A Novel Dataset for Real-Life Evaluation of Facial Expression Recognition Methodologies

Muhammad Hameed Siddiqi[1], Maqbool Ali[2], Muhammad Idris[2], Oresti Banos[2], Sungyoung Lee[2], and Hyunseung Choo[1(✉)]

[1] Department of Computer Science and Engineering, Sungkyunkwan University, Suwon, Korea
{siddiqi,choo}@skku.edu
[2] Department of Computer Engineering, Kyung Hee University, Suwon, Korea
{maqbool.ali,idris,oresti,sylee}@oslab.khu.ac.kr

Abstract. One limitation seen among most of the previous methods is that they were evaluated under settings that are far from real-life scenarios. The reason is that the existing facial expression recognition (FER) datasets are mostly pose-based and assume a predefined setup. The expressions in these datasets are recorded using a fixed camera deployment with a constant background and static ambient settings. In a real-life scenario, FER systems are expected to deal with changing ambient conditions, dynamic background, varying camera angles, different face size, and other human-related variations. Accordingly, in this work, three FER datasets are collected over a period of six months, keeping in view the limitations of existing datasets. These datasets are collected from YouTube, real world talk shows, and real world interviews. The most widely used FER methodologies are implemented, and evaluated using these datasets to analyze their performance in real-life situations.

Keywords: Facial expression recognition · Feature extraction · Feature selection · Recognition · YouTube · Real-world

1 Introduction

Existing FER methodologies utilized previous datasets and did not consider the real world challenges in their respective systems. For instance, two most commonly used expression datasets used for evaluating FER systems are Cohn-Kanade (CK) [5] dataset, and JAFFE dataset [7]. JAFFE dataset is collected from 10 different subjects (Japanese female), where CK dataset is collected from 97 subjects (university students). Both datasets are collected under controlled laboratory settings with constant lighting effects, camera stetting, and

H. Choo—Supported by the MSIP, Korea, under the G-ITRC support program (IITP-2015-R6812-15-0001) supervised by the IITP, and by the Priority Research Centers Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (NRF-2010-0020210).

R. Khoury and C. Drummond (Eds.): Canadian AI 2016, LNAI 9673, pp. 89–95, 2016.
DOI: 10.1007/978-3-319-34111-8_12

background. All of the images are taken from the frontal view of the camera, with tied hair in the case of JAFFE dataset, in order to expose all the sensitive regions. Furthermore, these datasets are pose-based, i.e., subjects performed the expressions exactly when they are asked to. Limited efforts have been put into designing a new dataset that is closer to real-life situations, probably because creating such a dataset is a very difficult and time consuming task. And this is where the contribution of this work lies.

Accordingly, in this work, we have defined a realistic and innovative dataset collected from YouTube, some real world talk shows, and some interviews that considered the above-mentioned limitations. From lab settings to a real-life environment, we defined three cases with increasing complexity. In all three cases, a large number of different subjects of different gender, race, and age were included. Also, the defined datasets have various sizes of the face that is related to proximity. From existing works, more recent methodologies were implemented.

2 Existing Standard FER Methodologies and Datasets

Standard Methods: For feature extraction, LBP [4], LDP [8], curvelet transform [14], and wavelet transform [10]; for feature selection, (LDA) [9], kernel discriminant analysis (KDA) [17], and generalized discriminant analysis (GDA) [15]; and for recognition, SVM(s) [8], HMM(s) [9] and HCRF(s) [11] were used.

Existing Datasets: The extended version of CMU-PIE dataset [12] was collected named Multi-PIE [3] that covered the limitations of CMU-PIE. However, multi-PIE is a pose-based dataset collected under a static illumination conditions. Similarly, the extended Cohn-Kanade (CK+) dataset [6] is the extension of CK dataset [5] which covered the limitations of CK dataset. This dataset consists of both pose-based and spontaneous expressions. However, this dataset has been collected under a controlled environment and though some subjects were at a 30-degree angle with the camera, the remaining subjects were with frontal view to the camera. Georgia Tech Face dataset [1] contains images of 50 people, which is a pose-based dataset and the images show frontal and/or tilted faces with different expressions. USTC-NVIE [16] consists of both pose and spontaneous expressions collected by more than 100 people. However, a visible and infrared thermal camera was used for dataset collection with a predefined lighting setup. Another important real world dataset named VADANA: Vims Appearance dataset [13] was collected to consider the research problem of gender and age in the area of FER. However, this dataset is a pose-based dataset, too. Likewise, another real world dataset named YMU (YouTube Makeup) dataset [2] was collected from the YouTube makeup tutorials consisting of 151 subjects. However, in this dataset, only females (having makeup) are involved that may cause the gender problem.

