Performance Evaluation of Emotion Recognition Methods on CK+ Dataset

Daniil Belkov, Konstantin Purtov, Vladimir Kublanov
Ural Federal University (UrFU)
Yekaterinburg, Russia
{d.d.belkov, k.s.purtov}@gmail.com, kublanov@mail.ru

***Abstract*— In this paper we evaluate performance of modern emotion recognition methods. Our task is to classify emotions as basic 8 categories: anger, contempt, disgust, fear, happy, sadness, surprise and neutral. CK+ dataset is used in all experiments. We apply Adaptive Boosting and Principal Component Analysis for dimensionality reduction and Support Vector Machine for classification. Size of train dataset is increased by use of few frames of sequences instead of one and vertical mirroring of faces. All images were normalized with mean centering and standardizing. In total 4428 images were used in experiment. The proposed method can work in real time and achieved average accuracy higher than 95%.**

I. INTRODUCTION

Human's facial expressions provide variable information about emotions and internal state of the person. Ability to automatically recognize facial expressions offers vast possibilities for video surveillance systems, systems that measure audience mood, public security and control systems, etc.

In past decade, a large number of systems for recognizing emotion [1,2,3] were developed with the use of modern methods of machine learning and high-quality databases, such as CK + [4], MMI [5], Jaffe [6].

These systems interpret the expression as one of the seven basic emotions (happiness, anger, contempt, disgust, sadness, surprise and fear). The interpretation is based on the analysis of facial expression using Facial Action Coding System (FACS) [7].

The purpose of this article is to create a system of monitoring of emotional state. To accomplish it we evaluate the accuracy of several emotion recognition algorithms.

II. RELATED WORK

A. Computer Emotion Recognition Toolbox (CERT)

CERT [2] is a fully automatic, real-time software tool that estimates facial expression both in terms of 19 FACS Action Units, as well as the 6 universal emotions.

This system uses its own face detector, which was trained using an extension of the Viola-Jones [9] approach. It employs GentleBoost as the boosting algorithm and WaldBoost for automatic cascade threshold selection. On the CMU+MIT dataset, CERT's face detector achieves a hit rate of 80.6%

After finding the face region, CERT estimates positions of 10 feature points: the corners of the eyes, eye centers, tip of the nose, the corners of the mouth and mouth center. Each facial feature detector, trained using GentleBoost, outputs the log-likelihood ratio of that feature being present at a location (x, y) within the face, to being not present at that location.

Given the set of 10 facial feature positions, the face patch is re-estimated at a canonical size of 96x96 pixels using an affine warp. The warp parameters are computed to minimize the L2 norm between the warped facial feature positions of the input face and a set of canonical feature point positions

The cropped 96x96-pixel face patch is then convolved with a filter bank of 72 complex-valued Gabor filters of 8 orientations and 9 spatial frequencies. The magnitudes of the complex filter outputs are concatenated into a single feature vector.

This feature vector is input to a separate linear support vector machine (SVM). SVM classifier provides distance of the input feature vector to the SVM's separating hyperplane, which can be interpreted as the intensity of certain facial movement.

To determine the 6 basic emotions and the neutral facial expression the classifier based on multivariate logistic regression was used. AU intensities and emotion ground truth labels were used for training classifier.

For this system the average accuracy is 0.93, the average F-measure is 0.79.

B. Facial expression recognition using radial encoding of local Gabor features and classifier synthesis

In this approach, [3] mouth and eye areas are manually labeled, then the face area of size 184 * 152 pixels is determined by these areas. Then, each image was divided into several local regions with a 50% overlap.

Next, to each of these local regions they apply the bank of Gabor filters with 3 scales and 8 orientations. The outputs of the filters were converted into a feature matrix.

Then, each filtered image was encoded by using a radial grid. The radial grid of resolution 18*5, with the center at the center of local region was used.

To reduce dimensionality of the feature matrix Principal Component Analysis and Fisher's linear discriminant were used.

For the classification of emotions K-nearest neighbor algorithm with k = 1 was used. The best result on the CK dataset was 91.51%.

