Facial Expression Recognition using Distance Importance Scores Between Facial Landmarks

Elena Ryumina[1,2][0000-0002-4135-6949], and Alexey Karpov[1][0000-0003-3424-652X]

[1] St. Petersburg Institute for Informatics and Automation
of the Russian Academy of Sciences (SPIIRAS), St. Petersburg, Russia
[2] ITMO University, St. Petersburg, Russia
ryumina_ev@mail.ru, karpov@iias.spb.su

Abstract. In this paper, we present a feature extraction approach for facial expressions recognition based on distance importance scores between the coordinates of facial landmarks. Two audio-visual speech databases (CREMA-D and RAVDESS) were used in the research. We conducted experiments using the Long Short-Term Memory Recurrent Neural Network model in a single corpus and cross-corpus setup with different length sequences. Experiments were carried out using different sets and types of visual features. An accuracy of facial expression recognition was 79.1% and 98.9% for the CREMA-D and RAVDESS databases, respectively. The extracted features provide a better recognition result compared to other methods based on the analysis of facial graphical regions.

Keywords: Visual Feature Extraction · Facial Landmarks · Facial Expression Recognition · Automatic Emotion Recognition

1 Introduction

Facial expressions are an important channel of nonverbal communication, so interest in automatic recognition of human emotions by facial expressions increases every year. This is also due to the fact that smart emotion recognition technologies are in demand and are introduced around the world, for example, automatic facial expression recognition systems are widely used in medicine [1], psychology [2], education [3], fraud detection [4], driver assistance systems [5], etc. In recent years, more research has focused on the analysis of facial expressions in a video [6–9] since video can transmit a change in facial expressions over time. Feature extraction is one of the most important steps in facial expression recognition systems by a video stream [10].

* This research is supported by the Russian Science Foundation (project No. 18-11-00145).

The main problems faced by researchers in the field of facial expression recognition are high variability in illumination, occlusions, gender, age, national origin, intra-class variation, inter-class similarities. The extraction of graphical facial regions, which is the most widely used approach, does not cope well with illumination and occlusion problems, while finding the coordinates of facial landmarks adapts well to illumination variation and partial occlusion.

In this work, we extracted the coordinates of facial landmarks from a video stream of two large-scale databases: CREMA-D and RAVDESS. Important features were calculated in the form of Euclidean distances between landmarks, and the importance of the features was evaluated using ensemble classifiers. We extracted features according to the algorithm presented in [11] to compare the effectiveness of the proposed approach. We formed the extracted features into sequences of different lengths, which were applied as a neural network input.

The rest of the article is organized as following: Section 2 presents analysis of existing approaches in the field of facial expression recognition and a brief overview of available emotional databases, Section 3 gives a new approach to feature extraction from the coordinates of facial landmarks, Section 4 shows the results of conducted experiments, Section 5 contains the discussion and conclusions.

2 Related Work

2.1 Facial Features

There are two main approaches to feature extraction from a video stream, namely: extracting facial graphical regions, where it is possible to save the raw images or use various methods of preprocessing images of faces [8, 9]; finding the coordinates of facial landmarks and extracting distances, angles, areas and other calculations with the coordinates found [12, 13].

Detection of facial landmarks in facial graphical images is performed by finding the points on regions of the mouth, eyebrows, eyes, nose, etc. This is easily implemented using pre-trained models from the Dlib library [14]. To date, there are a few research works based on finding and tracking landmarks [15, 16]. Emotii application on Android for audio-visual mood analysis is presented in [11]. Emotii recognizes the user's mood from the video by extracting coordinates of facial landmarks, distance from the coordinates to the "Center of Gravity" and calculating the face offset correction by finding the angle of the nose. A similar approach was previously proposed in [13]. OpenFace - an open source framework is described in [6]. OpenFace tracks facial landmarks, head position, gaze and evaluates facial Action Units (AU) [17]. This allows for analysis of facial behavior in real time. A face can be divided into regions of interest using the coordinates of the facial landmarks. The division of the face into 12 regions of interest is suggested in [18]. Regions are analyzed for changes in the intensity of each pixel using histograms. Determining pixel intensity allows tracking the changes in micro-expressions in successive images. The method of facial expression recognition based on 74 geometric features from (x, y)-coordinates, namely 11 distances and 26 areas for each coordinate is presented in [12].

