This CVPR workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

TikTok for good: Creating a diverse emotion expression database

Saimourya Surabhi Bhavik Shah Peter Washington Onur Cezmi Mutlu
Emilie Leblanc Prathamesh Mohite Arman Husic Aaron Kline
Kaitlyn Dunlap Maya McNealis Bennett Liu Nick Deveaux Essam Sleiman
Dennis P. Wall
Department of Pediatrics (Systems Medicine), Department of Biomedical Data Science, and
Department of Psychiatry and Behavioral Sciences, Stanford University
mourya@stanford.edu, shah.bhavik627@gmail.com, {peter100, cezmi} @stanford.edu, emilie.eam.leblanc@gmail.com
{ mohitep1, ahusic, akline, kdunlap2, mayamcne, bennettl, ndeveau, sleiessa, dpwall}@stanford.edu

Abstract

Facial expression recognition (FER) is a critical computer vision task for a variety of applications. Despite the widespread use of FER, there is a dearth of racially diverse facial emotion datasets which are enriched for children, teens, and adults. To bridge this gap, we have built a diverse expression recognition database using publicly available videos from TikTok, a video-focused social networking service. We describe the construction of the TikTok Facial expression recognition (FER) database. The dataset is extracted from 6428 videos scraped from TikTok. The videos consist of 9392 distinct individuals and labels for 15 emotion-related prompts. We were able to achieve a F1 score 0.78 for Ekman emotions on expression classification using transfer learning. We hope that the scale and diversity of the TikTokFER dataset will be of use to affective computing practitioners.

1. Introduction

Emotion recognition research has come a long way since Dr. Paul Ekman's [14] work on universal emotions, identifying the following "universal" emotions: Happiness, Sadness, Surprise, Anger, Disgust, Fear, and Contempt. A large body of Facial Expression Recognition (FER) research focuses on building algorithms to automatically identify emotions, and especially Ekman emotions, from modalities such as voice [6], text, faces, and video clips [48]. They rely on the availability of large datasets enriched with information such as categorical emotion labels, facial landmarks like the position of the nose or eyes, Facial Action Coding System (FACS) [13] action units detecting subtle changes in facial features, or continuous dimensions of valence, arousal, and dominance. Many of these datasets have been created and made publicly available for research purposes. The Cohn-Kanade dataset (CK) [30], one of the most used datasets in the field, consists of frontal and side views of 182 adults (18-50 years old, 69% female and 31% male, 81% European American) displaying 23 facial expressions, some of which are FACS coded and emotion-labeled by annotators. The CK dataset was then enhanced with 27% more subjects, revised emotion expression labels, and non-posed smiles seen in the CK+ dataset [43]. Other efforts, such as the Multimedia Understanding Group (MUG) dataset, [1] also attempt to gather both posed and naturalistic expressions. Unlike structured in-lab data collection efforts, AffectNet [46] gathered over 1 million facial images extracted from 3 major search engines using 1250 emotion-related keywords, also automatically increasing the number of distinct subjects. AFF-Wild2 [38], AM-FED [44], GIFGIF+ [8], EMOTIW [12] and the OMGEmotion [3] datasets also are compilations of real-world "in the wild" video clips or gifs. Addressing the racial imbalance in the datasets, some have focused their efforts on collecting data from specific ethnicities, such as JAFFE [29] and ISED [17].

FER algorithms are trained, tested, and validated on these available datasets. As described in detail by Ko [32], there are two main approaches for FER algorithms: using handcrafted features or generating features automatically through neural network outputs. The first approach relies on the extraction of facial components or landmarks in images, such as FACS action units and their spatial and temporal changes from videos. An expression classifier, such as a support vector machine or random forest, is then trained on these facial features. The second approach to FER relies on deep learning, extracting optimal features directly from the image or video data using convolutional neural networks (CNN) or a combination of CNN and RNN (recurrent neural networks) for temporal features of consecutive frames.

These expression recognition algorithms have many potential applications to not only improve the quality of human-computer interactions (e.g. through enhanced security cameras, online courses detecting frustration, or advanced driver assistance systems) but also to assist humans in their interactions with each other. Cultural differences, certain neurodevelopmental conditions such as Autism Spectrum Disorder, or blindness can affect our ability to understand the facial expressions of others. Initiatives like that of Buimer et al. [7] and SuperPowerGlass [10, 16, 31, 50, 51, 57], a wearable aid for the at-home therapy of children with autism, leverage video-based emotion recognition algorithms for clinical purposes and have had promising results. [9, 15] However, FER algorithms tend to suffer from the domain shift phenomena and therefore remain limited to datasets they are trained on. The performance of face and emotion recognition algorithms degrades when confronted with different ethnicities and age groups. To exemplify, Zhao et al. [60] noticed much higher accuracy on Finnish people in their dataset than on the Chinese subjects. Although algorithmic strategies are being developed to measure and adjust for these biases, building more diverse datasets remains a top priority.

To address the need for more diverse, balanced FER data, we leveraged TikTok challenges. We use publicly available recordings of emotion acting challenges to build a FER dataset containing racial diversity, tailored towards teens, and young adults. The rest of the paper is organized as follows: We describe the Tiktok-FER data set in Sec. 2. In Sec. 3 we describe how the data set is constructed by leveraging Amazon Mechanical Turk. Finally, in Sec. 4, we present analysis and a few simple experiments on the TikTokFER. Our goal is to show that the TikTokFER can serve as a useful resource for FER applications.