In contrast to these methods, we do not use AU, because of the small number of labeled samples for training the robust classifier. Therefore, we take the last 20% of the frames of each sequence as samples emotions.

III. Algorithm

The proposed system for the classification of emotions consists of 6 main stages as shown in Figure 1: (A) face detection, (B) detection of the key points, (C) face frontalization and extraction, (D) normalization of face images, (E) dimensionality reduction (F) classification of emotions. Next we briefly describe each of the stages.

A. Face detection

First we have to find a face on each image in the database. We use Viola Jones approach [9]. This algorithm is widely used for face detection because it has high accuracy and can process images in real time.

B. Detection of the key points

The next stage is to find the key points in a given face region. We use face alignment tool based on the algorithm that uses cascades of regression trees [8]. This tool outputs location estimates of 68 key points: contour of the face, the nose, the outer and inner sides of the lips, eyes, eyebrows.

Given the initial constellation of the (x, y) locations of the 68 facial features, the location estimates are refined using linear regression. Outputs of the key points detector marked by white circles within the face in Figure 1.

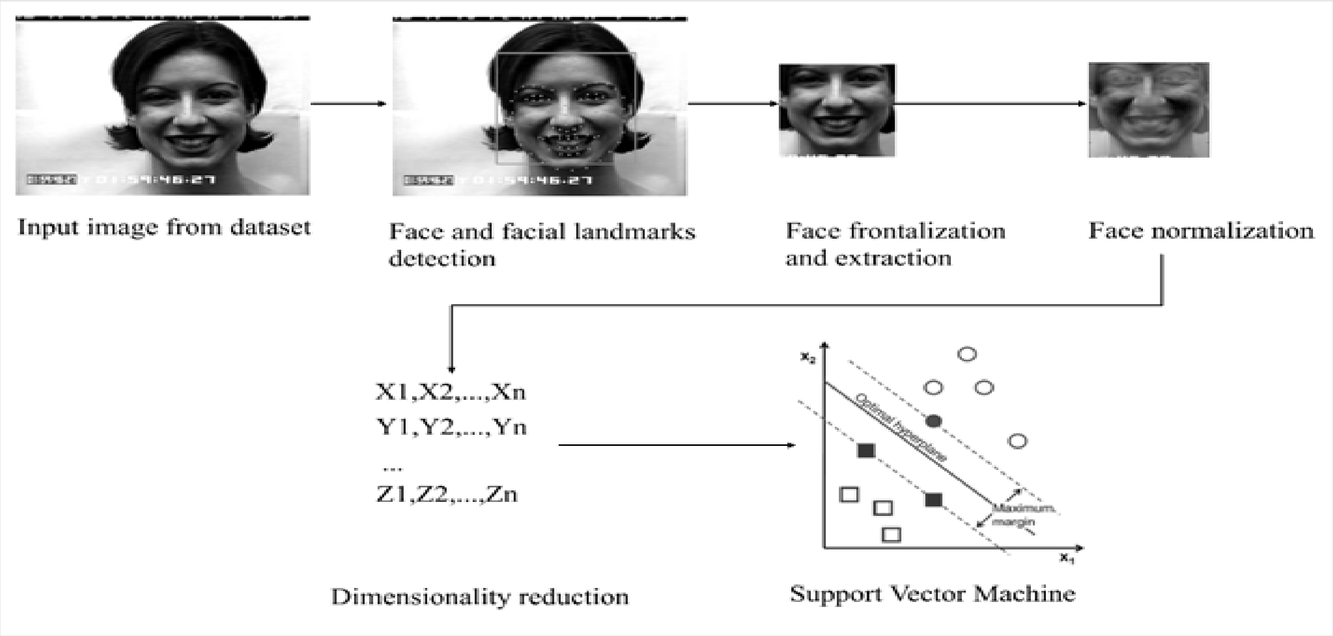

Fig. 1. Scheme of emotion recognition algorithm

C. Face frontalization and extraction

Given the face bounding rectangle and the key points we frontalize the face image with the affine transformation. Parameters of affine transformation are calculated to minimize the L2 norm between the warped facial feature positions of the input face and a set of canonical feature point positions.

