DAiSEE: Towards User Engagement Recognition in the Wild

Abhay Gupta
Microsoft India
Hyderabad, India
abhgup@microsoft.com

Arjun D'Cunha
IIT Hyderabad
Hyderabad, India
cs14btech11039@iith.ac.in

Kamal Awasthi
IIIT Vadodara
Vadodara, India
201451072@iiitvadodara.ac.in

Vineeth N Balasubramanian
IIT Hyderabad
Hyderabad, India
vineethnb@iith.ac.in

The difference between real and virtual worlds is shrinking at an astounding pace. With more and more users working on computers to perform a myriad of tasks from online learning to shopping, interaction with such systems is an integral part of life. In such cases, recognizing a user's engagement level with the system (s)he is interacting with can change the way the system interacts back with the user. This will lead not only to better engagement with the system but also pave the way for a more improved and personalized human-computer interaction. But in order to try to understand and respond to human behavior, new challenges surface. Challenges not limited to a single image but a sequence of images interwoven with one another. Here, a system must understand the underlying psychology of a subject under observation. It must go beyond what the image represents and make an attempt to understand reactions of the subject to stimuli given by the system. In doing so, several contemporary vision applications including advertising, healthcare, autonomous vehicles, e-learning and many more have the possibility of undergoing revolutionary advances. However, the lack of any publicly available dataset to recognize user engagement severely limits the development of methodologies that can address this problem. In addition, the development of intelligent algorithms for processing and handling visual data is also dependent on the existence and availability of such datasets.

The progress from research to consumer technologies in classical recognition problems in computer vision have been possible over the last decade due to the availability of large-scale datasets [3, 6, 11] which are made available to researchers and industry practitioners alike. The recognition of user engagement is increasingly relevant to a digital world that floods users with various kinds of content, and it is useful for systems to be "aware" of the user's engagement while providing content. For example, in e-learning: which parts of a lecture are confusing for most students who watch it? how engaged are students in a video? This work is an effort in this direction - to provide a dataset (and benchmark results) for a vision problem of contemporary relevance: user engagement recognition in e-learning environments. Though, there have been research efforts to develop methods for user engagement recognition [23][10], the resulting datasets are very small and not publicly available.

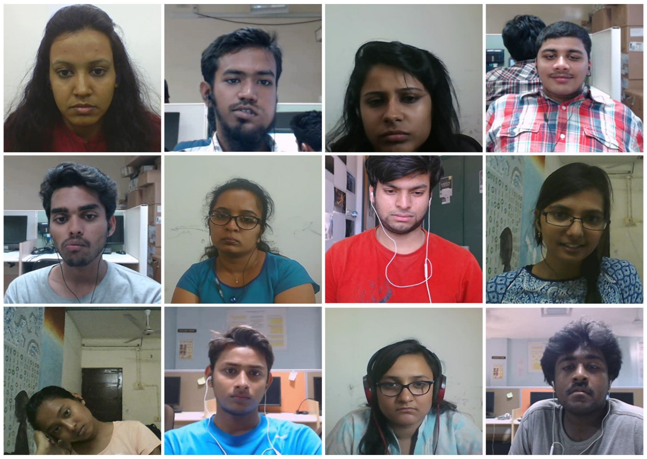

Figure 1: Examples of video frames from DAiSEE: The dataset captures real-world challenges of recognizing user engagement in natural settings.

In this work, we introduce DAiSEE (**D**ataset for **A**ffecti**v**e **S**tates in **E**-**E**nvironments) (please see Figure 1 for sample images), which will be made publicly available to the community for further research. In particular, we focus on engagement, frustration, confusion, and boredom as the affective states for this work, all of which are relevant to user engagement and concomitant applications. Considering that these affective states are subtle, the annotations for this dataset are crowd-sourced and then strengthened using a gold standard created using expert psychologists. (All annotations, including each individual crowd label, will also be shared along with this dataset.) We also benchmark the performance of standard video classification and deep learning-based models on this dataset to provide a baseline for further research. DAiSEE has the potential to be applied to the following domains, which may involve users seated and watching content on a display: (i) E-learning, to support personalized learning for an individual user, and thus, increase retention rate; (ii) Advertising, to gain insights into how engaging an advertisement is, and to then provide personalized advertising to customers; (iii) E-shopping, to understand user preferences in specific items or a larger domain (clothing, jewelry, electronics, etc.), and thus allow personalization in the shopping experience; or (iv) Autonomous vehicles, to capture the driver's engagement level, or to predict the driver's future actions based on his/her confusion or frustration levels.