To date, except for [11, 19], feature vectors have not been extracted from the CREMA-D [20] or RAVDESS [21] databases using facial landmarks. In [11], an accuracy was achieved by 96.3% for the RAVDESS with 7 classes (calmness was not considered) using the Support Vector Machine (SVM). In [19], authors proposed using facial landmarks to detect facial regions in the image with further conversion to grayscale. Then 32 features are extracted from the images using Gabor filters, which are combined with 68 positions of facial landmarks. After reading all frames from the video, the values of 2176 (32×68) features are averaged. An accuracy of 96.53% was obtained for the RAVDESS with 8 classes of emotional speech. An approach based on the 3D Convolutional Neural Networks (CNN) branch of a Two-Stream Inflated 3D ConNect and randomly resizing frames to increase data is described in [9]. The time context of images of detected faces is considered using Long Short-Term Memory network (LSTM). An accuracy was 66.8% and 60.5% for the CREMA-D and RAVDESS databases, respectively. The use of Haar features to detect facial regions with subsequent rotation of images at the same level of the pupils of the eye is proposed in [22]. An accuracy of 79.74% was achieved in 6 classes of emotional songs of the RAVDESS database using the pre-trained model of CNN Alex net.

2.2 Emotional Databases

Emotional database (EDB) is a key element in the emotion recognition task. EDB are divided into multimodal, bimodal or unimodal. Visual databases contain of images or video clips with facial expressions. Well-annotated data have a significant impact on the performance of machine learning classification algorithms. Most databases assume 5-7 basic emotions, namely happiness, sadness, anger, fear, disgust, surprise, and neutral. However, some databases include valence-arousal dimensions and AU codes. Also, EDBs are divided into ones collected in laboratory (imitating emotional expressions) and real ("in-the-wild" - natural emotional expressions) conditions. An extended overview of multimodal databases is presented in [23]. Several most popular of the existing EDBs are compared in Table 1.

For our experiments, we have selected and used two representative audio-visual databases with varying levels of emotional intensity: CREMA-D and RAVDESS.

CREMA-D database contains 7442 videos for speech, where 91 actors imitate 6 emotions, happiness (1271 videos), sadness (1271), anger (1271), fear (1271), disgust (1271), and neutral (1087). The cast has different ethnicities ranging in age from 20 to 74 years. The resolution of video clips is 480×360 with 30 frames per second. The database was evaluated by 2443 people for audio, video, and audiovisual data, where an accuracy of emotion recognition for the considered modalities was 40.9%, 58.2%, and 63.6%, respectively.

RAVDESS database contains 4904 videos for speech and songs, where 24 actors imitate 8 emotions, happiness (752 videos), sadness (752), anger (752), fear (752), disgust (384), surprise (384), neutral (376), and calmness (752). The resolution of video clips is 1280×720 with 30 frames per second. The database was evaluated by 247 people for audio, video, and audiovisual data, where an accuracy of emotion recognition for the considered modalities was 60%, 75% and 80%, respectively.

Table 1. A comparison of multimodal emotional databases

Database	# Subjects	# Emotions	# Videos	Specificity
CK+ [24]	123	7	593 sequences	AU codes
MMI [25]	75	5	over 2900 (videos + images)	AU codes. Various ethnicity
SAVEE [26]	4	7	480	60 markers on the faces
Oulu-CASIA [27]	80	6	480 sequences	Various illumination conditions and age (23 to 58 years old)
CREMA-D [20]	91	6	7442	Various age (20 to 74 years old), races and ethnicity
RAVDESS [21]	24	8	4904	Emotional speech and song
RAMAS [28]	10	7	564	Motion-capture data and physiological signals
Aff-Wild2 [29]	458	7	558	"In-the-wild" database. AU codes and valence-arousal dimensions.