3 Novel Datasets to Benchmark Real-World FER Systems

Emulated Dataset: In this dataset, the ordinary subjects performed expressions in a pose-based manner in a controlled lab environment. The subjects belonged to different colors, age, and ethnicities. The subject age ranges from 4 years to 60 years. In some of the cases, the images in some expressions are rotated using the camera for better accuracy of the system. The subjects include both males and females. Each expression has at least 165 images. The images used in the dataset are of size 240×320 and 320×240 pixels with facial frame.

Semi-naturalistic Dataset: In this dataset, the expressions are collected from the actors and actresses of Hollywood and Bollywood in their respective movies, where we had no control on expression timings, camera, lighting and background settings. The expressions have different views from different angles with glasses, hair open and close, and other obvious actions are collected in this dataset with dynamic settings. Each expression consists of at least 165 images. The dataset has the images of size 240×320 and 320×240 pixels with facial frame.

Naturalistic Dataset: In this dataset, subjects from various parts of the world, races, and ethnicities have been selected. The expressions are spontaneous that have been captured in natural and dynamic settings from real world talk-shows, interviews, and YouTube natural videos such as news and real world incidents. The total of 165 images have been considered for each expression. The age range of the subjects are from 18 to 50 years. Images used in the dataset are of size 240×320 and 320×240 pixels with facial frame. All the datasets include six basic expressions such as happy, sad, angry, normal, disgust, and fear. These datasets will be made available for future research to the research community.

4 Experimental Results

A comprehensive set of experiments were performed, in which the performance of each method was tested and validated using 10-fold cross-validation rule for each dataset. All the experiments were performed in Matlab using an Intel® Pentium® Dual-Core^{TM} (2.5 GHz) with a RAM capacity of 3 GB.

Experimental Analysis Using Emulated Dataset: The average recognition rates for each method when benchmarked on the emulated dataset are shown in Fig. 1 (first image). As it can be seen, the vast majority of the evaluated methods yield an average accuracy within the range 65 to 80 %, thus far from perfect recognition capabilities. Some combinations of feature extraction and selection methods provide better results than others, especially, LBP + KDA, LDP + GDA, Curvelet + KDA and Wavelets + KDA and GDA. The sort of feature extraction and selection technique used in the FER model turns to have a less clear impact than classification paradigm. In fact, highest accuracies are generally obtained by using HCRF, while poorest results are obtained for systems based on SVM.