2. Properties of TikTokFER

Social media and TikTok in particular have come under scrutiny in the last few years because of their lack of member data protection, generation of potential national security concerns, and their influence on the radicalization of the US political landscape. These geopolitical concerns led to the ban of the TikTok app in India in June 2020, its prohibition on all US government-issued devices by the US Navy and the US Army in December 2019, and calls to introduce US-based ownership of its parent company ByteDance. Nevertheless, since its launch in September 2016, TikTok's user base has grown considerably and has been installed on devices over 3 billion times worldwide. It passed the one billion milestone in February 2019, and it reached three billion in mid-2021, with 1 billion monthly active users as of January 2022. The TikTok app, which lets users view 15 second clips and publish their short videos leverages viral marketing methods such as challenges to engage a highly active community. Relying primarily on teenagers and young adults (41% of its users are between the ages of 16 and 24), TikTok has managed to attract a wide range of users from over 155 countries and is available in 35+ languages [58].

Figure 1. Fitzpatrick Scale

TikTok's huge and diverse audience is actively leveraged by brands through targeted marketing and influencer sponsoring. Political parties and governments have also started using this medium to communicate political [45] and public health messages, for instance during the COVID-19 crisis [4]. Educational initiatives have also shown promising results in engaging audiences through TikTok. "The Chemistry Collective," for example, was able to increase viewers' interest in chemistry by 82.7% [18] with their 16 educational TikTok clips. These initiatives illustrate the potential for TikTok to be used for the common good.

TikTok videos are typically short, fun recordings often involving music, dancing, or comedy. The vast majority of these videos are shot using a front facing mobile phone camera, where the recorder's face is in clear view. New challenges are continuously widely adopted across a diverse population, such as particular dances or skits. The nature of these videos, being first-person shot, are rich with changing, human facial expressions. TikTok is a great resource to access large amounts of diverse facial expressions for our dataset.

Scale TikTok FER dataset contains a total of 6482 videos from 9392 distinct individuals, labeled for 15 emotion-related prompts. We created a diverse and emotion-enriched subdataset of 1207 teens, containing 232 males and 975 females. To our knowledge, this is the largest and most diverse FER dataset targeting adolescent subjects and the first leveraging TikTok viral challenges for data collection.

Diversity An estimation of the age and gender of the subjects shows that 21.2% of them are male, 13.1% are teens (<18yo) and 65% are young adults (<30yo). We have used the Fitzpatrick scale in Figure 1 as a proxy for race of which 42.9% skin type 1, 9.2% skin type 2, 36.6% skin type 3, 2.4% skin type 4, and 8.9% skin type 6.

3. Construction of TikTokFER

TikTok has already been used in non-FER contexts for data analyses and to build training datasets for algorithms. Bandy et al [2] utilized social media to analyze the impact of call-to-action videos from more than 600 TikTok users and compare the visibility (i.e. play count) of these videos with other videos published by the same users. Tao et al [49] tested their algorithms on a dataset created by TikTok for a video-recommendation competition composed of 76,085 videos with their textual captions from 36,656 users, and the users' likes and interactions. TikTok video comments have also been leveraged for sentiment analysis [23] and an initiative by Jiang et al [22] uses a dataset of 500,000 short videos, some extracted from TikTok, for near-duplicate identification.

3.1. Collecting the data

In this work, we leverage TikTok's diverse and young user base to build a dataset enhanced for facial emotion recognition. To do so, we have identified two viral emotion-based challenges: Face challenge[1] by Zephyr (@zephyrean) posted on 2019-8-22 and Best Emoji Face[2] by Yangzom (@tseringyangzom17) posted on 2018-6-17. Both challenges rely on audio composed of a succession of emotion-related prompts. Each challenge participant uses the same audio as the original video that started the challenge and mimics the emotions or emojis when the audio prompts them to do so. As seen in figure 3, on February 4th,2021 TikTok's website indicated that Zephyr's Face challenge sound had been used in 285.1K videos and that Yangzom's Best Emoji Face sound had been used in 331.6K videos.

The Best Emoji Face challenge prompts the user to mimic the following emojis through visual and sound cues (e.g. *arrrr* sound when the user is prompted to mimic the angry-face emoji): 😘, 😛, 😬,😯, 🙄, 😖, 😠,😟,😩, 😭. For convenience, we translate these emojis to the following: "kissing emoji", "tongue pulled emoji", "clenched teeth emoji", "surprised emoji", "eye roll emoji", "cringe emoji", "angry emoji", "sad emoji", "very sad emoji" and "crying emoji". However, we are aware that cultural differences and context may change these emojis' interpretations.

The Face challenge is more straightforward and directly asks the user to mimic a "grinning face with normal eyes", a "grinning face with clenched teeth", a "slightly smiling face", a "winking face", a "disappointed face", a "thinking face", a "nauseated face", an "angry face", a "crying face" and a "clown face".

All the videos and data collected is publicly available through the TikTok website. No TikTok account is necessary for the collection of this data as the videos are accessible outside of the app, from the TikTok website. As we collected publicly available data, collection of informed consent was waived by Stanford University IRB.

3.2. Data processing pipeline

To create the TiktokFER database we built a data processing pipeline[3]. The data extraction process is done in four steps figure 2:

- Identifying the challenges.
- Extracting the raw videos.
- Splitting the videos into separate prompts. Extracting frames from prompt-specific videos.
- Detecting and extracting each face from each frame.

3.2.1 Extracting the raw videos

To extract the raw videos from the TikTok website, we used the scraper developed by GitHub user *drawrowfly*[4]. Using the scraper, we have gathered enhanced video information such as the epoch timestamp of the post, the user inputted text captions and hashtags associated, and the number of likes, shares, and, plays. Additionally, author level information was also extracted such as the author's unique identifier, alias, nickname, the number of accounts they are following, the number of users following the author, the total number of likes they have obtained, given, and the total number of videos they posted.