3 Proposed Method for Feature Extraction

The architecture of our proposed approach for feature extraction and facial expression recognition is depicted in Figure 1.

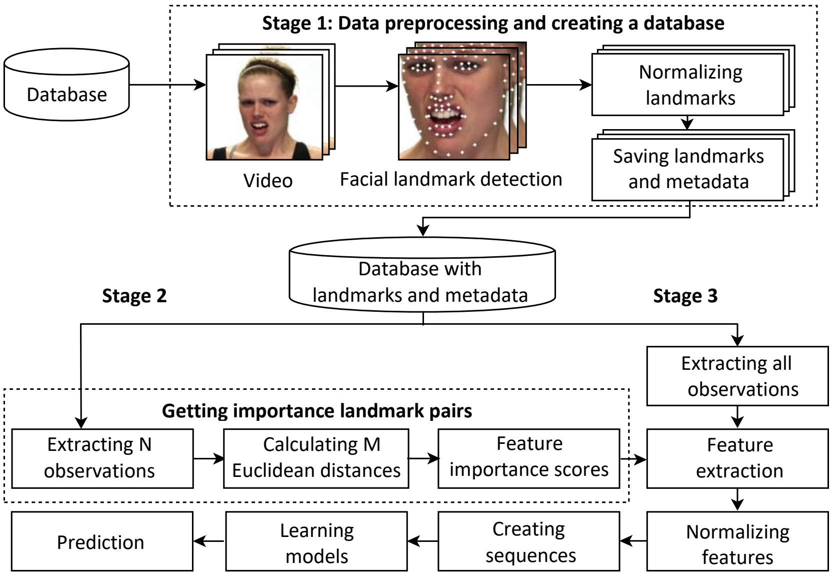

Stage 1: Data preprocessing and creating a database
Database
Video
Facial landmark detection
Normalizing landmarks
Saving landmarks and metadata
Database with landmarks and metadata
Stage 2
Stage 3
Extracting all observations
Getting importance landmark pairs
Extracting N observations
Calculating M Euclidean distances
Feature importance scores
Feature extraction
Prediction
Learning models
Creating sequences
Normalizing features

Fig. 1. Pipeline of our approach for feature extraction and facial expression recognition.

Data preprocessing and creating a database with facial landmarks and metadata are carried out at Stage 1. We used Dlib open source library [14] to find the coordinates of key facial landmarks. The detected coordinates were scaled to a resolution of 224×224 pixels since the video resolutions in the research datasets are different. Then we saved the received coordinates and metadata about the video and frames (database, video title, video duration, frame number, emotion) for subsequent extraction of features. As a result of processing, it was revealed that the average video duration for the CREMA-D database is 76 frames, and 122 frames for RAVDESS.

Feature importance scores are performed at Stage 2. We randomly took 120K observations from the considered databases. We extracted 2278 unique Euclidean distances between the coordinates of facial landmarks (for example, the distances between the coordinates of points 0 and 1 is equal to the distance between the coordinates of points 1 and 0, so only one of two possible combinations was taken into account). But since not all the distances considered have a positive impact on the decision-making of a classifier, it is necessary to leave only the most important features. The obtained observations were used as input to the ensemble classifiers Random Forest Classifier (RFC) [30], Extra Trees Classifier (ETC) [31] and AdaBoost Classifier (ABC) [32], which allow us to calculate feature importance scores. The parameters of classifiers are shown in Table 2.

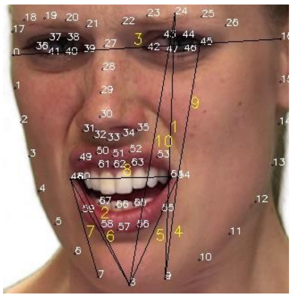
Table 2. Parameters of the machine classifiers applied.