Affective State	Before	After
Engagement	0.73	0.54
Boredom	0.85	0.81
Confusion	0.58	0.43
Frustration	0.46	0.36

Table 1: Mean closeness to Gold Standard before and after removing bad annotators

Dataset Capture: To capture DAiSEE, the e-learning environment used included a full HD web camera (1920x1080, 30 fps, focal length 3.6mm, 78° field of view) mounted on a computer focusing on student users watching videos. A custom application was created that presented a subject with 2 different videos (20 minutes total in length), one educational and one recreational to capture both focused and relaxed settings, allowing for natural variations in user's engagement levels. To model unconstrained settings, the subjects had the option to scroll through the videos. DAiSEE comprises of 9068 video snippets, each 10 seconds long (This duration was chosen based on inferences observed by Whitehill et al. in [23] that 10-second clips were best suited for annotation.). DAiSEE contains recordings captured from 112 users for recognizing affective states that tend to be associated with instructional videos, namely engagement, boredom, confusion, and frustration (Movatived by [18] in intelligent tutoring systems). Every video snippet in DAiSEE is annotated with four level of labels, namely very low, low, high and very high for each of the affective states, similar to [23]. We followed this labeling strategy to avoid the "neutral" state since early experiments showed that crowd annotators often preferred to choose "neutral" as a state when unsure. Data was collected across varied locations such as dorm rooms, labs and library and 3 different illumination settings (light, dark and neutral); with a male-to-female ratio of ≈ 2:1. In order to avoid the Hawthorne effect [14] [16], also referred to as the observer effect, the subjects were not informed of the recording prior to the session; however, the subject's consent was taken at the session end and the videos were saved only when approval was given.

Subtle affective states such as user engagement are subjective and vary based on the viewer's discretion. Hence, we relied on "wisdom-of-the-crowd" for our annotations in the dataset as in most recent computer vision dataset [3, 11, 22]. We use Figure Eight (previously CrowdFlower) for annotating video snippets in the dataset (each video snippet was annotated by 10 users, inspired from [26, 9]). To improve the quality of annotations, we created a gold-standard using a subset of DAiSEE. Expert psychologists were asked to annotate this subset and only crowd annotators who had high correlation with the gold standard were retained in the finally used labels. We used the popular Dawid-Skene [2] aggregation algorithm on the annotations to obtain the final ground-truth label for each affective state in each video snippet. The improvement in the labels after cleaning the annotations can be viewed in Table 1

Benchmark Evaluation: We used a number of state-of-the-art deep learning models to benchmark DAiSEE (shown in Figure 2) on a 60-20-20 train-test-validation split, which is also shared for benchmarking. We ran experiments using two types of models: *Static* (Frame classification) and *Dynamic* (video classification), and the results of these experiments can be seen in Figure 2. The models chosen were

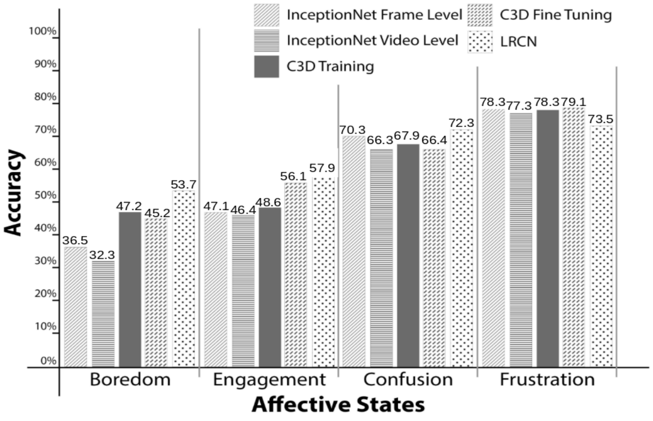

Figure 2: Benchmark results on DAiSEE (top-1 accuracy averaged over three trials). We used [20, 21, 5] as benchmarking models.