Then we extract face patch size of 96 * 96 pixels from the original image. Each patch is then presented in the form of vector length 9216.

D. Normalization of the face image

Each resulting face image is normalized using the following strategy. First, from each pixel of the image we subtract the mean value of pixels in the image. Then, from each pixel we subtract the average value of that pixel on all the images in the database and divide the resulting value by the standard deviation of that pixel on all the images.

E. Dimensionality reduction

1) AdaBoost: AdaBoost algorithm trains a cascade of weak classifiers in iterative manner modifying weights of training samples. The weights are changed according to the correctness of object classification at the current stage. If the object was classified incorrectly, then its weight increases, otherwise reduces. The next predictor in the cascade will focus more on those objects that were incorrectly classified at the previous stage.

Decision stumps or decision trees are usually used as the weak classifiers. In this paper we use decision trees with tree depth equal to 4. The number of weak classifiers in a cascade is set to 50.

Since we use the decision trees as the weak classifiers, we can evaluate the importance of each feature, i.e. in this case the importance of values of each feature vector component. The basic idea is to evaluate the significance of each feature for separating in the tree nodes. In the case of decision trees with depth greater than 1 we determine the importance of each feature to a tree, then determine the final importance by averaging these values. Features in the top of the tree have a greater weight than the features used in the bottom, as they separate more objects.

Given the obtained importance of the features we can reduce the dimensionality of space by taking into account the most informative features.

In this study, we use all the features that have the importance greater than 0. The total number of such features is 348.

2) PCA: Principal component analysis (PCA) is a method that is used in data mining to reduce the dimensionality. Images often contain redundant information that has little importance in the analysis.

This method reduces the dimensionality of data and hence reduce the computational load during their processing, while losing a minimal amount of information.

First, training set is centered on the mean value. Then the covariance matrix is calculated.

From covariance matrix the Eigen values and eigenvectors are calculated. The eigenvector with the highest Eigen values is the principal component. Eigen values are ordered in ascending manner to form feature matrix. Eigenvectors with low Eigen values can be dropped. The principal component data set is done by multiplying transposed data set value and transposed feature vectors.

In this paper we used the PCA with number of principal components set to 348 for comparison with AdaBoost and with number of principal components set to 1800 as it provides the best results.

F. The classification of emotions

For the classification of emotions, we use a support vector machine algorithm [10]. This algorithm allows to separate data in a high dimensional space using hyperplane located at a maximum margin of all classes.

We use SVM which implements the “one-against-one” classification strategy. In this approach, the classifier decides for each pair of classes. The final classification is made by voting. The "RBF" function is selected as a core function of the classifier.

IV. Database

CK+ dataset consists of a frame sequences. In each sequence there is a person performing a certain facial expression. Resolution of all frames is 640 * 480 pixels.

Participants were 18 to 50 years of age, 69% female, 81%, Euro-American, 13% Afro-American, and 6% other groups. CK + contains the sequences for 123 subjects. The sequence length varies from 10 to 60 frames. Each of the sequences contains images from onset (neutral frame) to peak expression (last frame). The peak frame was reliably FACS coded for facial action units.

Emotion label is based on the subject’s impression of each of the 7 basic emotion categories: anger, contempt, disgust, fear, happy, sadness and surprise.

Labels are defined according to the FACS codes and the Emotion Prediction Table from the FACS manual. In total 327 sequences have emotion labels. We use these sequences in the experiment.

V. Experiment

We divide the CK+ dataset as follows: first 6% of sequence of frames used as a neutral facial expression, last 20% used as a facial expression that represents a certain emotion. This approach allows us to increase size of dataset. After extraction all samples were reviewed manually to exclude any class overlapping

Furthermore, each image was mirror reflected by vertical axis. Such approach provides increasing the size of dataset and increasing the robustness of the algorithm. In the end, there are 4428 face images in the dataset.

