Review of Different Combinations of Facial Expression Recognition System

Abd_Almuhsen, F. Almudhafer¹, Zainab A. Khalaf²

¹ Basrah University, College of Education for Pure Sciences, Department of Computer Science, Basrah, Iraq
² Basrah University, College of Science, Department of Mathematics, Basrah, Iraq

¹ abdalmudhafer@gmail.com
² zainab.ali2004@yahoo.com

Abstract. The facial expression recognition (FER) system is a classifier system that attempts to recognize facial expressions based on the analysis of emotion behaviour on the face. The FER system can be implemented by using one classifier or combining multi feature extraction and/or multi classifiers. In general, FER is used with one classifier system to find the best label. Although a classification system is commonly used to find the most likely facial expression, it still produces substantial numbers of errors due to several factors that influence the FER result, such as data quantity, and environmental conditions (i.e. illumination and noise). Therefore, combined multi feature extraction methods and/or multi classifier systems are useful to avoid the single classifier errors. Multi feature extraction or a multi classifier system combination are used to take advantage of different system hypotheses to find an accurate result. This paper is a survey of the latest system combination techniques being used to enhance the classification performance in the FER system; the most recent studies are presented.

1. Introduction

The human face reflects internal feelings in an immediate time frame, due to its important role in human personal communication. After viewing the face, the identification of the person, sex, expression, etc. will be recognized. Thus, facial expressions provide sensitive signals about feelings, and play a major role in human interaction and nonverbal communication [1]. Although there is a wide range of possible facial expressions, psychologists have identified six fundamental ones (happiness, sadness, surprise, anger, fear and disgust) that are universally recognized. It is obvious that a system capable of performing automatic recognition of human emotions is a desirable task for a set of applications such as human-computer interaction,
security, affective computing, etc. Figure 1 is an example of these six fundamental expressions in addition to a normal facial expression.

Figure 1: An example of basic facial expressions

Recently many papers have proposed the combination of more than one classifier when designing systems for pattern classification with high performance that work in a hybrid way. The reason caused the increased need in multi-classifier systems is the acknowledgement that the classic methods for implementing a pattern recognition system that directed to select the best classifier, suffer from some basic drawbacks. The basic drawback is the hardness in selecting the suitable classifier for the classification task, unless a deep former knowledge on the data is available. In addition, using one classifier won’t let the exploitation of the complete distinctive data to be encapsulated by other classifiers [1]. The vision behind a combined classifiers is to merge multiple machine learning technique, in order to improve the system performance as best as available. In other words, a combination approach basically consists by two, three or more classifiers. For example, the first classifier takes the inputted data and generates initial findings. The second one takes those findings as the input and produce the final results [2]. Results based aggregation can be used if more than two classifiers are used to have the best results that represent the collective opinion of the whole system [3]. In facial expression recognition systems, a combination is used with more than one approach that seeks the general goal of the highest accuracy that can be reached. Two main approaches to be discussed are the prior-combination and the post-combination [4], as shown in figure 2.

Figure 2: Typical system for FERS with combination

Prior-combination works on the feature extraction phase, which means that it might be a combined extraction method or a combined extracted features. Both ways will lead to a new set of features different to the initial sets. And many classifiers can be used to detect and extract features from the face such as local binary pattern, Hidden Markova model, the AdaBoost classifiers [5], Principle component analysis and eigenvectors [6]. The post-combination works on either enhancing the classifier results with the assistance of a second classifier or
works on gathering more than one result in order to vote on the most frequent results to be chosen as the final result.

![Figure 3: Typical Prior-combination System](image)

Figure 3 demonstrates a typical scenario of the prior-combination that can occur when using two different extraction methods after the pre-processing step. This will lead to extraction of a two feature set that can be combined to provide a new set with more distinctive aspects.