[1] https://www.tiktok.com/music/Face-challenge-6727954870776695557

[2] https://www.tiktok.com/music/Best-Emoji-Face-6568049491721587461

[3] https://github.com/walllab/tiktok_FER

[4] https://github.com/drawrowfly/tiktok-scraper

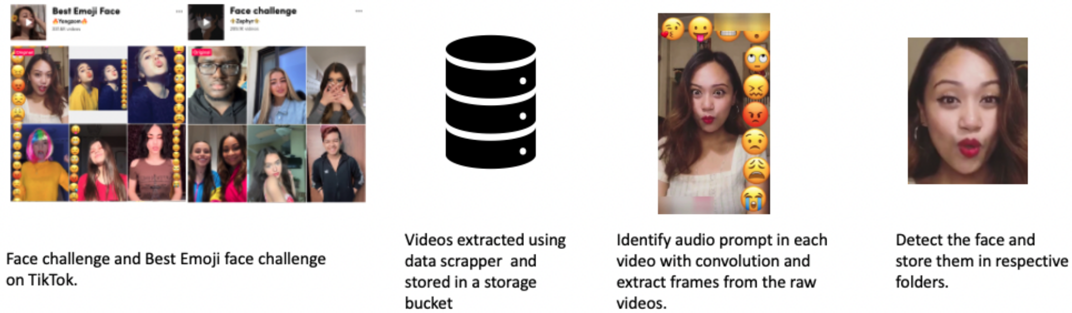
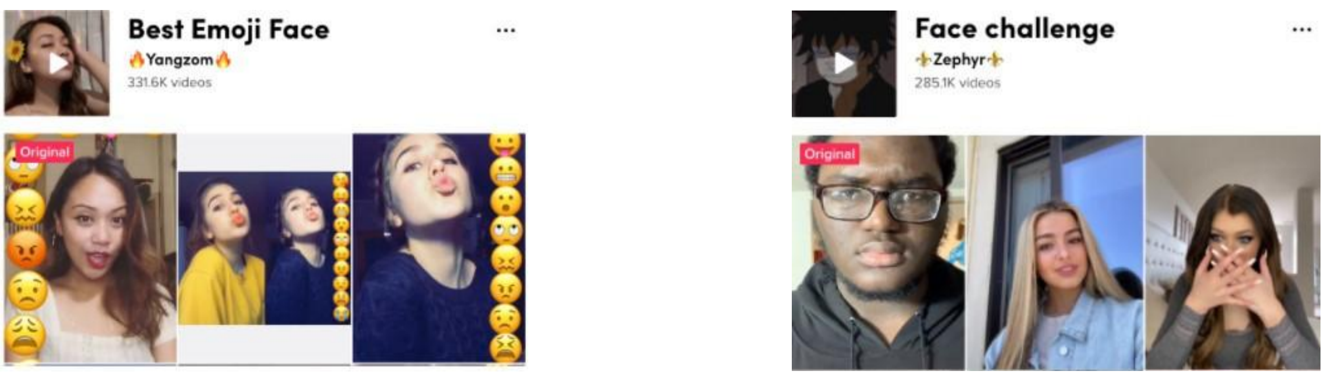


Figure 2. Schema of the data processing pipeline



(a) Best Emoji Face Challenge.

(b) Face Challenge.

Figure 3. TikTok challenges

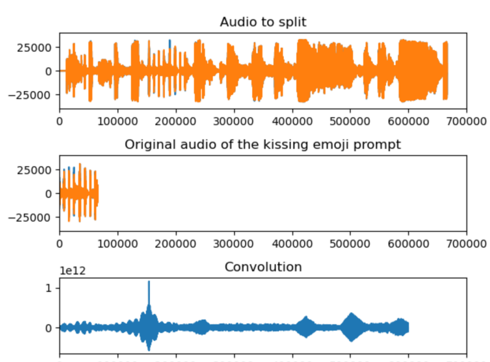


Figure 4. Use of convolution to find the location of an original prompt audio in a video to split (example with the kissing emoji audio prompt)

3.2.2 Splitting the videos into separate prompts.

The videos consist of different facial expressions, and the transition between them is accompanied by different sound prompts that signal the user to mimic the given face/emoji. These signals are the same across all videos in a challenge and we use them to detect the beginning of an expression. We achieve this by first collecting all of the unique prompts in a challenge. We then "search" for these sets of prompts in the audio signals of all videos. This "search" is performed by a signal processing tool convolution. In convolution, we compute the inner product of our target signal (prompt) with a portion of the main signal (audio signal from the entire video) and we perform this computation for the entirety of the main signal and do this by sliding the target over the main signal, one sample at a time. When we achieve a perfect overlap between the prompt and the audio track, the inner product is maximized, hence we find the start time of the prompt. This method is also referred to as a matched filter, and it is an optimal detection algorithm in our conditions. More formally, for a target signal h[t] and main signal x[t] the result of convolution operation at time t∗ is:

$$y[t^*] = \sum_{k=0}^{N-1} h[k]x[t^* + k]$$

where N is the length of the target signal. We calculate y for the all timestamps and search for the timestamp $y[t] = \sum_i h[i]^2$. If such point is found, corresponding index minus half the signal duration gives us the starting point of the prompt. In the following case, the first function is the original video's pre-split audio prompt (the audio signal of "grinning face with clenched teeth" for example) and the second is the audio signal of the video we wish to split. The convolution between the two audio signals is maximal if there is no ambient noise when the original video's audio prompt matches the location of the same audio prompt in the video we wish to split. As seen in figure 4, the kissing emoji audio prompt can be found in the audio to split at timestep 154,399, i.e. when the convolution of both audio signals is maximal.