To obtain different sets we randomly shuffled and divide our set to training and testing subsets 10 times. Subsets divided at a ratio of 70:30. Each training subset contains 3104 images and testing subset contains 1324 images.

To evaluate performance of proposed algorithms we use accuracy and F_1-score. All classifier's decisions may be divided into four groups:

- TP — true positive decision;
- TN — true negative decision;
- FP — false positive decision;
- FN — false negative decision.

Accuracy can be found as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

F-measure can be found as:

$$F_1 = 2 \cdot \frac{precision \times recall}{precision + recall}$$

Here precision and recall are

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

VI. Results

Since the experiment was performed on 10 different subsets, we present the average results among all datasets for all algorithms used in experiment. Results shown in Tables I, II, III for various methods of reducing dimension.

The results show that the emotion of contempt has the lowest recognition rate in all three cases. This is probably due to the fact that the source database sequences have very few images with that emotion. The emotions of happiness give the best recognition rate by F_1-score. The emotions anger, fear and surprise have slightly lower score than happiness, but better than 0.95 for AdaBoost classification. The neutral facial expression state had the lowest F_1-score value in all three cases, because it is intersect with all cases of emotions.

The worst results for all categories of emotions for F_1-score is obtained by SVM with first 328 components of PCA presented in Table II. In this case the best average F_1-score is 0.82, the worst is 0.38 which is very small.

To verify the superiority of AdaBoost over PCA for important feature selection, we increase the count of first PCA components trying get better results. It is increased by about 5 times to 1800 components. The obtained results is much better than for 348 PCA components, but not such good as AdaBoost.

In PCA cases the intersection between the basics emotions is minimal (excluding neutral), This is possible because of our dataset is small, and people in one emotion state not present in the another. So it can be used as additional condition for checking the category of emotions for small dataset.

VII. Conclusion

We evaluated the accuracy of the proposed algorithms of emotion classification using various evaluation criteria.

Two methods show sufficiently high accuracy. Given the obtained values of the accuracy and F_1-score criteria we can consider these algorithms as state-of-the-art.

The main difference between the algorithms is in use of various approaches to reduce the dimensionality. AdaBoost algorithm does better with dimensionality reduction as it outputs the lesser number of components which provide a high classification accuracy results. Consequently, less time is required to classifier learning and, thus, less time for emotion classification in testing and working stages.

In our future work, we plan to increase the database and improve the quality of classification by using the convolution neural networks, as well as the integration of the system with the analysis of physiological state by video.

TABLE I. RESULTS OF SVM CLASSIFICATION USING ADABOOST WITH 348 COMPONENTS

	Anger	Contempt	Disgust	Fear	Happy	Sad	Surprise	Neutral	Average
Anger	98,42	0	0	0	0	1,97	0	0,04	
Contempt	0	76,88	0	0	0	0	0,73	0,04	
Disgust	0,08	0	98,14	0	0	0	0	0	
Fear	0,08	0	0	100	0,23	0	0	0,04	
Happy	0	0	0	0	99,77	0	0	0	
Sad	0,15	0	0	0	0	86,76	0	0,15	
Surprise	0	0	0	0	0	0	98,65	0	
Neutral	1,28	23,13	1,86	0	0	11,27	0,62	99,74	
Precision	0,98	0,99	1,00	1,00	1,00	1,00	1,00	0,72	0,96
Recall	0,98	0,77	0,98	1,00	1,00	0,87	0,99	1,00	0,95
F_1-score	0,98	0,87	0,99	1,00	1,00	0,93	0,99	0,84	0,95
Accuracy	99,53	96,95	99,74	99,96	99,97	98,25	99,82	95,18	98,67