![Figure 4: Typical Post-combination System](image)

In a similar, but not exactly the same method, classifiers are being combined by using more than one classifier with the same features set. Each result from those classifiers are used as a vote to select the best result as detailed in figure 4. In another scenario, classifiers are combined to assess in handling part of the tested data, the first classifier will recognize part of the tested data according to a predefined condition and pass the remaining to the second classifier while the results are sent to the result aggregation pool, as detailed in figure 5. A study to compare classifier combination strategies was presented in Kuncheva’s work [7]. Techniques like average, minimum, maximum, median, and majority votes were detailed in a theoretical way.

![Figure 5: An example of Post-combination](image)

The remaining parts of this paper are in four sections: Prior-Combination for features extraction and Post-Combination Classifier of recent studies are detailed in sections 2 and 3. Discussions and the conclusions follow in sections 4 and 5, respectively.
2. Prior-Combination for Features Extraction

Several studies, where more than one methodology was used, worked on extracting features with the best distinctive rate. For example, Sun et al. [8] proposed a robust facial expression recognition approach, directed to find the region of interest in the face to train a robust face-specific of Convolutional Neural Networks (CNN). By using the similar aspects between the facial areas within the ROI, an architecture stated aims to enhance the performance while predicting targets. The proposed architecture depends on deep learning to fine tune the new model of neural networks, the fine tune step is made according to a previously trained deep network to achieve the required performance. Also, the researchers worked to increase and improve the training process in the deep CNN by using data augmentation strategy. Kalsam et al. [9] designed a combination for a spatial bag of features (SBoFs) with spatial scale-invariant feature transform (SBoF-SSIFT). The SBoFs descriptor generated a feature vector with fixed length for all images used irrespective of their size and the SSIFT was designed by combining scale-invariant feature transform (SIFT) with speeded up robust transform features (SURF). The (SBoF-SSIFT) speeded up the transform process and enhanced the recognition ability of facial expressions. Because those features are independent of rotation, scale, translation, projective transforms, and when partial illumination might occur. For the recognition phase, K-Nearest Neighbor and Support vector machines were used. In another use of the bag of features combination, a feature extraction method was proposed by Sun and Lv [10] to be used in the facial expression recognition from a single image frame. The hybrid features used a combination of SIFT and deep learning features of different extraction level from the images, by using a trained CNN model. Also, Mahmood et al. [11] worked on facial variations and the complexity of appearance. In their study, the researchers attempted to improve the system accuracy by using Radon transform and Gabor wavelet transform. Facial detection was tested by the oval parameter method and facial tracking was achieved by implementing vertex mask generation. Radon transform and Gabor transform filters were applied to extract a variable set of features. Finally, self-organized maps with neural networks were used as the recognizing engine to measure the six facial expressions. Other researchers worked on using the distances between the facial landmarks, and on a triangular structure induced with three points in circumcenter, in center and centroid, which was considered as the geometric primitive to extract the required features. Information gained from those features were used for the task of discrimination of expressions using the Multilayer Perceptron (MLP) classifier in images containing facial expressions available in more widely known databases [12]. On the other hand, Sun worked on the idea of extracting optical flow from the changes between the highest expression intensity in the face image and the normal expression face image as the temporal information of a facial expression. He used the grey image of the facial expression as the spatial information. Also, a multi-channel Deep Spatial-Temporal feature Fusion neural Network (MDSTFN) was presented to perform the spatiotemporal deep feature extraction with fusion from the static images [13].

