Indian Semi-Acted Facial Expression (iSAFE) Dataset for Human Emotions Recognition

Shivendra Singh and Shajulin Benedict(✉)

Indian Institute of Information Technology Kottayam, Kottayam, Kerala, India
shivendra15@alumni.iiitkottayam.ac.in, shajulin@iiitkottayam.ac.in
http://www.iiitkottayam.ac.in/shajulin.php

Abstract. Human emotion recognition is an imperative step to handle human computer interactions. It supports several machine learning based applications, including IoT cloud societal applications such as smart driving or smart living applications or medical applications. In fact, the dataset relating to human emotions remains as a crucial pre-requisite for designing efficient machine learning algorithms or applications. The traditionally available datasets are not specific to the Indian context, which lead to an arduous task for designing efficient region-specific applications. In this paper, we propose a new dataset that reveals the human emotions that are specific to India. The proposed dataset was developed at the IoT Cloud Research Laboratory of IIIT-Kottayam – the dataset contains 395 clips of 44 volunteers between 17 to 22 years of age; face expressions were captured when volunteers were asked to watch a few stimulant videos; the facial expressions were self annotated by the volunteers and they were cross annotated by annotators. In addition, the developed dataset was analyzed using ResNet34 neural network and the baseline of the dataset was provided for future research and developments in the human computer interaction domain.

Keywords: Human computer interaction · Affective computing · Human emotions · Facial expression recognition

1 Introduction

Perceptual processing of facial expressions plays a major role while expressing emotions. Emotion exercises a powerful influence on human action [5]. A better emotion recognition, obviously, brings better expression to life in animated movies and *animojis*. In human computer interaction, therefore, deciding the course of action heavily depends on emotions. In fact, the identification of emotion is applied in several domains such as computer vision, artificial Intelligence and emotional science; it could help in medical science [17], it could automate the selection of the best suited music for user's current mood and it could also provide a better music visualization than animated patterns [4].

S. M. Thampi et al. (Eds.): SIRS 2019, CCIS 1209, pp. 150–162, 2020.
https://doi.org/10.1007/978-981-15-4828-4_13

Ekman and Friesen in his research work presented six basic emotions such as *Anger, Disgust, Fear, Happiness, Sadness, Surprise* that are universally present among humans irrespective of cultural differences [6]. But, due to the differences in the physical appearance of people from different race there is a need for region specific databases. The authors, in their work, had identified six basic emotions based on their study on the isolated culture of people from the Fori tribe in Papua New Guinea in 1972. The tribe members were able to identify these six emotions on the pictures.

Precisely, a well-annotated media content of facial expressions is needed for training and testing machine learning algorithms which could be applied in recognition systems or IoT cloud applications. This paper focuses on developing a dataset that highlights the emotions in the Indian scenario. In this dataset, we have collected temporal and spatial expressions that are extracted from the recorded video when stimulant video was watched by volunteers. Temporal information is utilized to identify the emotions and capture more information when compared to the spacial information. We have disclosed an exploratory survey on the available databases. In addition, we have analyzed the importance of the proposed dataset by conducting experiments at the IoT Cloud research laboratory of our premise.

The rest of the paper is organized as follows: Sect. 2 reviewed the available databases in line to human emotion recognition; Sect. 3 described the details on the proposed database and its attributes; Sect. 5 explained about the experimental setup and the analysis report on the database; and, finally, Sect. 6 reports on the conclusion of the paper.

2 Literature Survey

Machine learning and data science domains, including IoT cloud domain [15,16], in general, demand high quality datasets for accurate modeling or accurate predictions. Research works were carried out in the past to create human emotion datasets. For instances, a large label based dataset was developed through croud sourcing by [2] for representation learning in the facial emotion recognition domain. However, the images available in the dataset were only 48 * 48 in dimension. Similarly, in 1998, a Japanese database [12] of 213 images was developed with 7 different emotions by 10 posers.

Recently, Cohn-Kanade's AU-coded facial expression database [10] has become quite popular among researchers. It is a French database which contains 486 clips by 97 posers. Later, they have released it's extension known as CK+ database which contains 593 clips of 123 posers. The database has also provided a baseline for future research and for pursuing comparative studies.