The frames from each video are extracted once all the videos have been split into prompt sub-clips. We have limited ourselves to 2 frames per second.

3.2.3 Detecting and extracting each face from each frame.

To detect the face in each frame we used the RetinaFace [10] algorithm with the MobileNet-0.25 backbone. If no face is detected in the video taken in landscape mode we rotated the frames 90 degrees to account for the orientation. As provided in Table 1, on manual testing of 200 frames, the algorithm showed a 100% true positive rate and only had 4 false positives out of the total 223 faces in the frames. We have considered all human faces (excluding paintings) as a true positive and any detection of non-human faces (such as emojis) as false positives. The frames contained multiple faces, either because of multiple people in the video or duet style videos, and did not contain any faces . The faces were then aligned based on detected facial key points.

3.3. Crowdsourcing the labels

Crowdsourcing has been proven to be an affordable and effective way to label large amounts of data, including complex social human behaviors [52–56]. To build a gold standard labeled data set, we leverage crowdsourced workers to label a set of images collected after the face extraction and alignment process. We used Amazon Mechanical Turk (AMT) to label the data. In each of our labeling tasks, we present AMT workers with a set of candidate images and examples to help them understand the task. We ask the workers to verify whether each image contains a face in the frame. If there is no face in the frame, no other label should be returned and the rating process ends. If there is a face in the frame, the rater indicates their estimation of the individual's skin color, age, and gender in the second, third and fourth labels. In the final labels, the rater indicates their estimation of the emotion expressed by the individual within two lists of possible emotions: Ekman emotions and beyond Ekman which are complex non-Ekman emotions. The second list of possible emotions (complex non-Ekman emotions) is only presented to the rater if they have not selected any from the first list (Ekman emotions). It is crucial to set up a quality control system to ensure this accuracy. Human users make mistakes and not all users follow the instructions. Users do not always agree with each other, especially for more subtle or confusing images. The solution to these issues is to have multiple users independently label the same image. An image is considered positive only if it gets a convincing majority of the votes.

3.4. Data Validation

Identity resolution is necessary to identify the total number of individuals in the dataset, estimate their age, gender, and skin color, and validate the quality of the emotion labels. We model the dataset as a weighted undirected graph where each node corresponds to an image and connection weights are assigned based on the similarity of the two faces. For measuring similarity, we use convolutional neural networks to extract facial features and calculate the cosine distance between vectors for each individual to build the graph. In particular, we use the Arcface algorithm [11], which is a widely used facial recognition model, for feature extraction. We then threshold the edge weights and only keep the ones exceeding them to reduce the computational complexity of the clustering algorithm. This threshold is selected with a cross-validation process performed on a subset of the dataset. To accurately count the number of individuals in the dataset, we use the Chinese whispers (CW) [5] algorithm. The identity resolution is done in three steps (Figure 5) :

- Running the CW algorithm to identify the distinct individuals within the same video.

- Aggregating facial features per individual per video and using cosine similarity to identify identical individuals in other distinct videos.

- Using cosine similarity with a higher threshold to remove exact duplicates across videos.

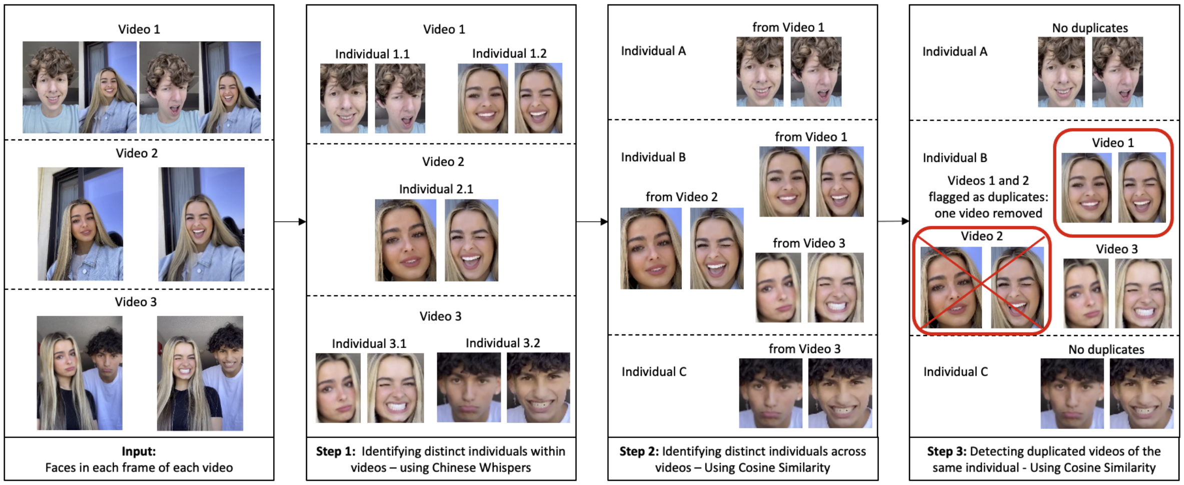


Figure 5. Identity resolution process on the TikTok dataset

#	Best Emoji Face	Face Challenge	Total
frames used	100	100	200
faces in the frames	110	113	223
faces correctly identified	110	113	223
faces missed	0	0	0
faces incorrectly identified	3	1	4

Table 1. Performance of the RetinaFace algorithm on randomly sampled frames

To validate this process we collected 100 faces, split evenly between both challenges. We manually checked these 100 videos and noticed that there were: 50 distinct individuals, 25 of them were present in two distinct videos (but not duplicated videos) and 25 of them were present as duplicates. The aim is to accurately detect clusters corresponding to the same individual and clusters corresponding to exact duplicates. Since we know the ground truth of the labels, we use the adjusted Rand [21] index for clustering evaluation. The adjusted Rand index is a consensus measure, measuring the similarity between two assignments, ignoring permutations, and adjusting for the chance. As seen in Table 2, the detection of unique individuals yielded an adjusted Rand index of 0.9

4. Results and Analysis

In this section, we present an analysis of the TikTokFER dataset and provide baseline performances on certain discriminative tasks.