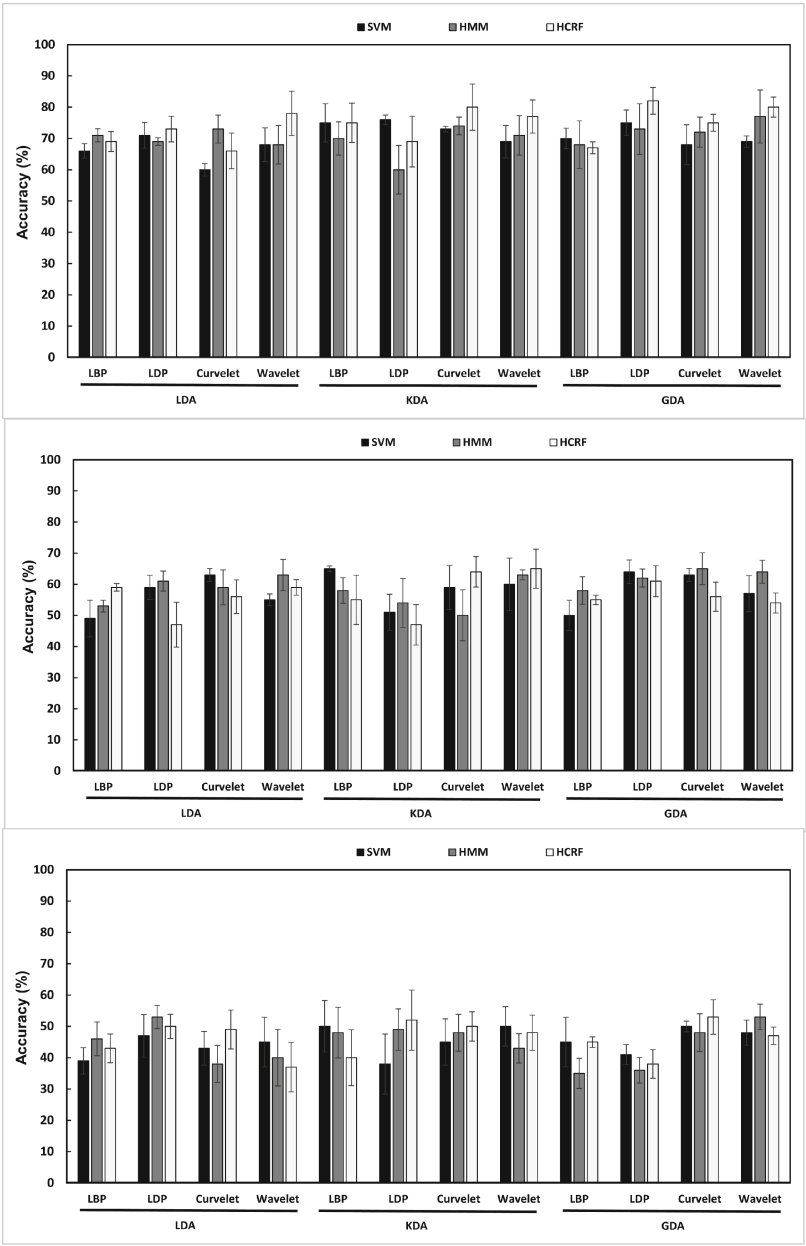

Fig. 1. The average (bar and standard deviation whiskers) classification rates from the evaluation of the standard FER methods using the defined emulated (first image), semi-naturalistic (middle image), and naturalistic (last image) datasets. The top legend presents the recognition paradigm. The horizontal axis labels present the standard feature extraction methods used for each experiment, while, the underlined shows respectively the standard feature selection methods.

Experimental Analysis Using Semi-naturalistic Dataset: Figure 1 (middle image) depicts the performance values of each standard method for the semi-naturalistic case. At first sight, a significant drop in the performance is observed with respect to the ideal scenario. Here, the performance of the evaluated models span from 45 % to 65 %, which is unacceptable for realistic FER applications. The combinations of feature extraction and selection methods that yield best results are different to the ones highlighted for the emulated case. Concretely, Wavelet + KDA and LDP + GDA seem to provide the best performance for all classification paradigms. Conversely to the emulated scenario, no classification paradigm is observed to prevail over the others for the semi-naturalistic case.

Experimental Analysis Using Naturalistic Dataset: The accuracy results corresponding to the third evaluation scenario, i.e., the naturalistic case, are shown in Fig. 1 (last image). As it could be expected, the performance of all models is dramatically reduced with respect to the emulated or ideal scenario, with accuracies that range between less than 40 % to 55 %. Although marginalizing across feature extraction, selection, and classification techniques is of arguable value given the low accuracy values, it may be said that best combinations are LDP + LDA, and Curvelet/Wavelet + KDA/GDA. Similarly, no clear conclusions can be derived from the analysis of the prevalence of the classification models, although highest results tend to be obtained by using the HCRF. Despite, the combinations of feature extraction and selection methods that yield best results are different to the ones highlighted for the emulated case.

5 Conclusion

Human FER has emerged as a fascinating research area during the last two decades. However, accurate FER in real world scenarios is still a challenging work. Most of the previous FER methodologies achieved high recognition rate using all the previous datasets. However, most of these datasets were collected under predefined setups. And, these methodologies showed poor performance when applied on real world datasets. Several factors that effects the accuracy of the FER methodologies include varying light conditions and dynamic variation of the background.