In the meantime, Affect Net has emerged as one of the widely known dataset among researchers in recent years. This database contains over 1 million images [13]. The authors of Affect Net have collected images by querying three major search engines using 1250 keywords relating to emotions in six different languages. Out of 1 million images, 40% of images are manually annotated and they are annotated by applying machine learning based training model on the manually annotated images.

A few researchers have attempted to create an online-based database – Daniel et al. [11] recorded clips of emotions over internet. The participants were subjected to watch commercial products and the emotions were captured through webcams. This particular database contains 242 facial videos. In addition, there were efforts in the past to capture emotions using physiological signals [9] and audio signals [3].

Region specific datasets are crucial for several machine learning based applications in order to achieve quality results. For instance, Happy et al. [7] have presented an Indian dataset named as Indian Spontaneous Expression Dataset (ISED) – ISED bestowed lots of emphasis on the spontaneity of emotions; it contained near frontal face clips of recorded videos of 50 participants (29 male, 21 female) when exposed to stimulant videos; ISED utilized clips rather than a single image, which helped to capture temporal as well as spatial sequence of changes in behaviour, while capturing emotions. ISED dataset is more relevant to our work. However, the ISED dataset has only four emotions.

Our proposed iSAFE dataset, in this paper, provides temporal information and contains the required 6 basic emotions for facial recognitions. The dataset is generated and stored in databases which could be collected over internet via. http protocol.

3 iSAFE - Dataset Creation Approach

This section describes the approach of creating iSAFE – i.e., the protocols involved, stimulant video utilized, and the volunteers involved (see Fig. 1).

In order to capture high quality videos, proper lighting conditions were arranged in the recording room. Hearing devices were provided so that volunteers could capture intense details of stimulant clips. The stimulant clips were played on a laptop, which was kept in front of the volunteer, and the camera was placed behind it in order to capture the emotions.

3.1 Camera

Nikon D5000 camera with auto ISO and exposure settings was utilized for recording videos. The video was recorded at high resolutions. For instance, in our experiments, we have utilized a 1920 × 1080 resolution with the rate of 60 frames per second.

3.2 Protocols

Emotions were captured using the above mentioned camera while setting up an open environment rather than a closed laboratory. Before exposing stimulant videos to the volunteers, some general conversations were done by the experimenter who knows the subject well in order to make volunteers to feel comfortable with the exercises. In fact, the volunteers were not monitored or interrupted while watching the stimulant clips.

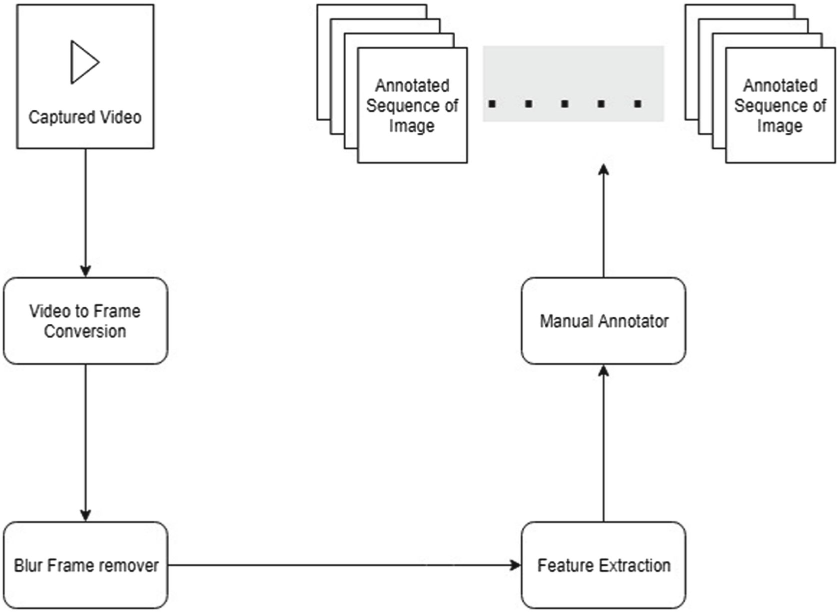

Captured Video
Video to Frame Conversion
Blur Frame remover
Feature Extraction
Manual Annotator
Annotated Sequence of Image
. . . . .
Annotated Sequence of Image

Fig. 1. Data pre-processing and annotation steps for iSAFE

3.3 Stimulant Video

Videos for inducing emotions were carefully selected and tested before the experiments were conducted by an experienced person. They were downloaded from various sources on internet and emotion inducing parts were trimmed and merged together. However it was experienced that stimulant videos were not much effective in stimulating sadness and anger as they were so for happiness, disgust, fear and surprise.