TABLE II. RESULTS OF SVM CLASSIFICATION WITH USING FIRST 348 PCA COMPONENTS

	Anger	Contempt	Disgust	Fear	Happy	Sad	Surprise	Neutral	Average
Anger	60,60	0	0	0	0	0	0	0	
Contempt	0	50,63	0	0	0	0	0	0	
Disgust	0	0	35,85	0	0	0	0	0	
Fear	0	0	0	59,44	0	0	0	0	
Happy	0	0	0	0	69,94	0	0	0	
Sad	0	0	0	0	0	45,35	0	0,09	
Surprise	0	0	0	0	0	0	49,33	0	
Neutral	39,40	49,38	64,15	40,56	30,06	54,65	50,67	99,91	
Precision	1,00	1,00	1,00	1,00	1,00	1,00	1,00	0,23	0,90
Recall	0,61	0,51	0,36	0,59	0,70	0,45	0,49	1,00	0,59
F_1-score	0,75	0,67	0,53	0,75	0,82	0,62	0,66	0,38	0,65
Accuracy	92,28	90,51	88,01	92,07	94,00	89,59	90,29	58,88	86,95

TABLE III. Results of SVM classification with using first 1800 PCA components

	Anger	Contempt	Disgust	Fear	Happy	Sad	Surprise	Neutral	Average
Anger	91,95	0	0	0	0	0	0	0	
Contempt	0	65,63	0	0	0	0	0	0,46	
Disgust	0	0	88,22	0	0	0	0	0,22	
Fear	0	0	0	85,77	0	0	0	0,02	
Happy	0	0	0	0	97,50	0	0	0	
Sad	0	3,75	0	0	0	78,59	0	0,31	
Surprise	0	0	0	0	0	0	94,94	0	
Neutral	8,05	30,63	11,78	14,23	2,50	21,41	5,06	98,99	
Precision	1,00	0,99	1,00	1,00	1,00	0,95	1,00	0,51	0,93
Recall	0,92	0,66	0,88	0,86	0,98	0,79	0,95	0,99	0,88
F_1-score	0,96	0,79	0,94	0,92	0,99	0,86	0,97	0,68	0,89
Accuracy	98,87	95,27	98,32	98,01	99,64	96,50	99,28	88,11	96,75

Acknowledgment

We would like to thank the reviewers for their comments, and colleagues of the Research Medical and Biological Engineering Center of High Technologies, UrFU.

This work was supported by Act 211 Government of the Russian Federation, contract 02.A03.21.0006 and partially supported by UrFU scholarship for outstanding achievements in science and Russian Foundation for Assistance to Small Innovative Enterprises (FASIE).

References

[1] Peng Yang, Qingshan Liu, and Dimitris N. Metaxas. Boosting encoded dynamic features for facial expression recognition. Pattern Recognition Letters, 30:132–139, 2009.

[2] Gwen Littlewort1, Jacob Whitehill1, Tingfan Wu1, Ian Fasel2, Mark Frank3, Javier Movellan1, and Marian Bartlett1. The Computer Expression Recognition Toolbox (CERT). Automatic Face & Gesture Recognition and Workshops.2011

[3] Wenfei Gu, Cheng Xiang, Y.V.Venkatesh, Dong Huang ,Hai Lin. Facial expression recognition using radial encoding of local Gabor features and classifier synthesis. Pattern recognition. 2011

[4] 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 Computer Vision and Pattern Recognition Workshop on Human-Communicative Behavior, 2010.

[5] Maja Pantic, Michel Valstar, Ron Rademaker, and Ludo Maat. Webbased database for facial expression analysis. In International Conference on Multimedia and Expo, 2005.

[6] M.Kamachi,M.Lyons,J.Gyoba,TheJapaneseFemaleFacialExpression (JAFFE) Database. [Online]. Available: http://www.kasrl.org/jaffe.html.

[7] P. Ekman and W. Friesen. The Facial Action Coding System: A Technique For The Measurement of Facial Movement. Consulting Psychologists Press, Inc., San Francisco, CA, 1978.

[8] Vahid Kazemi, Josephine Sullivan. One Millisecond Face Alignment with an Ensemble of Regression Trees. CVPR. 2014

[9] Paul Viola and Michael Jones. Robust real-time face detection. International Journal of Computer Vision, 2004.

[10] Chang, Chih-Chung and Lin, Chih-Jen. LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology. 2011