeated this process 1000 times for each video and calculated the standard deviation of the recalculated average rating. Figure 4a shows how the standard deviation of the consensus rating across videos varies if we use either five or all annotators for each video. This analysis shows that having more annotators leads to smaller standard deviations in the consensus rating which can lead to more accurate representations of the ground truth emotion in the videos.

Additionally, We investigated how observers' responses varied across videos by calculating the standard deviation across observers for each video. Figure 4b shows the standard deviations across videos. We find that the standard deviations for both valence and arousal dimensions were small with valence having an average standard deviation of $\mu = 0.248$ and a median of 0.222 and arousal having an average standard deviation of $\mu = 0.248$ and a median of 0.244, which are comparable with the valence and arousal rating variance from EMOTIC [3].

5. Familiraity and Enjoyment Ratings

Familiarity and enjoyment ratings were collected for each video across participants which is shown in Figure 5. Familiarity and enjoyment ratings for video IDs 0-83 were collected in a scale of 1-5 and 1-9, respectively. Familiarity and enjoyment ratings for video IDs 83-123 were collected prior to the planning of the VEATIC dataset and were col-

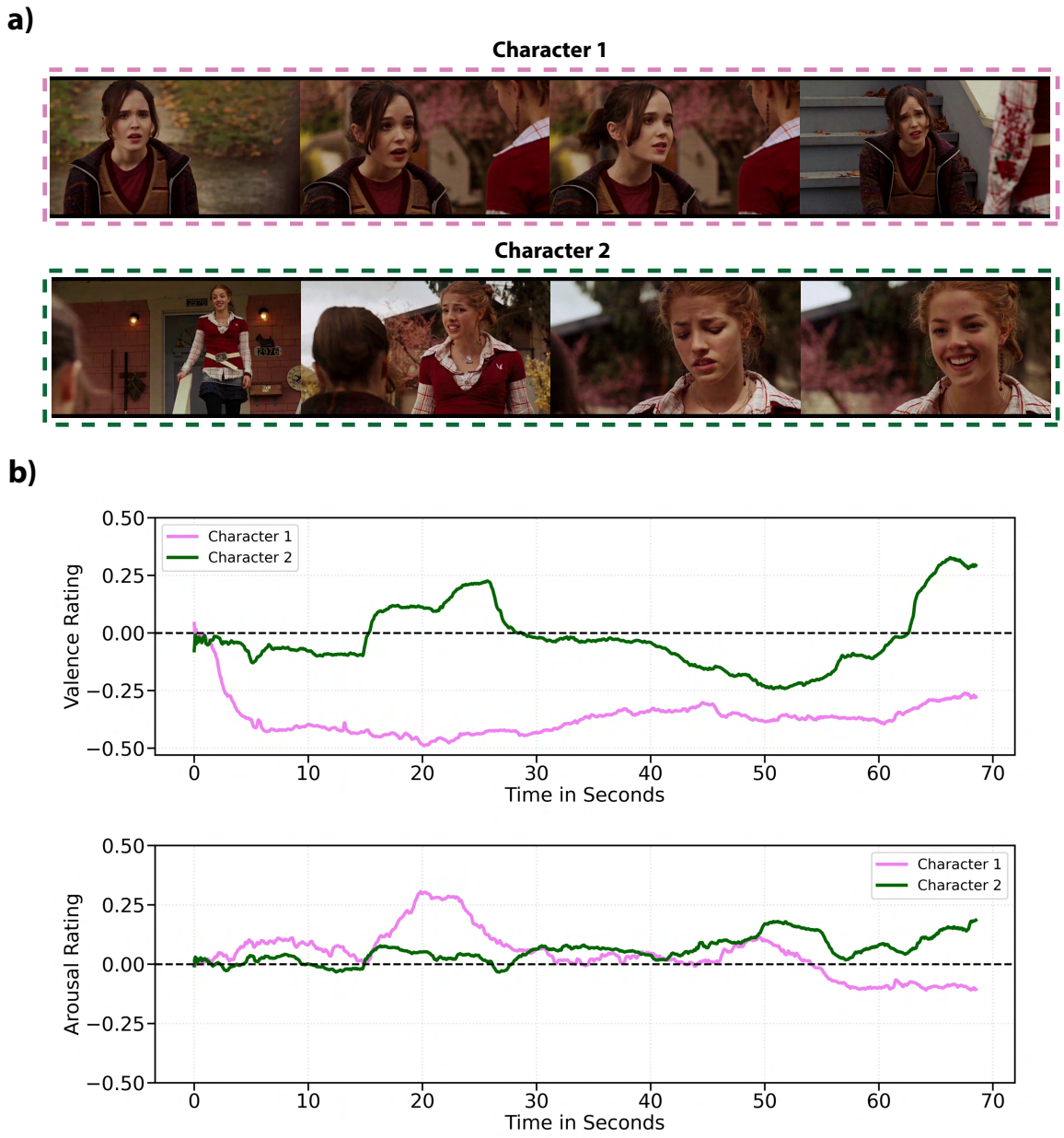