Classifier	Optimised Parameters
RFC	n_jobs=3, n_estimators=500, warm_start=True, max_depth=6, min_samples_leaf=2, max_features=sqrt
ETC	n_jobs=3, n_estimators=500, max_depth=8, min_samples_leaf=2
ABC	n_estimators=n_estimators, learning_rate=0.75

We obtained feature importance scores from 3 different classifiers and averaged them to find the mean. We then set three different thresholds of importance (0.0009, 0.001 and 0.002) and obtained three different feature sets, whose importance scores exceed the corresponding threshold. This resulted in sets of 368, 259 and 104 features, respectively. An algorithm for processing landmark pairs is open-sourced[†]. The 10 most important distances between the coordinates of facial landmarks with their feature importance scores are depicted in Figure 2 using the example of a frame from the RAVDESS database.

The high score was obtained by the distance between facial landmarks 9 and 24 and amounted to 0.014. As one can see from the figure, most of the 10 important distances are in the lower part of the face.

[†] The annex to article "Facial Expression Recognition using Distance Importance Scores Between of FacialLandmarks", https://elenaryumina.github.io/GraphiCon_2020/



1	0.014
2	0.008
3	0.006
4	0.006
5	0.006
7	0.005
8	0.005
9	0.005
10	0.005

Fig. 2. Top-10 important distances and their feature importance scores.

Feature extraction of various sets, learning models, and obtaining predictions are carried out at Stage 3. We calculated 368, 259 and 104 Euclidean distances for all observations. 272 features were extracted using the algorithm presented in [11] to compare the effectiveness of the proposed approach. Thus, 5 different feature sets with dimensionalities 136 (68 (x, y)-coordinates of facial landmarks), 272, 104, 259 and 368 were obtained, which were normalized by the average values and standard deviations of the features of the training set. This improves accuracy of facial expression classification. Feature vectors were applied as LSTM input, which includes two LSTM layers with 128 and 256 output neurons, and a dropout rate of 0.5 after each layer, the last layer is a fully connected layer with the number of neurons equal to the number of classes and with softmax activation function. The number of epochs for all experiments was 30. Adam was chosen as the optimizer with a learning rate of 0,001 and a weight decay of 0,00005. The size of batches was 64. The parameters were determined using a grid search at the first training stage.

4 Experimental Results

The experiments were carried out using the LSTM Recurrent Neural Network. Different sequence length of features was applied as LSTM input. First, we set the sequence length equal to the average video duration (76, 122). If the video duration was less than the average duration, then the arrays were supplemented with zeros to the desired length, if it was longer than the average duration, then frames were selected in steps equal to video length divided by the average video duration. Also, the sequence length was set equal to the number of frames per second (30). Then video sequences were divided into sections of 30 frames, if the section was less than 30 frames, then the array was supplemented with zeros, so all the frames were considered. We divided the datasets into 10 roughly identical sets to perform cross-validation. The reported results are the average of these 10 sets. We conducted experiments when training on one dataset and testing on another dataset. Since the CREMA-D dataset does not contain the emotions surprise and calmness and the average video duration of the CREMA-D database is 76 frames, so all emotions and sequence length of 122 frames were considered only when cross-validation for the RAVDESS database. Accuracy results for feature vectors with dimension 136 components and experiment numbers are shown in Table 3.

The best accuracy was achieved with a sequence length of 76, this is especially noticeable when training and testing on various databases, for the RAVDESS an increase in the accuracy by 9.05% was achieved, for the CREMA-D - 5.60%. The results show that the model trained on the CREMA-D database gives the better accuracy on unfamiliar samples compared to the model trained on the RAVDESS database. The results of accuracy and improvement obtained using feature vectors with dimensions 272, 104, 259 and 368 components are presented in Table 4. An absolute improvement in accuracy is considered relative to accuracy obtained without feature extraction from the coordinates of facial landmarks. Thus, 8 experiments (with setups presented in Table 3) were performed for each set of features.

Table 3. Accuracy results for feature vectors of 136 components.