4.1. Analysis

Face Challenge and Best Emoji Face Challenge resulted in a total of 6,428 videos (2,447 and 3,981, respectively). After the data preprocessing and identity resolution, there are 92,389 distinct faces in the TikTokFER dataset.(Table 4) The TikTokFEER data set consists of 15 emotions: Angry, Clenched-teeth, Clown-face, Cringe,Cry, Disappointed, Disgust, Eye-roll, Kiss, Nauseated-face, Sad, Surprise, Smiling, Thinking face and Winking face. 5 of these emotions, namely: Angry, Disgust, Sad, Surprise, Smiling are Ekman emotions and the rest are complex beyond Ekman emotions. Figure 6 & 7 show the distribution of expressions in the dataset.

4.2. Results

To perform an initial analysis to gain an idea of the emotion-classification power of TikTok, we first develop a model using ResNet-50 [19], pre-trained on ImageNet. We replace the last linear to match our output size of 15 and allow all the layers to be trained. We use Adam optimizer with learning rate 0.001, default parameters and cosine annealing with warm restarts [42]. We utilize label smoothing [47] to avoid overconfidence. We implemented this model and its training procedures in PyTorch and performed training on a single NVIDIA Tesla P100 GPU.

We tested the performance of our mode on popu-

Challenge	Number of Input Faces	Number of Distinct Individuals	Adjusted Rand Index
Face challenge	100 (including 25 duplicates and 25 faces from same individuals but a different timestamp)	50	0.89
Best Emoji Face	100 (including 25 duplicates and 25 faces from same individuals but a different timestamp)	50	0.91
Both	200 (including 50 duplicates and 50 faces from same individuals but a different timestamp)	100	0.9

Table 2. Performance of the Chinese whispers algorithm

lar FER benchmarks for 5 Ekman emotions present in our dataset and results are provided in Table 3. The analysis on classification accuracy shows that TikTok-FER dataset can provide significant predictive power for expression classifications tasks.

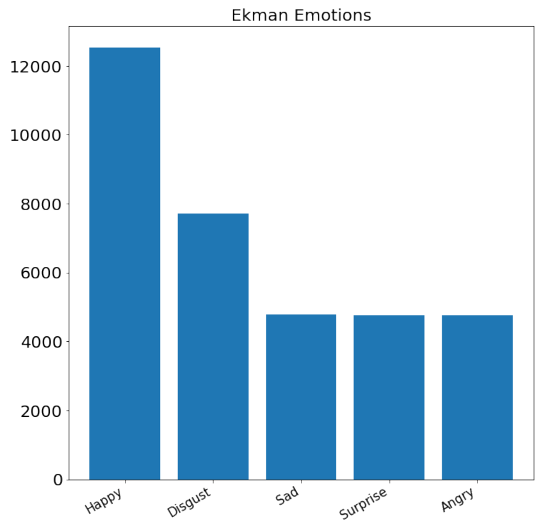


Figure 6. Distribution of Ekman emotions in the dataset

5. Discussion and Future Work

We anticipate that TikTokFER will become a useful resource for a broad range of FER-related research. Most directly, TikTokFER can become a standard training resource for FER. Most of today's FER recognition algorithms have focused on smaller data sets that are not diverse. TikTokFER, on the other hand, contains a large number of images for Ekman emotion classes. One interesting research direction could be to study the evolution of emotion across various age, gender, and skin color groups. (2) Using Tiktok FER dataset to build a personalized FER algorithm using meta-learning methods. Current emotion classifiers fail on pediatric populations [20, 24]. Using FER models which are tuned for pediatric populations can improve digital interventions for children with affective conditions such as autism [25–28].

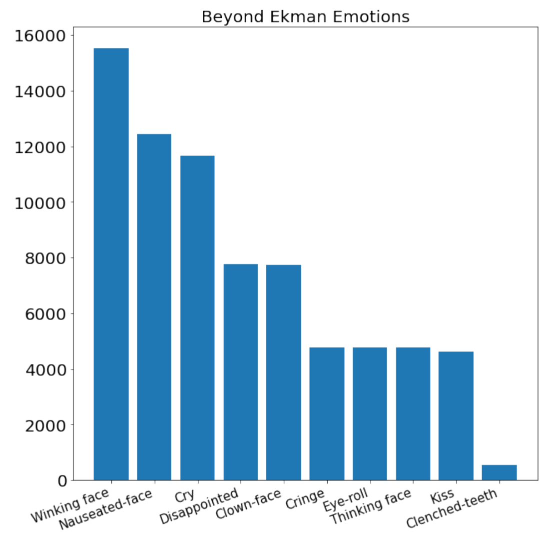


Figure 7. Distribution of beyond Ekman emotions in the dataset

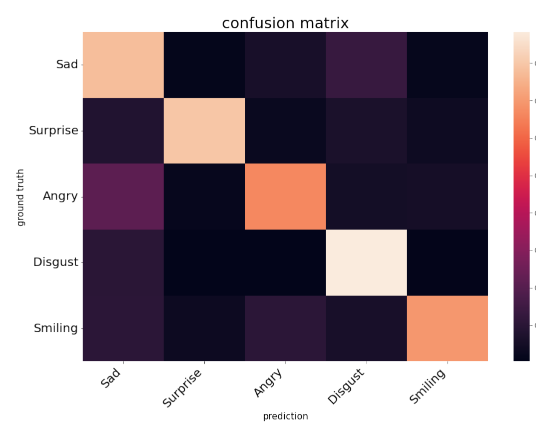


Figure 8. Confusion matrix for Ekman emotions on validation set

Some potential drawbacks of the dataset are that the emotions are not natural or genuine, as they are acted out. In the challenge, the TikTokers try to pose based on the emoji but emoticons are not equal to emotions. Still, it is possible to glean useful information about a diverse number of emotions with this data.