3.4 Volunteers

All volunteers were between 17 to 22 years of age. Out of 44 volunteers, 25 volunteers were male and 19 were female. They were educated about the six basic emotions before recording the emotions. At the end of the recording, consent were obtained from them in order to utilize the videos for the research and education purposes. To do so, they were asked to sign in the "Data Usage Agreement" (DUA) document. All the volunteers were fit and healthy while recording the emotions. However, no formal medical tests were conducted to the volunteers for the experiments.

4 iSAFE - Data Pre-processing and Annotations

In this section, we have described the data pre-processing stages and the annotation approaches carried out during the process of creating iSAFE database.

All the clips are split from the recorded video of volunteers while watching a few stimulant clips. The emerging facial expressions are self annotated by the volunteers and are cross annotated by an annotator. Both of the annotations are entered into the database. Due to the presence of sequence of images of emotions, the movement of facial muscles provides better classifications over time. A brief description of the iSAFE database is provided in Table 1.

Table 1. Brief description of iSAFE database

Number of clips	395
Number of volunteers	44
Age	17–22 years
Emotion labels	Happy Sad Surprise Disgust Fear Anger Uncertain No-Emotion
Number of annotator	2 + 1 (self-annotation)
Race and ethnicity	Indo-Aryan and Dravidian (Asian)

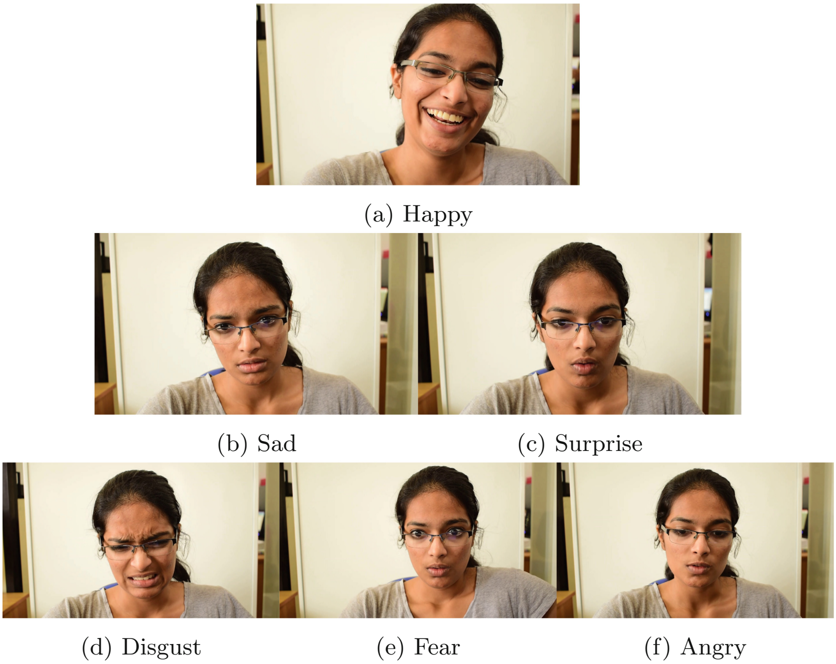
The sequence of emotions was carefully annotated in iSAFE database. In fact, the frames, at the start of the captured video, would not provide sufficient emotions when compared to the last frames. In addition, the feature extractions might lead to inaccuracy due to the presence of objects such as hand, spectacles, and so forth. In short, the sequence of extracted emotions was carefully annotated in the iSAFE database (as shown in Fig. 2).

4.1 Conversion of Recorded Video to Frames

The recorded videos of volunteers are trimmed out to collect the regions where the presence of expressions are detected. Later, these trimmed video clips are converted to images by taking each frame out of the video clips. Python codes were written in order to convert video clips to frames.