In this work, we have defined three kinds of datasets named emulated, semi-naturalistic, and naturalistic datasets. The defined datasets considered most of the limitations of the existing datasets in real world scenarios. These datasets are collected from real world talk shows, interviews, and YouTube. We have evaluated some well-known existing standard FER methodologies using the defined datasets. All the standard methodologies were tested and validated using 10-fold cross-validation rule. It can be seen that all the methodologies showed least performance on semi-naturalistic and naturalistic datasets.

Therefore, it is desirable that in future we will propose new methods to improve the accuracy of FER systems in real-life scenarios.

References

1. Chen, L., Man, H., Nefian, A.V.: Face recognition based on multi-class mapping of fisher scores. Pattern Recogn. **38**(6), 799–811 (2005)
2. Dantcheva, A., Chen, C., Ross, A.: Can facial cosmetics affect the matching accuracy of face recognition systems? In: 2012 IEEE Fifth International Conference on Biometrics: Theory, Applications and Systems (BTAS), pp. 391–398. IEEE (2012)
3. Gross, R., Matthews, I., Cohn, J., Kanade, T., Baker, S.: Multi-pie. Image Vis. Comput. **28**(5), 807–813 (2010)
4. Hablani, R., Chaudhari, N., Tanwani, S.: Recognition of facial expressions using local binary patterns of important facial parts. Int. J. Image Process. (IJIP) **7**(2), 163 (2013)
5. Kanade, T., Cohn, J.F., Tian, Y.: Comprehensive database for facial expression analysis. In: Fourth IEEE International Conference on Automatic Face and Gesture Recognition, Proceedings, pp. 46–53. IEEE (2000)
6. 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 (CVPRW), pp. 94–101. IEEE (2010)
7. Lyons, M., Akamatsu, S., Kamachi, M., Gyoba, J.: Coding facial expressions with gabor wavelets. In: Third IEEE International Conference on Automatic Face and Gesture Recognition, Proceedings, pp. 200–205. IEEE (1998)
8. Rivera, A.R., Castillo, R., Chae, O.: Local directional number pattern for face analysis: face and expression recognition. IEEE Trans. Image Process. **22**(5), 1740–1752 (2013)
9. Siddiqi, M.H., Ali, R., Idris, M., Khan, A.M., Kim, E.S., Whang, M.C., Lee, S.: Human facial expression recognition using curvelet feature extraction and normalized mutual information feature selection. Multimedia Tools Appl. **75**(2), 935–959 (2016)
10. Siddiqi, M.H., Ali, R., Khan, A.M., Kim, E.S., Kim, G.J., Lee, S.: Facial expression recognition using active contour-based face detection, facial movement-based feature extraction, and non-linear feature selection. Multimedia Syst. **21**(6), 541–555 (2015)
11. Siddiqi, M.H., Ali, R., Khan, A.M., Park, Y.-T., Lee, S.: Human facial expression recognition using stepwise linear discriminant analysis and hidden conditional random fields. IEEE Trans. Image Process. **24**(4), 1386–1398 (2015)
12. Sim, T., Baker, S., Bsat, M.: The cmu pose, illumination, and expression database. IEEE Trans. Pattern Anal. Mach. Intell. **25**(12), 1615–1618 (2003)
13. Somanath, G., Rohith, M.V., Kambhamettu, C.: Vadana: a dense dataset for facial image analysis. In: 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), pp. 2175–2182. IEEE (2011)
14. Tang, M., Chen, F.: Facial expression recognition and its application based on curvelet transform and pso-svm. Optik-Int. J. Light Electron Opt. **124**(22), 5401–5406 (2013)
15. Uddin, M.Z., Kim, T.-S., Song, B.C.: An optical flow feature-based robust facial expression recognition with hmm from video. Int. J. Innovative Comput. Inf. Control **9**(4), 1409–1421 (2013)

16. Wang, S., Liu, Z., Lv, S., Lv, Y., Wu, G., Peng, P., Chen, F., Wang, X.: A natural visible and infrared facial expression database for expression recognition and emotion inference. IEEE Trans. Multimedia **12**(7), 682–691 (2010)
17. Wu, Q., Zhou, X., Zheng, W.: Facial expression recognition using fuzzy kernel discriminant analysis. In: Wang, L., Jiao, L., Shi, G., Li, X., Liu, J. (eds.) FSKD 2006. LNCS (LNAI), vol. 4223, pp. 780–783. Springer, Heidelberg (2006)