No.	Training Database	Testing Database	Classes	Sequence length	Accuracy (%)
1	CREMA-D	RAVDESS	6	76	66.69
2	CREMA-D	CREMA-D	6	76	76.65
3	RAVDESS	CREMA-D	6	76	47.88
4	RAVDESS	RAVDESS	8	122	97.80
5	CREMA-D	RAVDESS	6	30	57,64
6	CREMA-D	CREMA-D	6	30	76.58
7	RAVDESS	CREMA-D	6	30	42.28
8	RAVDESS	RAVDESS	8	30	97.59

Table 4. Accuracy (A, %) and absolute improvement (Delta) values for various feature vectors.

No.	104 comp.		259 comp.		272 comp.		368 comp.	
	A	D	A	D	A	D	A	D
1	68.07	1.38	69.43	2.74	66.75	0.06	69.59	2.90
2	77.87	1.22	79.07	2.42	77.37	0.72	78.03	1.38
3	49.54	1.66	49.91	2.03	47.41	-0.47	49.15	1.27
4	98.65	0.85	98.86	1.06	97.84	0.04	98.41	0.61
5	59.00	1.36	58.43	0.79	57.71	0.07	57.44	-0.20
6	77.64	1.06	78.14	1.56	77.20	0.62	77.60	1.02
7	43.81	1.53	44.83	2.55	42.83	0.55	43.74	1.46
8	98.22	0.63	98.37	0.78	98.12	0.53	98.14	0.55

By feature extraction from the coordinates of facial landmarks using the method proposed in [11], the accuracy of facial expression recognition was increased, a growth rate of over 1%. As can be seen from the table, feature vectors with dimensions of 259 components provide a greater growth in an accuracy value than feature vectors with dimensions of 104 and 368 components. In doing so, the accuracy exceeds the one obtained for feature vectors with dimensions 272 and 136 components. This confirms the effectiveness of the proposed approach. The classification accuracy was 79.07% and 98.68% by using cross-validation with the average video duration and feature vectors of dimension 259 components for the CREMA-D and RAVDESS databases, respectively. An accuracy of 69.43% and 49.91% was achieved with a dimension of feature vectors 259 components and a sequence length of 76 by training and testing on different databases for the RAVDESS and CREMA-D, respectively.

Table 5 shows a comparison of our accuracy with other solutions proposed in the recent literature.

Table 5. Comparison of the proposed method with existing approaches.

Method	Classes	Accuracy, %
CREMA-D		
Cao et al. 2014 [20]	6	58.2
Ghaleb et al. 2020 [9]	6	66.8
Proposed, seq. length 30	6	78.1
Proposed, seq. length 76	6	**79.1**
RAVDESS		
Livingstone et al. 2018 [21]	8	75.0
Ghaleb et al. 2020 [9]	8	60.5
He et al. 2019 [22]	6	79.7
Alshamsi et al. 2019 [11]	7	96.3
Jaratrotkamjorn et al. 2019 [19]	8	96.5
Proposed, seq. length 30	8	98.4
Proposed, seq. length 122	8	**98.9**

As can be seen from the table, our approach is superior to modern results in the task of classifying facial expressions on the CREMA-D and RAVDESS datasets. So, using facial landmarks significantly increases the accuracy of facial expression recognition compared to methods based on the analysis of facial graphical regions.

5 Conclusions

In the paper, we have studied various feature extraction methods calculated using coordinates of facial landmarks. The research was conducted on two large-scale datasets CREMA-D and RAVDESS containing various human's emotions with different degrees of intensity. The highest recognition accuracy was achieved after carrying out the following proposed processing steps. 68 detected coordinates of facial landmarks were scaled to an area of 224×224 since some videos have different resolutions. 2278 unique Euclidean distances were calculated between 68 facial landmarks. Three configurations with different number of facial distances were studied that have the greatest importance score and accurately characterize changes in facial expressions. LSTM has been applied to capture long-term dependence of frame-by-frame changes in facial expressions for different sequence lengths and feature sets. We analyzed the impact of different feature sets on facial expression recognition using both a single corpus (10-folds cross-validation experiments) and cross-corpus setups.