A comparison of DAiSEE against other facial emotion recognition datasets is shown in Table 2. DAiSEE is uniquely positioned due to the following unique features:

- It is the first publicly available dataset for studying user

Database	Database Information	# of Subjects	Condition	Affect Modelling
Oulu-CASIA-NIR-VIS [25]	2,880 videos, Frontal view	80	Posed	7 standard emotions
AR Face Database [13]	4,004 images, Frontal view	154	Posed	7 standard emotions
CK+ [12]	593 images, Frontal & 30° images	123	Controlled, Posed	30 AUs, 7 emotions
MultiPie [8]	750,000 images	337	Controlled, Posed	7 emotions
MMI [17]	2900 videos, Frontal and side views	25	Controlled, Posed	31 AUs, 6 Basic Expressions
AFEW [4]	1832 videos	330	Wild	7 emotions
Aff-Wild [24]	500 YouTube videos	500	Wild	Valence and arousal (continuous)
EmotioNet [1]	1,000,000 images with landmarks	100,000	Wild	12 AUs, 23 AU based emotion categories
AffectNet [15]	1,000,000 images with landmarks	450,000	Wild	8 emotions, Valence, Arousal
Belfast Database [19]	1,400 videos	256	Wild	Disgust, Fear, Anger, Surprise, Frustration
DAiSEE (This Work)	**9,068 videos**	**112**	**Wild**	**Engagement, Boredom, Confusion, Frustration**

Table 2: Summary and Characteristics of Several Datasets in Affect Recognition

engagement and related affective states, providing a holistic measure of human-computer interaction.

- The subjects are captured "in-the-wild", thus emulating real-life, natural and spontaneous emotions.
- Engagement and related affective states are generally prolonged in nature and cannot be captured with a single frame. To facilitate their study, we are releasing video snippets and annotations for all video snippets.

Affective State	Accuracy
Engagement	51.07%
Boredom	35.89%
Confusion	57.45%
Frustration	73.09%

Table 3: Benchmarking EmotionNet on DAiSEE

Analysis: To further analyze DAiSEE, we used EmotionNet [1], a CNN pre-trained on CK+ [12] and KDEF [7] and used for traditional emotion recognition as a benchmarking model. We then fine-tuned the model on DAiSEE and the results are summarized in Table 3.

DAiSEE presents challenges in benchmarking and study of affective states (see Figure 3 and Figure 4) because of the natural state in which the videos are captured contributing to factors such as low illumination, lack of frontal face posture or facial occlusion, rapid changes in affective states for some subjects, and the correlation between the different affective states - for example, the complementary nature of boredom and engagement or the correlation between confusion and frustration. With its size, robustness and variety of affective states considered, we believe that DAiSEE can be extended for applications in: (i) e-advertising - for more personalized advertisements; (ii) e-shopping - for personalized user recommendations; (iii) e-healthcare - for detecting early signs of disorders such as ADHD/autism, and (iv) autonomous vehicles - capturing driver's engagement levels or predicting driver's future actions based on confusion/frustration levels; but not be limited to them.

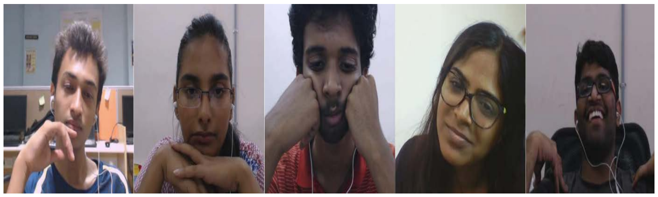

Figure 3: Variety in DAiSEE: All shown frames are assigned *Very High Engagement* label but express it through various poses. Illumination changes are visible from light to dark in the above images, as we move from left to right.

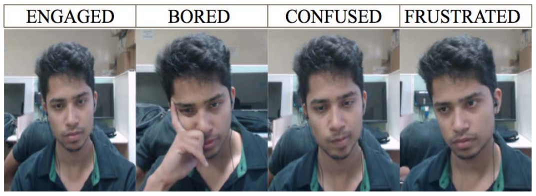

Figure 4: Range of affective states observed in a single 10-second video snippet, making it difficult for classification

We believe that DAiSEE will provide the research community with challenges in context-based inference and development of suitable machine learning methods for related tasks, thus providing a springboard for further research. We hope that DAiSEE especially assists researchers in the domain of e-learning, creating better and more responsive systems to help improve human-computer interaction. A more comprehensive version of the paper is available at arXiv and the dataset is available for download here.