Figure 3. Example of different ratings of the same video in VEATIC. (a). The two selected characters. (b). The continuous emotion ratings of corresponding characters. The same color indicates the same character. A good emotion recognition algorithm should infer the emotion of two characters correspondingly given the interactions between characters and the exact same context information.

lected on a different scale. Familiarity and enjoyment ratings for video IDs 83-97 were collected on a scale of 0-5 and familiarity/enjoyment ratings were not collected for video IDs 98-123. For analysis and visualization purposes, we rescaled the familiarity and enjoyment ratings for video IDs 83-97 to 1-5 and 1-9, respectively, to match video IDs 0-83. To resale the familiarity values from 0-5 to 1-5 we performed a linear transformation, we first normalized the data between 0 and 1, then we multiplied the values by 4 and added 1. We rescaled the enjoyment values from 0-5 to 1-9 similarly by first normalizing the data between 0 and 1, then we multiplied the values by 8 and added 1. As a result, the average familiarity rating was 1.61 while the average enjoyment rating was 4.98 for video IDs 0-97.

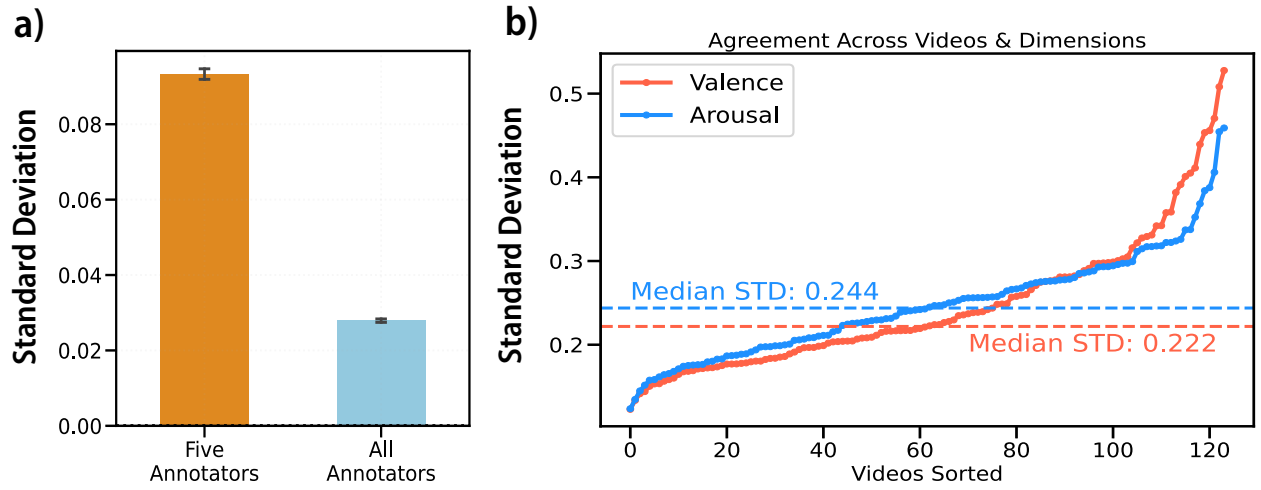


Figure 4. a)Five annotators' response standard deviation vs. all annotators'. b) Annotators' responses standard deviation of each video. Red and blue solid lines indicate the standard deviation of annotators' responses of valence and arousal on each video respectively. The results are sorted based on the standard deviation for visualization purposes. The dashed lines show the medians of standard deviations. The mean values for standard deviations of valence and arousal are the same with $\mu = 0.248$.