The experimental results showed that an absolute improvement of the recognition accuracy is achieved with an average video duration and the feature set of 259 components. This suggests that 259 components better generalize changes in facial expressions both in a single corpus and cross-corpus setup. The best recognition accuracy results of 79.1% and 98.9% were obtained with a single corpus for the CREMA-D and RAVDESS datasets, respectively. Our results of facial expression recognition outperform state-of-the-art results for the same datasets and experimental setups.

In our future work, we are going to apply the proposed approach to some other widely used databases, such as CK+, Aff-Wild2, etc.

References

1. Nijsse, B., Spikman, J. M., Visser-Meily, J. M., de Kort, P. L., van Heugten, C. M.: Social Cognition Impairments in the Long Term Post Stroke. Archives of Physical Medicine and Rehabilitation. vol. 100, no. 7, pp. 1300–1307 (2019)
2. Chen, L., Wu, M., Zhou, M., Liu, Z., She, J., Hirota, K.: Dynamic emotion understanding in human-robot interaction based on two-layer fuzzy SVR-TS model. IEEE Transactions on Systems, Man, and Cybernetics: Systems. vol. 50, no. 2, pp. 490–501 (2017)
3. Ninaus, M., Greipl, S., Kiili, K., Lindstedt, A., Huber, S., Klein, E., Moeller, K.: Increased emotional engagement in game-based learning–A machine learning approach on facial emotion detection data. Computers & Education. vol. 142, pp. 103641 (2019)
4. Prasad, N., Unnikrishnan, K., Jayakrishnan, R.: Fraud Detection by Facial Expression Analysis Using Intel RealSense and Augmented Reality. In: 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS), pp. 919–923 (2018)
5. Izquierdo-Reyes, J., Ramirez-Mendoza, R. A., Bustamante-Bello, M. R., Navarro-Tuch, S., Avila-Vazquez, R.: Advanced driver monitoring for assistance system (ADMAS). International Journal on Interactive Design and Manufacturing (IJIDeM). vol. 12, no.1, pp. 187–197 (2018)
6. Baltrušaitis, T., Robinson, P., Morency, L. P.: Openface: an open source facial behavior analysis toolkit. In: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV). Lake Placid, NY, USA, pp. 1–10 (2016)
7. Jannat, R., Tynes, I., Lime, L. L., Adorno, J., Canavan, S.: Ubiquitous emotion recognition using audio and video data. In: Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers, pp. 956–959. (2018)
8. Fan, Y., Lu, X., Li, D., Liu, Y.: Video-based emotion recognition using CNN-RNN and C3D hybrid networks. In: Proceedings of the 18th ACM International Conference on Multimodal Interaction, pp. 445–450 (2016)
9. Ghaleb, E., Popa, M., Asteriadis, S.: Multimodal and Temporal Perception of Audio-visual Cues for Emotion Recognition. In: 2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII). Cambridge, United Kingdom, pp. 552–558 (2019)
10. Ryumina E.V., Karpov A.A. Analytical review of methods for emotion recognition by human face expressions. Scientific and Technical Journal of Information Technologies, Mechanics and Optics. vol. 20, no. 2, pp. 163–176 (2020) (in Russian)
11. Alshamsi, H., Kepuska, V., Alshamsi, H., Meng, H.: Automated Facial Expression and Speech Emotion Recognition App Development on Smart Phones using Cloud Computing. In: 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON). Vancouver, BC, Canada, pp. 730–738 (2018)
12. Nasir, M., Jati, A., Shivakumar, P. G., Nallan Chakravarthula, S., Georgiou, P.: Multimodal and multiresolution depression detection from speech and facial landmark features. In: Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge, pp. 43–50 (2016)
13. Van Gent, P.: Emotion Recognition Using Facial Landmarks Python DLib and OpenCV. A tech blog about fun things with Python Embed. Electron (2016)