References

[1] C. F. Benitez-Quiroz, R. Srinivasan, A. M. Martinez, et al. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *CVPR*, pages 5562–5570, 2016. 3

[2] A. P. Dawid and A. M. Skene. Maximum likelihood estimation of observer error-rates using the em algorithm. *Applied statistics*, pages 20–28, 1979. 2

[3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009. 1, 2

[4] A. Dhall, R. Goecke, J. Joshi, M. Wagner, and T. Gedeon. Emotion recognition in the wild challenge 2013. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, pages 509–516. ACM, 2013. 3

[5] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*, 2015. 2

[6] M. Everingham, L. V. Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge, 2010. 1

[7] E. Goeleven, R. De Raedt, L. Leyman, and B. Verschuere. The karolinska directed emotional faces: a validation study. *Cognition and emotion*, 22(6):1094–1118, 2008. 3

[8] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-pie. *Image and Vision Computing*, 28(5):807–813, 2010. 3

[9] H. Han, C. Otto, X. Liu, and A. K. Jain. Demographic estimation from face images: Human vs. machine performance. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 37(6):1148–1161, 2015. 2

[10] J. Hernandez, Z. Liu, G. Hulten, D. DeBarr, K. Krum, and Z. Zhang. Measuring the engagement level of tv viewers. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, pages 1–7. IEEE, 2013. 1

[11] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014*, pages 740–755. Springer, 2014. 1, 2

[12] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pages 94–101. IEEE, 2010. 3

[13] A. M. Martinez. The ar face database. *CVC Technical Report24*, 1998. 3

[14] R. McCarney, J. Warner, S. Iliffe, R. Van Haselen, M. Griffin, and P. Fisher. The hawthorne effect: a randomised, controlled trial. *BMC medical research methodology*, 7(1):1, 2007. 2

[15] A. Mollahosseini, B. Hasani, and M. H. Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *arXiv preprint arXiv:1708.03985*, 2017. 3

[16] T. Monahan and J. A. Fisher. Benefits of observer effects: lessons from the field. *Qualitative Research*, 10(3):357–376, 2010. 2

[17] M. Pantic, M. Valstar, R. Rademaker, and L. Maat. Web-based database for facial expression analysis. In *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, pages 5–pp. IEEE, 2005. 3

[18] R. Rajendran. *Enriching the Student Model in an Intelligent Tutoring System*. PhD thesis, The IITB-Monash Research Academy, 2014. 2

[19] I. Sneddon, M. McRorie, G. McKeown, and J. Hanratty. The belfast induced natural emotion database. *IEEE Transactions on Affective Computing*, 3(1):32–41, 2012. 3

[20] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *Computer Vision and Pattern Recognition, 2016. CVPR 2016*. IEEE, 2016. 2

[21] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. ICCV, 2015. 2

[22] G. Van Horn, S. Branson, R. Farrell, S. Haber, J. Barry, P. Ipeirotis, P. Perona, and S. Belongie. Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 595–604, 2015. 2

[23] J. Whitehill, Z. Serpell, Y.-C. Lin, A. Foster, and J. R. Movellan. The faces of engagement: Automatic recognition of student engagementfrom facial expressions. *Affective Computing, IEEE Transactions on*, 5(1):86–98, 2014. 1, 2

[24] S. Zafeiriou, A. Papaioannou, I. Kotsia, M. Nicolaou, and G. Zhao. Facial affect“in-the-wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 36–47, 2016. 3

[25] G. Zhao, X. Huang, M. Taini, S. Z. Li, and M. PietikäInen. Facial expression recognition from near-infrared videos. *Image and Vision Computing*, 29(9):607–619, 2011. 3

[26] D. Zhou, S. Basu, Y. Mao, and J. C. Platt. Learning from the wisdom of crowds by minimax entropy. In *Advances in Neural Information Processing Systems*, pages 2195–2203, 2012. 2