3. Post-Combination Classifier System

On the other side of the combination diagram, the classifier is combined for the recognition task and to enhance the classifier performance to recognize features with high or low distinctively. Jain et al. [14] proposed combining sequential information while using Recurrent Neural Network (RNN) to spread gained information. The CNN model was used for the extraction of features in order to fix all of the CNN parameters and to be able to eliminate the regression layer. For the processing, after the image passes to the network, 200-dimensional vectors are extracted from the fully-connected layers. All vectors will go through a node of the RNN; and finally, all the nodes of the RNN returns a results of valence label. On the other hand, Vo and Lee [15] Present a new methodology works on the hierarchical representation called hierarchical collaborative representation-based classification (HCRC). The researchers
depend on a classifier designed with two stages, first is to use deep convolutional neural network (DCNN) which works on extracting distinctive features from the image. And the second stage is to combine the HCRC with a model of local ternary patterns (LTP) in order to enhance the classifier performance to be robust with noisy conditions. Kar and Babu worked on proposing a combination system with both features and recognition to classify facial expressions; the proposed system has three steps. In the beginning, ripple transform type II is used to find the features from the facial area in the image. This approach is efficiency and working with both edges and textures. The second step, a principal component analysis (PCA) along with linear discriminant analysis (LDA) method are used to get more discriminative features. In the final step, the recognition is done by the least squares variant of support vector machine (LS-SVM) with a radial based function (RBF) kernel as demonstrated in [16]. Another use of the CNN combination to figure out the video copy might happen by suggesting a 3D-CNN architecture with parallel condition to handle multi-class recognition by implementing one 3D-CNN in combination with multiple two-class classifier. Eventually, the 3D-CNN is proposed as the two classes’ classifier in any of the two-class classification [17].

4. Discussions

In order to understand the differences between the methods used to design a FER system, the results of the methods mentioned are summarized as a compression in the table below in terms of method, type, database and accuracy. There are many databases used in the FER system, such as: the Japanese Female Facial Expression (JAFFE) [18], the Extended Cohen Kennedy (CK+) [19], the Facial Expression Database 2013 (FER-2013) [20], Media Understanding Group (MUG) [21] and many other databases can be found in the literature.

| Ref No. | Year | Method Proposed | Combination Type | Databases | Accuracy |
|---------|------|----------------|------------------|-----------|----------|
| 8       | 2020 | ROI partitioning | Prior            | JAFFE, CK+, FER-2013, WCD | 53.77%, 94.67%, 40.13%, 37.25% |
| 9       | 2018 | Hybrid feature descriptors | Prior | CK+, JAFFE | 98.5%, 98.3% |
| 10      | 2019 | Combining deep and shallow features | Prior | CK+, JAFFE, MMI | 94.13%, 48.9%, 53.81% |
| 11      | 2018 | 1D transform and GWT | Prior | CK, AT&T | 86%, 83.7% |
| 12      | 2020 | Landmark triangulation | Prior | CK+, JAFFE, MMI, MUG | 99.08%, 97.18%, 96.87%, 97.26 |
| 13      | 2019 | Deep spatial-temporal feature fusion | Prior | CK+, MMI, RaFD | 98.38%, 99.17%, 99.59 |
| 14      | 2018 | Hybrid deep neural networks | Post | JAFFE, MMI | 94.91%, 92.07% |
| 15      | 2018 | Hierarchical collaborative representation | Post | AR, Yale | 99.9%, 99.5% |
| 16      | 2019 | Ripplet II + LS SVM | Post | JAFFE, CK+ | 99.46%, 98.97% |