14. King, D. E.: Dlib-ml: A machine learning toolkit. Journal of Machine Learning Research. vol. 10, no. Jul, pp. 1755–1758 (2009)
15. Gite, B., Nikhal, K., Palnak, F.: Evaluating facial expressions in real time. In: 2017 Intelligent Systems Conference (IntelliSys). London, UK, pp. 849–855 (2017)
16. Al-Omair, O. M., Huang, S. A: Comparative Study of Algorithms and Methods for Facial Expression Recognition. In: 2019 IEEE International Systems Conference (SysCon). Orlando, FL, USA, pp. 1–6 (2019)
17. Ekman, P., Friesen, W. V.: Facial action coding system: Investigator's guide. Consulting Psychologists Press (1978)
18. Li, Q., Zhan, S., Xu, L., Wu, C.: Facial micro-expression recognition based on the fusion of deep learning and enhanced optical flow. In: Multimedia Tools and Applications. vol. 78, no. 20, pp. 29307-29322. Springer, (2019). https://doi.org/10.1007/s11042-018-6857-9
19. Jaratrotkamjorn, A., Choksuriwong, A.: Bimodal Emotion Recognition using Deep Belief Network. In: 2019 23rd International Computer Science and Engineering Conference (ICSEC), pp. 103–109 (2019)
20. Cao, H., Cooper, D. G., Keutmann, M. K., Gur, R. C., Nenkova, A., Verma, R.: CREMA-D: Crowd-sourced emotional multimodal actors dataset. IEEE Transactions on Affective Computing. vol. 5, no. 4, pp. 377–390 (2014)
21. Livingstone, S. R., Russo, F. A.: The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PloS One. vol. 13, no. 5, e0196391 (2018)
22. He, Z., Jin, T., Basu, A., Soraghan, J., Di Caterina, G., Petropoulakis, L.: Human emotion recognition in video using subtraction pre-processing. In: Proceedings of the 2019 11th International Conference on Machine Learning and Computing, pp. 374–379 (2019)
23. Siddiqui, M.F.H., Javaid A.Y: A Multimodal Facial Emotion Recognition Framework through the Fusion of Speech with Visible and Infrared Images. Multimodal Technologies and Interaction. vol. 4, no. 3:46, pp. 1–20 (2020)
24. Lucey, P., Cohn, JF., 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, San Francisco, CA, USA, pp. 94–101 (2010)
25. Valstar, M., Pantic, M.: May. Induced disgust, happiness and surprise: an addition to the mmi facial expression database. In: Proc. 3rd Intern. Workshop on EMOTION (satellite of LREC): Corpora for Research on Emotion and Affect, pp. 65–70 (2010)
26. Haq, S., Jackson, P.J.: Multimodal emotion recognition. Machine audition: principles, algorithms and systems, pp. 398–423 (2010)
27. Zhao, G., Huang, X., Taini, M., Li, S.Z., PietikäInen, M.: Facial expression recognition from near-infrared videos. Image and Vision Computing. vol. 29, no. 9, pp.607–619 (2011)
28. Perepelkina O., Kazimirova E., Konstantinova M.: RAMAS: Russian Multimodal Corpus of Dyadic Interaction for Affective Computing. Springer International Publishing. vol. 11096, pp. 501–510 (2018)
29. Kollias, D., Zafeiriou, S.: Aff-wild2: Extending the aff-wild database for affect recognition. arXiv preprint arXiv:1811.07770 (2018)
30. Breiman, L.: Random forests. Machine learning. vol. 45, no. 1, pp. 5–32 (2001)
31. Geurts, P., Ernst, D., Wehenkel, L.: Extremely randomized trees. Machine learning. vol. 63, no. 1, pp. 3–42 (2006)
32. Hastie, T., Rosset, S., Zhu, J., Zou, H.: Multi-class adaboost. Statistics and its Interface. vol. 2, no. 3, pp. 349–360 (2009)