4.2 Removal of Blurry Frames

To ensure the quality of database, blurred and shaky frames were removed from the database. For removing the blur in an image, there exists a few techniques [14]. Out of those techniques, fast fourier transform and laplacian of image performs well according to the literature. Hence, we have utilized the variance of



(a) Happy

(b) Sad

(c) Surprise

(d) Disgust

(e) Fear

(f) Angry

Fig. 2. Different emotions of a volunteer

laplacian of image to find out the blurred frame as it provides single floating point value to represent the blur in a given image. If the variance of image reaches below a certain threshold, the image is considered to be shaky or blurry; otherwise, the image is graded as a good image (see Fig. 3).

Laplacian approach focuses on the regions of a picture containing the rapid change in the intensity and sometimes it is utilized for edge detections. Our assumption here is that if an image has a high variance, then there is a wide spread of responses. If there is very low variance, then there is a tiny spread of responses, indicating that there are very less edges in the image. This is due to the fact that the more an image is blurred, the lesser the edges are present.

4.3 Feature Extraction

With visual data, the extracted features of images are also provided with the database. Features from the images are extracted using openface [1]. The features that were extracted from the images include a subset of facial Action Units (AUs), intensity of AUs, 45 Facial landmarks in 2D and 3D, head pose and Eye gaze movement data and so forth.

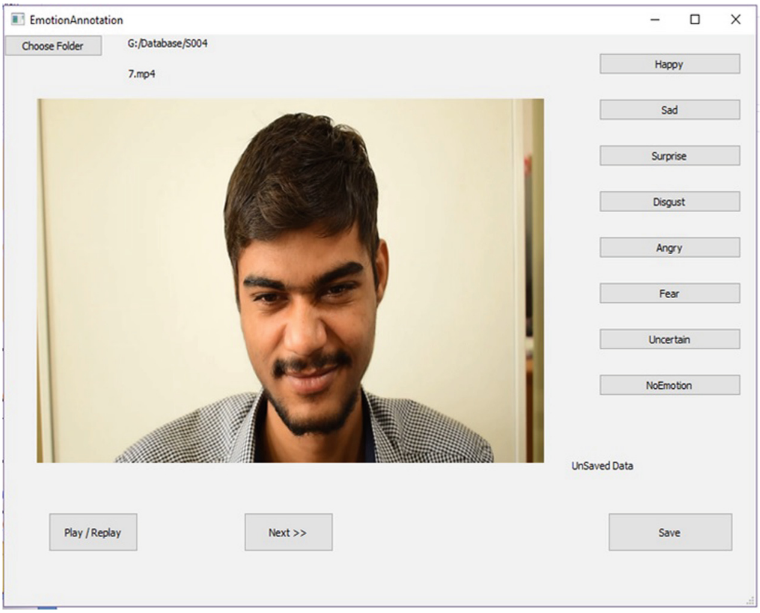
The volunteers involved in creating the iSAFE database are aware that they would be filmed for extracting required features. However, they are not instructed

(a) Good image

(b) Shaky image

Fig. 3. Blurred and good image detected using variance of laplacian of images method

to act. Hence, iSAFE database is considered to be a Semi-Acted Feature Extraction based Indian dataset.



G:/Database/S004

7.mp4

UnSaved Data

Fig. 4. GUI for manual annotations

The entire dataset is available for future reference and research works. Interested candidates could sign an End User License Agreement (EULA) and obtain the dataset through the github resource links of our lab.

4.4 Manual Annotation

Manual annotation of the database is done by a trained annotator and by the volunteers itself. The trained annotator, in our approach, is the psychologist of our premise who is trained in assessing the human emotions.

In order to identify the emotions experienced by the volunteer, they were asked to self-annotate these data. The self-annotation was carried out by an user-interface portal as depicted in Fig. 4.

As the assessment of emotion plays an important role in the success of creating database, the annotations were carried out very carefully. In addition, the annotation score of all the annotators are available in the database.

5 iSAFE Analysis and Baseline Creation

We have also analysed the database and proposed a baseline for the classification of images into various emotions. In general, convolutional neural networks (CNN) are the state of the art algorithms for image classification. We have utilized ResNet34 (Residual Networks with 34 layers) convolutional neural network for the evaluations.

In succinct, ResNet34 explores the bigger parameter space of images in a deep mode in order to solve the vanishing gradiant problems. To do so, convolution and pooling steps are iteratively targeted on the input images [8].

For the baseline in iSAFE, we have removed some images from each sub-sessions. And, multiple images from the same session had provided a good augmentation with some small variation in images (see Fig. 6).