Dataset	F1(macro)	Accuracy
Tiktok-validation (all 15 expressions)	0.47	0.52
Tiktok-validation	0.78	0.79
Affwild2-validation [33–40, 59]	0.20	0.29
CK+ [43]	0.65	0.73
CAFE [41]	0.50	0.499

Table 3. Predictive performances on 5 Ekman emotions

Challenge	Videos	Frames	Distinct Face	Distinct Individuals
Face challenge	3,981	74,922	37,181	3,571
Best Emoji Face	2,447	139,578	54,567	5,875
Total	6,482	214,500	92,389	9,392

Table 4. TikTokFER dataset

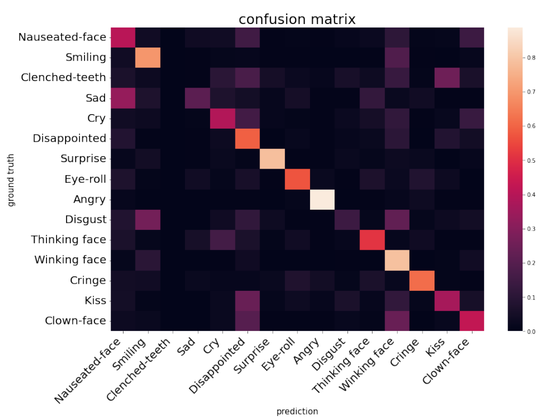


Figure 9. Confusion matrix for all emotions on validation set

6. Acknowledgements

We would like to thank all members of the community who provided valuable feedback throughout the process of defining and collecting the dataset. The work was supported in part by funds to DPW from the National Institutes of Health (1R01EB02502501,1R01LM013364-01, 1R21HD09150001, 1R01LM013083), the National Science Foundation (Award 2014232), The Hartwell Foundation, Bill and Melinda Gates Foundation, Coulter Foundation, Lucile Packard Foundation, Auxiliaries Endowment, the ISDB Transform Fund, the Weston Havens Foundation, and program grants from Stanford's Human Centered Artificial Intelligence Program, Precision Health and Integrated Diagnostics Center, Beckman Center, Bio-X Center, Predictives and Diagnostics Accelerator, Spectrum, Spark Program in Translational Research, MediaX, and from the Wu Tsai Neurosciences Institute's Neuroscience:Translate Program. We also acknowledge generous support from David Orr, Imma Calvo, Bobby Dekesyer and Peter Sullivan. PW would like to acknowledge support from Mr. Schroeder and the Stanford Interdisciplinary Graduate Fellowship (SIGF) as the Schroeder Family Goldman Sachs Graduate Fellow.

References

[1] Niki Aifanti, Christos Papachristou, and Anastasios Delopoulos. The mug facial expression database. In *11th International Workshop on Image Analysis for Multimedia Interactive Services WIAMIS 10*, pages 1–4. IEEE, 2010. 1

[2] Jack Bandy and Nicholas Diakopoulos. # tulsaflop: A case study of algorithmically-influenced collective action on tiktok. *arXiv preprint arXiv:2012.07716*, 2020. 3

[3] 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

[4] Corey H Basch, Grace C Hillyer, and Christie Jaime. Covid-19 on tiktok: harnessing an emerging social media platform to convey important public health messages. *International journal of adolescent medicine and health*, 2020. 2

[5] Chris Biemann. Chinese whispers-an efficient graph clustering algorithm and its application to natural language processing problems. In *Proceedings of TextGraphs: the First Workshop on Graph Based Methods for Natural Language Processing*, pages 73–80, 2006. 5

[6] Damian Borth, Rongrong Ji, Tao Chen, Thomas Breuel, and Shih-Fu Chang. Large-scale visual sen-

timent ontology and detectors using adjective noun pairs. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 223–232, 2013. 1

[7] Hendrik P Buimer, Marian Bittner, Tjerk Kostelijk, Thea M Van Der Geest, Abdellatif Nemri, Richard JA Van Wezel, and Yan Zhao. Conveying facial expressions to blind and visually impaired persons through a wearable vibrotactile device. *PloS one*, 13(3):e0194737, 2018. 2

[8] Weixuan Chen, Ognjen Oggi Rudovic, and Rosalind W Picard. Gifgif+: Collecting emotional animated gifs with clustered multi-task learning. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 510–517. IEEE, 2017. 1

[9] Jena Daniels, Nick Haber, Catalin Voss, Jessey Schwartz, Serena Tamura, Azar Fazel, Aaron Kline, Peter Washington, Jennifer Phillips, Terry Winograd, et al. Feasibility testing of a wearable behavioral aid for social learning in children with autism. *Applied clinical informatics*, 9(01):129–140, 2018. 2

[10] Jena Daniels, Jessey N Schwartz, Catalin Voss, Nick Haber, Azar Fazel, Aaron Kline, Peter Washington, Carl Feinstein, Terry Winograd, and Dennis P Wall. Exploratory study examining the at-home feasibility of a wearable tool for social-affective learning in children with autism. *NPJ digital medicine*, 1(1):1–10, 2018. 2, 5