Table 1: A compression of prior and post-combinations

This study focuses on the effect of using combination system either prior- or post-combination system instead of using individual system in facial expression recognition. In order to show the advantages of combination system in enhancing the facial expression results, the experimental results of the recent related works are compared. For the CK+ data set, the
researchers in [22] used SIFT shallow features to extract the facial features, and the accuracy was 79%. On the other hand, the study in [23] extracted the deep emotions features in feature extraction stage, the obtained accuracy was 80.07%. While, the study of [10] used the same data set by applying the combination of deep and shallow features. The results were 94.82% as accuracy. Furthermore, Zhang et al. [24] used SIFT features for feature extraction on the CK+ dataset, and their accuracy was 95.8%. Whereas, after combining a spatial bag of features (SBoFs) with spatial scale-invariant feature transform (SBoF-SSIFT) in research [9], the accuracy enhanced to 98.5% on the same dataset. For the mix of two datasets: MMI and JAFFE, Jain et al. [14] used CNN as classifier system and the obtained accuracy was 76.51%. Then, the researchers combined two deep learning classifiers CNN-RNN, and the accuracy enhanced to 91.20%. In addition, the researchers used the Hybrid CNN-RNN model with the ReLU, and a significant performance achieved with 94.46%. Support Vector Machine (SVM) facial expression classification system has been used by Sohail and Bhattacharya [25]. In this study, fifteen different feature points used to face identification. The obtain accuracy of 92% on the JAFFE dataset, and 86.33% on the CK dataset. These facial expression classification results using SVMs as a classifier positively illustrate the strengths of SVMs for emotion recognition. In another study, CNN with five convolutional layers to extract a feature vector was used by Ouellet [26], and then the vector feed to a SVM for classification. The researcher got 94.4% accuracy rate on CK+ Dataset. Ruiz-Garcia et al. [27] used for recognition a hybrid model combined CNN for feature extraction and SVM for classification. They tested it model on the CK+ dataset and achieved a classification performance of 95.87%. In recap, we can conclude that the combination of two (or more) types of systems or methods (either prior- or post-combination) could significantly enhance the overall result of facial expression detection. Consequently, the combination system can be considered as an indicator for promising results.

5. Conclusions

This paper is prepared to review the latest facial expression recognition system studies which worked on using a combination of either features extraction or the recognition of expressions. FERS is used in many applications such as entertainment, medical, security, mental disorders and many other fields. Therefore, it is important to always have a high accuracy in recognizing facial expression. Multiple studies have been briefly reviewed in the previous sections to build a comprehensive view of recent theories proposed to enhance system performance. Prior- and Post-combination techniques were mentioned, either by using more than one method to combine results or features, or by enhancing the classifiers effectiveness and performance or the features extractor by the assistance of another one. In recap, different algorithm produce different results due to the robustness of the algorithms that are vary from one algorithm to another. Hence, using combination system gives promising results comparing with the individual system due to the combination system takes the advantages of different approaches to produce reliable result.

References

[1] F. Roli, G. Giacinto, and G. Vernazza, (2001, July). Methods for designing multiple classifier systems. In International Workshop on Multiple Classifier Systems (pp. 78–87). Springer, Berlin, Heidelberg.

[2] M. Mohandes, M. Deriche and S. O. Aliyu, "Classifiers Combination Techniques: A Comprehensive Review," in IEEE Access, vol. 6, pp. 19626-19639, 2018.

[3] A. M. Chacko, & Kumar, K. A. (2016). Offline Malayalam Character Recognition: A Comparative Study Using Multiple Classifier Combination Techniques. In Information Systems Design and Intelligent Applications (pp. 69-77). Springer, New Delhi.
[4] Z. A. Khalaf, (2015). Broadcast News Segmentation Using Automatic Speech Recognition System Combination with Rescoring and Noun Unification, PHD thesis, Universiti Sains Malaysia, Malaysia.

[5] W. M. Sanjaya, , D. Anggraeni, A. Juwardi, & M. Munawaroh, (2018, September). Design of Real Time Facial Tracking and Expression Recognition for Human-Robot Interaction. In Journal of Physics: Conference Series (Vol. 1090, No. 1, pp. 012044). IOP Publishing.

[6] M. Azriansyah, N. Hartuti, M. Fachrurrozi, & B. A. Tama, (2019, March). A Study about Principle Component Analysis and Eigenface for Facial Extraction. In Journal of Physics: Conference Series (Vol. 1196, No. 1, pp. 012010). IOP Publishing.

[7] L. I. Kuncheva, (2002). A theoretical study on six classifier fusion strategies. IEEE Transactions on pattern analysis and machine intelligence, 24(2), pp. 281–286.

[8] X. Sun, P. Xia, L. Zhang, and L. Shao, (2020). A ROI-guided deep architecture for robust facial expressions recognition. Information Sciences. v 522, pp (35-48). Elsevier

[9] T. Kalsum, S. M. Anwar, M. Majid, B. Khan, and S. M. Ali, (2018). Emotion recognition from facial expressions using hybrid feature descriptors. IET Image Processing, 12(6), 1004–1012.