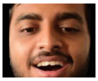
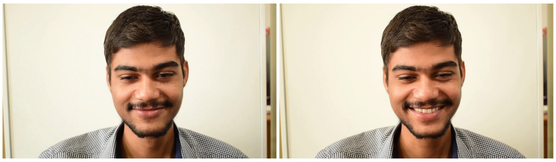


Fig. 5. Masked image



(a) Frame 1 of same sub-session (b) Frame 10 of same sub-session

Fig. 6. Different frame from same sub-session (Augmented data)

5.1 Training, Testing and Validation Dataset

Three annotations are provided for each sub-session which includes self annotation, annotation from trained professional and annotation from trained unprofessional. The agreement between self annotation and trained professional is shown in Fig. 7 and agreement between trained unprofessional and trained professional is shown in Fig. 8. The agreements were represented as confusion matrices.

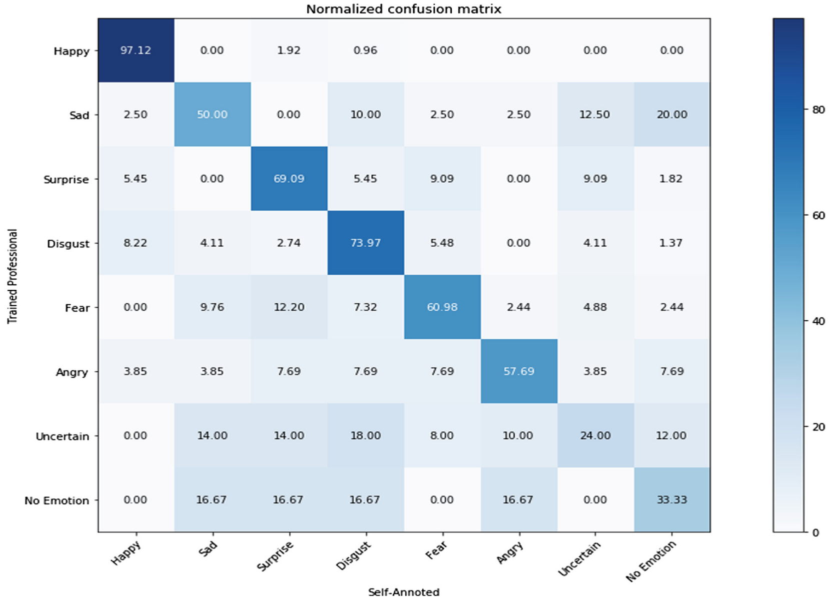


Fig. 7. Agreement between self annotation and trained professional

From the analysis we can find out that there is xx% agreement between self annotation and professional while xx% agreement between unprofessional and professional. Notably, annotating happy emotion observed 94.23% agreement between trained and un-trained professional; 97.12% agreement between trained and self-annotated emotions. Uncertain class of annotation was not agreed upon effectively – 40% agreement between trained and untrained professional, and 24% agreement between trained and self-annotated professionals.

Distribution of processed images used for training was skewed towards happy emotions so that we could make the distribution in a even manner. In addition, we have dropped some of the augmented frames of happy emotions in order to reduce the bias. Uncertain class was creating confusion while training models so we had dropped that class too.

Initially, we started training on images without cropping out faces. However, it ended up with very poor results – i.e., we observed around 30% accuracy for the training models. Subsequently, we have cropped out faces from the image