[11] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019. 5

[12] Abhinav Dhall, Roland Goecke, Shreya Ghosh, Jyoti Joshi, Jesse Hoey, and Tom Gedeon. From individual to group-level emotion recognition: Emotiw 5.0. In *Proceedings of the 19th ACM international conference on multimodal interaction*, pages 524–528, 2017. 1

[13] Paul Ekman and Wallace V Friesen. Facial action coding system. *Environmental Psychology & Nonverbal Behavior*, 1978. 1

[14] Paul Ekman, Wallace V Friesen, Maureen O'sullivan, Anthony Chan, Irene Diacoyanni-Tarlatzis, Karl Heider, Rainer Krause, William Ayhan LeCompte, Tom Pitcairn, Pio E Ricci-Bitti, et al. Universals and cultural differences in the judgments of facial expressions of emotion. *Journal of personality and social psychology*, 53(4):712, 1987. 1

[15] Nick Haber, Catalin Voss, Jena Daniels, Peter Washington, Azar Fazel, Aaron Kline, Titas De, Terry Winograd, Carl Feinstein, and Dennis P Wall. A wearable social interaction aid for children with autism. *arXiv preprint arXiv:2004.14281*, 2020. 2

[16] Nick Haber, Catalin Voss, and Dennis Wall. Making emotions transparent: Google glass helps autistic kids understand facial expressions through augmented-reaiity therapy. *IEEE Spectrum*, 57(4):46–52, 2020. 2

[17] SL Happy, Priyadarshi Patnaik, Aurobinda Routray, and Rajlakshmi Guha. The indian spontaneous expression database for emotion recognition. *IEEE Transactions on Affective Computing*, 8(1):131–142, 2015. 1

[18] Clare Hayes, Katherine Stott, Katie J Lamb, and Glenn A Hurst. "making every second count": utilizing tiktok and systems thinking to facilitate scientific public engagement and contextualization of chemistry at home, 2020. 2

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[20] Cathy Hou, Haik Kalantarian, Peter Washington, Kaiti Dunlap, and Dennis P Wall. Leveraging video data from a digital smartphone autism therapy to train an emotion detection classifier. *medRxiv*, 2021. 7

[21] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of classification*, 2(1):193–218, 1985. 6

[22] Qing-Yuan Jiang, Yi He, Gen Li, Jian Lin, Lei Li, and Wu-Jun Li. Svd: A large-scale short video dataset for near-duplicate video retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5281–5289, 2019. 3

[23] Hao Jiaxiang. Building domain specific lexicon based on tiktok comment dataset. *arXiv preprint arXiv:2012.08773*, 2020. 3

[24] Haik Kalantarian, Khaled Jedoui, Kaitlyn Dunlap, Jessey Schwartz, Peter Washington, Arman Husic, Qandeel Tariq, Michael Ning, Aaron Kline, and Dennis Paul Wall. The performance of emotion classifiers for children with parent-reported autism: quantitative feasibility study. *JMIR mental health*, 7(4):e13174, 2020. 7

[25] Haik Kalantarian, Khaled Jedoui, Peter Washington, Qandeel Tariq, Kaiti Dunlap, Jessey Schwartz, and Dennis P Wall. Labeling images with facial emotion and the potential for pediatric healthcare. *Artificial intelligence in medicine*, 98:77–86, 2019. 7

[26] Haik Kalantarian, Khaled Jedoui, Peter Washington, and Dennis P Wall. A mobile game for automatic emotion-labeling of images. *IEEE transactions on games*, 12(2):213–218, 2018. 7

[27] Haik Kalantarian, Peter Washington, Jessey Schwartz, Jena Daniels, Nick Haber, and Dennis Wall. A gamified mobile system for crowdsourcing video for autism research. In *2018 IEEE international conference on healthcare informatics (ICHI)*, pages 350–352. IEEE, 2018. 7

[28] Haik Kalantarian, Peter Washington, Jessey Schwartz, Jena Daniels, Nick Haber, and Dennis P Wall. Guess what? *Journal of healthcare informatics research*, 3(1):43–66, 2019. 7

[29] Miyuki Kamachi, Michael Lyons, and Jiro Gyoba. The japanese female facial expression (jaffe) database. *Availble: http://www. kasrl. org/jaffe. html*, 01 1997. 1

[30] Takeo Kanade, Jeffrey F Cohn, and Yingli Tian. Comprehensive database for facial expression analysis. In *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580)*, pages 46–53. IEEE, 2000. 1

[31] Michael Ning Arman Husic Peter Washington Catalin Voss Kaitlyn L. Dunlap Yordan Penev Emilie Leblanc Nick Haber Kline, Aaron and Dennis Wall. "superpower glass: An augmented reality intervention for improving social deficits in children with autism spectrum disorder.". *INSAR*, 2020. 2

[32] Byoung Chul Ko. A brief review of facial emotion recognition based on visual information. *sensors*, 18(2):401, 2018. 1

[33] Dimitrios Kollias. Abaw: Valence-arousal estimation, expression recognition, action unit detection & multi-task learning challenges. *arXiv preprint arXiv:2202.10659*, 2022. 8

[34] D Kollias, A Schulc, E Hajiyev, and S Zafeiriou. Analysing affective behavior in the first abaw 2020 competition. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)(FG)*, pages 794–800. 8

[35] Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou. Face behavior a la carte: Expressions, affect and action units in a single network. *arXiv preprint arXiv:1910.11111*, 2019. 8

[36] Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou. Distribution matching for heterogeneous multi-task learning: a large-scale face study. *arXiv preprint arXiv:2105.03790*, 2021. 8

[37] Dimitrios Kollias, Panagiotis Tzirakis, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, Björn Schuller, Irene Kotsia, and Stefanos Zafeiriou. Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond. *International Journal of Computer Vision*, pages 1–23, 2019. 8