[10] X. Sun, and M. Lv, (2019). Facial expression recognition based on a hybrid model combining deep and shallow features. Cognitive Computation, 11(4), 587–597.

[11] M. Mahmood, A. Jalal and H. A. Evans, "Facial Expression Recognition in Image Sequences Using 1D Transform and Gabor Wavelet Transform," 2018 International Conference on Applied and Engineering Mathematics (ICAEM), Taxila, 2018, pp. 1-6.

[12] A. Nandi, P. Dutta, and M. Nasir, (2020). Recognizing human emotions from facial images by Landmark Triangulation: A combined circumcenter-incenter-centroid trio feature-based method. In Algorithms in Machine Learning Paradigms (pp. 147–164). Springer, Singapore.

[13] N. Sun, Q. Li, R. Huan, J. Liu, and G. Han, (2019). Deep spatial-temporal feature fusion for facial expression recognition in static images. Pattern Recognition Letters, 119, pp. 49–61.

[14] N. Jain, S. Kumar, A. Kumar, P. Shamsolmoali, and M. Zareapoor, (2018). Hybrid deep neural networks for face emotion recognition. Pattern Recognition Letters, 115, pp. 101–106.

[15] D. M. Vo, and S. W. Lee, (2018). Robust face recognition via hierarchical collaborative representation. Information Sciences, 432, pp. 332–346.

[16] N. B. Kar, K. S. Babu, A. K. Sangaiah, and S. Bakshi, (2019). Face expression recognition system based on ripplet transform type II and least square SVM. Multimedia Tools and Applications, 78(4), pp. 4789–4812.

[17] J. Li, H. Zhang, W. Wan, and J. Sun, (2018). Two-class 3D-CNN classifiers combination for video copy detection. Multimedia Tools and Applications, pp. 1–13.

[18] M. J. Lyons, S. Akamatsu, M. Kamachi, J. Gyoba, and J. Budynek, “The japanese female facial expression (JAFFE) database,” in Proceedings of third international conference on automatic face and gesture recognition, 1998, pp. 14–16.

[19] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. 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. IEEE, 2010, pp. 94–101.

[20] P. L. Carrier, A. Courville, I. J. Goodfellow, M. Mirza, & Y. Bengio (2013). FER-2013 face database. Universit de Montral.

[21] N. Alfanti, C. Papachristou, & A. Delopoulos (2010, April). The MUG facial expression database. In 11th International Workshop on Image Analysis for Multimedia Interactive Services WIAMIS 10 (pp. 1-4). IEEE.

[22] R. Azhar, D. Tuwohingide, D. Kamudi & N. Suciati (2015). Batik image classification using sift feature extraction, bag of features and support vector machine. Procedia Computer Science, 72, 24-30.

[23] T. Randhavane, U. Bhattacharya, K. Kapsaskis, K. Gray, A. Bera and D. Manocha, "Learning Perceived Emotion Using Affective and Deep Features for Mental Health Applications," 2019 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct), Beijing, China, 2019, pp. 395-399.

[24] T. Zhang, W. Zheng, Z. Cui, Y. Zong and Y. Li, Spatial-Temporal Recurrent Neural
[25] A. S. M. Sohail. & P. Bhattacharya (2011). Classifying facial expressions using level set method based lip contour detection and multi-class support vector machines. International Journal of Pattern Recognition and Artificial Intelligence, 25(06), 835-862.

[26] S. Ouellet (2014). Real-time emotion recognition for gaming using deep convolutional network features. arXiv preprint arXiv:1408.3750.

[27] A. Ruiz-Garcia, Elshaw, M., Altahhan, A., & Palade, V. (2018). A hybrid deep learning neural approach for emotion recognition from facial expressions for socially assistive robots. Neural Computing and Applications, 29(7), 359-373.