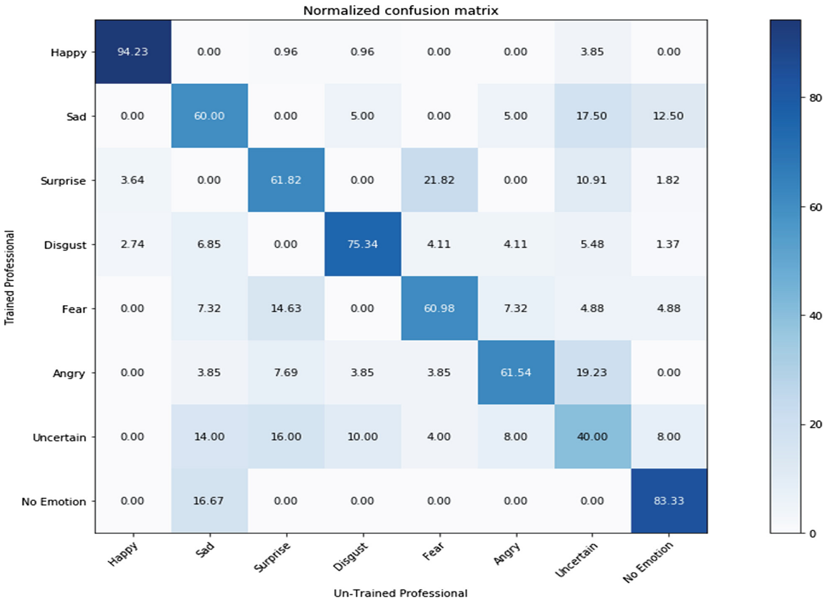


Fig. 8. Agreement between trained and untrained professional

and masked the unnecessary area with black masks as shown in Fig. 5. By this approach, we could improve the accuracy and drop the anticipated losses.

The processed dataset for evaluation is divided randomly into 60%, 30%, and 10% for training, validation and testing respectively.

5.2 Evaluation Metrics

For multi class classification, the predicted class labels should perfectly match with actual class labels. Accuracy score gives a closer insight towards the performance of model. It performs well on class balanced dataset. For binary classification, accuracy score can be defined as

$$AccuracyScore = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

where, TP equals True Positives; TN equals True Negatives; FP equals False Positives; and, FN equals False Negatives.

We are using the $AccuracyScore$ for evaluating the proposed model. Figures 9 and 10 illustrate the skewness of data with class imbalance dataset and without class imbalance datasets.

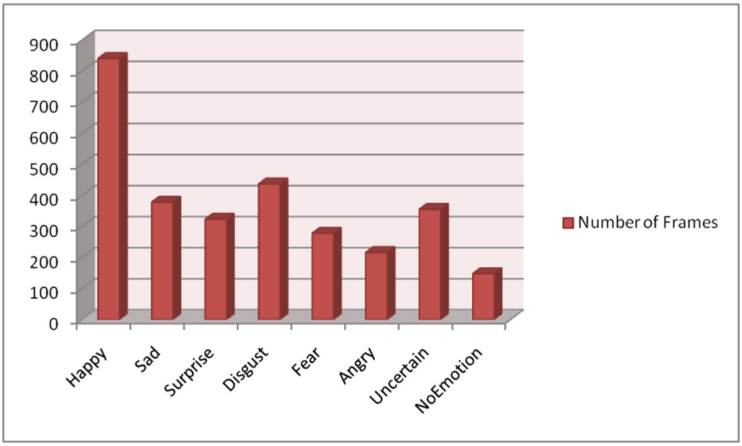


Fig. 9. Data with class imbalance

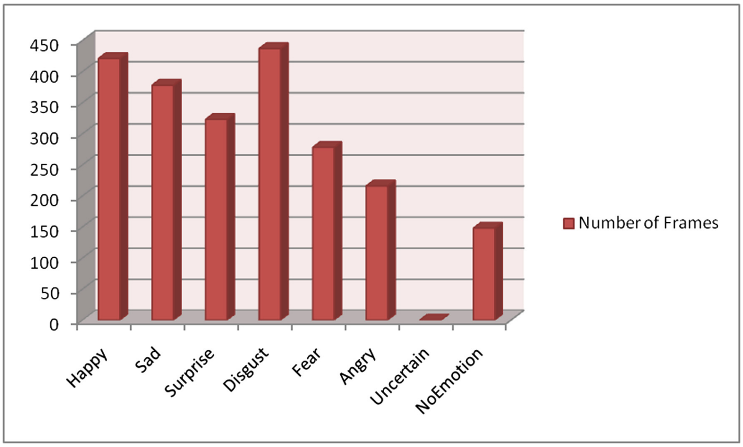


Fig. 10. Data without class imbalance

5.3 ResNet Analysis Results

Our Model took around 50 min on an Nvidia K80 Tesla GPU with 12 GB memory and reached to 98.32 % of accuracy on validation dataset; and, 92% on test dataset.