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

[39] Dimitrios Kollias and Stefanos Zafeiriou. Affect analysis in-the-wild: Valence-arousal, expressions, action units and a unified framework. *arXiv preprint arXiv:2103.15792*, 2021. 8

[40] Dimitrios Kollias and Stefanos Zafeiriou. Analysing affective behavior in the second abaw2 competition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3652–3660, 2021. 8

[41] Vanessa LoBue and Cat Thrasher. The child affective facial expression (cafe) set: Validity and reliability from untrained adults. *Frontiers in psychology*, 5:1532, 2015. 8

[42] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 6

[43] Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *2010 ieee computer society conference on computer vision and pattern recognition-workshops*, pages 94–101. IEEE, 2010. 1, 8

[44] Daniel McDuff, Rana Kaliouby, Thibaud Senechal, May Amr, Jeffrey Cohn, and Rosalind Picard. Affectiva-mit facial expression dataset (am-fed): Naturalistic and spontaneous facial expressions collected. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 881–888, 2013. 1

[45] Juan Carlos Medina Serrano, Orestis Papakyriakopoulos, and Simon Hegelich. Dancing to the partisan beat: a first analysis of political communication on tiktok. In *12th ACM Conference on web science*, pages 257–266, 2020. 2

[46] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2017. 1

[47] Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. When does label smoothing help? *Advances in neural information processing systems*, 32, 2019. 6

[48] Ruthger Righart and Beatrice De Gelder. Rapid influence of emotional scenes on encoding of facial expressions: an erp study. *Social cognitive and affective neuroscience*, 3(3):270–278, 2008. 1

[49] Zhulin Tao, Yinwei Wei, Xiang Wang, Xiangnan He, Xianglin Huang, and Tat-Seng Chua. Mgat: Multimodal graph attention network for recommendation. *Information Processing & Management*, 57(5):102277, 2020. 3

[50] Catalin Voss, Jessey Schwartz, Jena Daniels, Aaron Kline, Nick Haber, Peter Washington, Qandeel Tariq, Thomas N Robinson, Manisha Desai, Jennifer M Phillips, et al. Effect of wearable digital intervention for improving socialization in children with autism spectrum disorder: a randomized clinical trial. *JAMA pediatrics*, 173(5):446–454, 2019. 2

[51] Catalin Voss, Peter Washington, Nick Haber, Aaron Kline, Jena Daniels, Azar Fazel, Titas De, Beth McCarthy, Carl Feinstein, Terry Winograd, et al. Superpower glass: delivering unobtrusive real-time social cues in wearable systems. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*, pages 1218–1226, 2016. 2

[52] Peter Washington, Haik Kalantarian, Jack Kent, Arman Husic, Aaron Kline, Emilie Leblanc, Cathy Hou, Cezmi Mutlu, Kaitlyn Dunlap, Yordan Penev, et al. Training affective computer vision models by crowdsourcing soft-target labels. *Cognitive Computation*, 13(5):1363–1373, 2021. 5

[53] Peter Washington, Haik Kalantarian, Qandeel Tariq, Jessey Schwartz, Kaitlyn Dunlap, Brianna Chrisman, Maya Varma, Michael Ning, Aaron Kline, Nathaniel Stockham, et al. Validity of online screening for autism: crowdsourcing study comparing paid and unpaid diagnostic tasks. *Journal of medical Internet research*, 21(5):e13668, 2019. 5

[54] Peter Washington, Emilie Leblanc, Kaitlyn Dunlap, Yordan Penev, Aaron Kline, Kelley Paskov, Min Woo Sun, Brianna Chrisman, Nathaniel Stockham, Maya Varma, et al. Precision telemedicine through crowdsourced machine learning: testing variability of crowd workers for video-based autism feature recognition. *Journal of personalized medicine*, 10(3):86, 2020. 5

[55] Peter Washington, Emilie Leblanc, Kaitlyn Dunlap, Yordan Penev, Maya Varma, Jae-Yoon Jung, Brianna Chrisman, Min Woo Sun, Nathaniel Stockham, Kelley Marie Paskov, et al. Selection of trustworthy crowd workers for telemedical diagnosis of pediatric autism spectrum disorder. In *BIOCOMPUTING 2021: Proceedings of the Pacific Symposium*, pages 14–25. World Scientific, 2020. 5

[56] Peter Washington, Qandeel Tariq, Emilie Leblanc, Brianna Chrisman, Kaitlyn Dunlap, Aaron Kline, Haik Kalantarian, Yordan Penev, Kelley Paskov, Catalin Voss, et al. Crowdsourced feature tagging for scalable and privacy-preserved autism diagnosis. *medRxiv*, 2020. 5

[57] Peter Washington, Catalin Voss, Aaron Kline, Nick Haber, Jena Daniels, Azar Fazel, Titas De, Carl Feinstein, Terry Winograd, and Dennis Wall. Superpowerglass: a wearable aid for the at-home therapy of children with autism. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, 1(3):1–22, 2017. 2

[58] Werner Geyser. Tiktok statistics – 63 tiktok stats you need to know Fact Sheet N°282, 2022. https://influencermarketinghub.com/tiktok-stats/, Last accessed on 2022-03-30. 2

[59] Stefanos Zafeiriou, Dimitrios Kollias, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, and Irene Kotsia. Aff-wild: Valence and arousal 'in-the-wild'challenge. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*, pages 1980–1987. IEEE, 2017. 8

[60] Guoying Zhao, Xiaohua Huang, Matti Taini, Stan Z Li, and Matti PietikäInen. Facial expression recognition from near-infrared videos. *Image and Vision Computing*, 29(9):607–619, 2011. 2