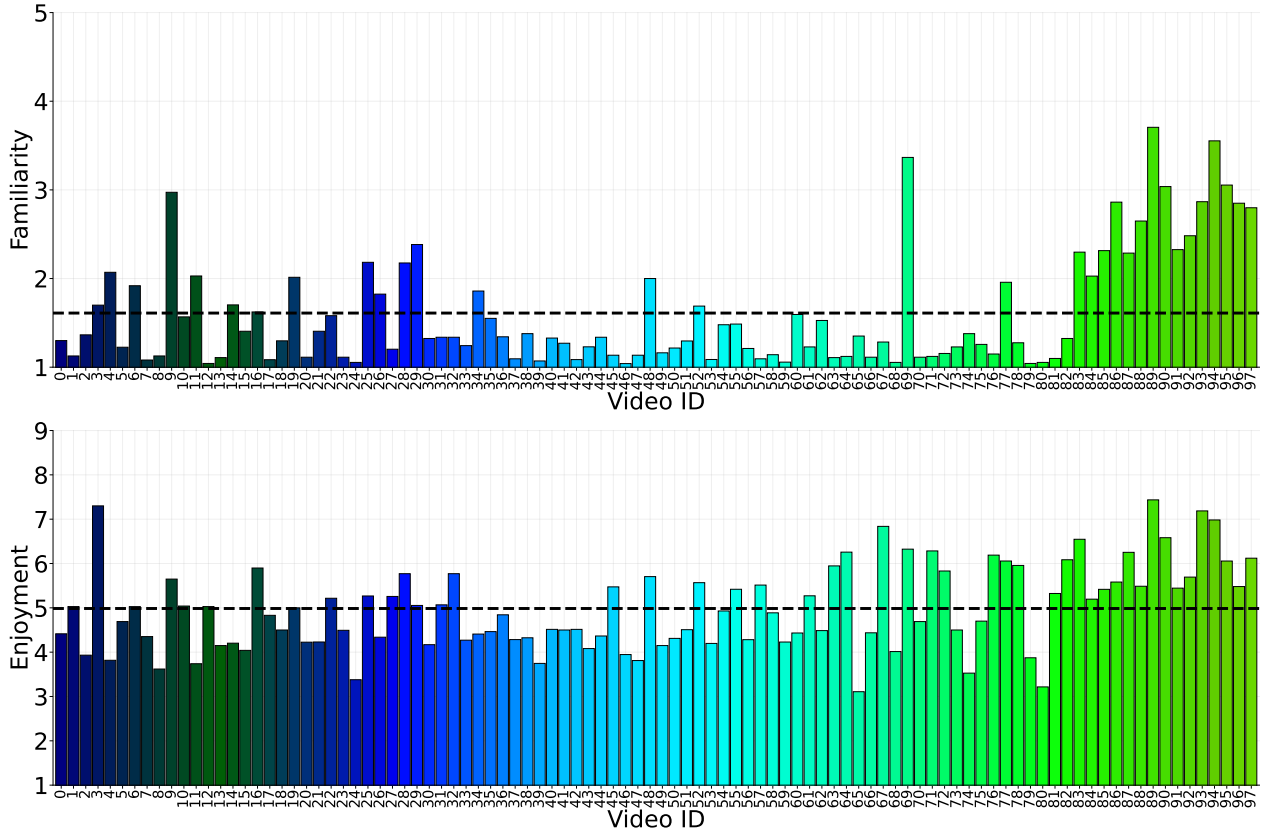


Figure 5. Familiarity and enjoyment ratings across all videos. Each bar represents the average familiarity or enjoyment rating reported by all participants who annotated the video. The average rating across all videos is depicted by the horizontal dashed line in both figures. Video IDs are shown on the x-axis.

References

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

[2] Jean Kossaifi, Georgios Tzimiropoulos, Sinisa Todorovic, and

Maja Pantic. Afew-va database for valence and arousal estimation in-the-wild. *Image and Vision Computing*, 65:23–36, 2017.

[3] 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.