6 Conclusions

Emotion recognition using facial expressions by computer needs a quality facial expression databases with reliable annotations. The proposed iSAFE fulfills a number of aspects of the desired requirements. It has mild to strong facial expressions, and it's quality build can help researchers to develop algorithms for the recognition of human emotions in practical situations – especially, in emerging applications such as IIoT applications. Several protocols and strategies are adopted for the creation of the database which have been briefly described in the paper. Stimulant clips were presented to keep the subject engaged and to induce spontaneous emotions. Due to the presence of large number of volunteers and the variation of expression across time, several expressions of the same participants are included in iSAFE. The recorded videos were further processed and trimmed into small videos and further converted to sequence of images. Blurry frames are also removed from the sequence. The video clips of the database are annotated carefully by annotators and by the volunteers itself.

In data science, the amount of data provides direct advantage in availing concrete results and leverages to train model better. iSAFE is still very small in size; it requires more recordings for appending to the database. Some evaluation of database needs to be carried out in order to provide reference evaluation results for researchers. And, an increase in the number of annotations will help the database to become more reliable.

Acknowledgement. The authors thank IIIT Kottayam officials for granting space and support in order to carry out this research work at IoT Cloud research lab of IIIT Kottayam.

References

1. Baltrusaitis, T., Zadeh, A., Lim, Y.C., Morency, L.: OpenFace 2.0: facial behavior analysis toolkit. In: 2018 13th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2018) (FG), pp. 59–66, May 2018
2. Barsoum, E., Zhang, C., Ferrer, C.C., Zhang, Z.: Training deep networks for facial expression recognition with crowd-sourced label distribution. In: ACM International Conference on Multimodal Interaction (ICMI) (2016)
3. Busso, C., et al.: IEMOCAP: interactive emotional dyadic motion capture database. Lang. Resour. Eval. **42**(4), 335 (2008)
4. Chen, C.-H., Weng, M.-F., Jeng, S.-K., Chuang, Y.-Y.: Emotion-based music visualization using photos. In: Satoh, S., Nack, F., Etoh, M. (eds.) MMM 2008. LNCS, vol. 4903, pp. 358–368. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-77409-9_34
5. Dolan, R.J.: Emotion, cognition, and behavior. Science **298**(5596), 1191–1194 (2002)
6. Ekman, P.: Universals and cultural differences in facial expressions of emotion, pp. 207–283 (1971)
7. Happy, S.L., Patnaik, P., Routray, A., Guha, R.: The Indian spontaneous expression database for emotion recognition. IEEE Trans. Affect. Comput. **8**(1), 131–142 (2017)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. CoRR, abs/1512.03385 (2015)
9. Koelstra, S., et al.: DEAP: a database for emotion analysis; using physiological signals. IEEE Trans. Affect. Comput. **3**(1), 18–31 (2012)

10. Lucey, P., Cohn, J.F., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I.: 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, pp. 94–101, June 2010
11. McDuff, D., Kaliouby, R., Senechal, T., Amr, M., Cohn, J.F., Picard, R.: Affectiva-MIT facial expression dataset (AM-FED): naturalistic and spontaneous facial expressions collected “In-the-Wild”. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 881–888, June 2013
12. Kamachi, M., Gyoba, J., Lyons, M.J., Akemastu, S.: Coding facial expressions with gabor wavelets, pp. 200–205 (1998)
13. Mollahosseini, A., Hassani, B., Mahoor, M.H.: AffectNet: a database for facial expression, valence, and arousal computing in the wild. CoRR, abs/1708.03985 (2017)
14. Pech-Pacheco, J.L., Cristobal, G., Chamorro-Martinez, J., Fernandez-Valdivia, J.: Diatom autofocusing in brightfield microscopy: a comparative study. In: Proceedings 15th International Conference on Pattern Recognition, ICPR 2000, vol. 3, pp. 314–317, September 2000
15. Gowtham, N., Benedict, S., Giri, D., Sreelakshmi, N.: Real time water quality analysis framework using monitoring and prediction mechanisms. In: IEEE CiCT2018, pp. 1–6 (2018). https://doi.org/10.1109/INFOCOMTECH.2018.8722381
16. Ajith, S., Kumar, S., Benedict, S.: Application of natural language processing and IoTCloud in smart homes. In: Proceedings of IEEE-ICCT2019 (2019)
17. Tivatansakul, S., Ohkura, M., Puangpontip, S., Achalakul, T.: Emotional healthcare system: emotion detection by facial expressions using Japanese database. In: 2014 6th Computer Science and Electronic Engineering Conference (CEEC), pp. 41–46, September 2014