ZERO – JURNAL SAINS MATEMATIKA DAN TERAPAN
Volume 3 No. 1 2019
Page : 1-11
P-ISSN: 2580-569X
E-ISSN: 2580-5754

IMPROVING CNN FEATURES FOR FACIAL EXPRESSION RECOGNITION

Ahmet Serdar Karadeniz[1], Mehmet Fatih Karadeniz[2], Gerhard-Wilhelm Weber[3], Ismail Husein[4]

Marketing and Economic Engineering, U1. Strzelecka 11 60-965 Poznan, Poland; and IAM, METU, 06800 Ankara, Turkey
[1]Hacettepe University Department of Computer Engineering, 06800 Beytepe Ankara, Turkey
[2]Cankaya University Department of Computer Engineering, Ankara, Turkey
[3]Poznan University of Technology Faculty of Engineering Management, Chair of
ahmet.karadeniz@hacettepe.edu.tr
[4]Program Studi Matematika, Universitas Islam Negeri Sumatera Utara Medan

Abstract Facial expression recognition is one of the challenging tasks in computer vision. In this paper, we analyzed and improved the performances both handcrafted features and deep features extracted by Convolutional Neural Network (CNN). Eigenfaces, HOG, Dense-SIFT were used as handcrafted features. Additionally, we developed features based on the distances between facial landmarks and SIFT descriptors around the centroids of the facial landmarks, leading to a better performance than Dense-SIFT. We achieved 68.34 % accuracy with a CNN model trained from scratch. By combining CNN features with handcrafted features, we achieved 69.54 % test accuracy.

Key Word*:* Neural network, facial expression recognition, handcrafted features

1. INTRODUCTION

Facial expression recognition (FER) is a system for inferring the emotions of people from images. Given an image of the face of a person, expression of the person needs to be determined automatically. FER is challenging as there are large variations in face images. These variations include personal attributes such as age, gender and ethnicity. Moreover, various poses, illumination and occlusions also make the problem harder.

FER systems are divided into two parts as static and dynamic facial expression recognition [1]. In static FER, expressions are determined from still images. For the dynamic case, sequences of frames are used to determine facial expressions. Furthermore, datasets are also divided into two types according to how they are collected. Some of the datasets are collected in controlled lab environments and some are collected from web or movies, which are also called the datasets for facial expression recognition in the wild [2, 3, 4]. In this paper, we focus on the static FER problem with a dataset collected from web. In Section 2, we

describe the details of the dataset we used for our experiments.

In the last decade, Convolutional Neural Networks have been used in solving many tasks in computer vision. In this work, we trained a CNN model and improved its accuracy with handcrafted features. Starting from simple features, we moved towards more advanced features to train a model and evaluated their performances both separately and with each other. Finally, we combined them with the features extracted by CNN and trained an SVM which resulted in an improvement of %1.20 test accuracy. In addition to the existing features, we used SIFT descriptors around the centroids of facial landmarks for each separate landmark region of the face. Then we combined these descriptors with pairwise distances between facial landmarks, achieving a higher accuracy than Dense-SIFT which was used in the previous state of the art. Final test accuracy we achieved is 69.54%.

DATA SET

Deep Neural Networks are used to be known as to require large amount of data. However, the amount of data in the facial expression recognition datasets are not large when compared to the other datasets such as ImageNet or COCO. A comprehensive list of datasets for FER is provided by Li et al. in their survey [1]. One of the largest and also challenging datasets in this list is FER- 2013 which is the dataset used in the experiments discussed in this work.

FER-2013 is one of the subchallenges of *ChallengesinRepresentation Learning* which was hosted by Kaggle under ICML Workshop [5]. The dataset was created from web using Google image search with different emotional keyword queries. Emotional keywords were also combined with gender, age and ethnicity. They used OpenCV face detection for automatically cropping faces and human labelers corrected failures in this operation. According to Goodfellow et al., human accuracy on this dataset was $65 \pm 5\%$ [5].

Figure 1. Sample images from FER-2013 [1].

There are 28709 training images, 3589 validation and 3589 test images in this dataset. All of the images are 48 × 48 grayscale images. In training and validation sets, there are 4462 anger, 492 disgust, 4593 fear, 8110 happy, 5483 sad, 3586 surprise, 5572 neutral face images. There are 7 emotion classes which are anger, disgust, fear, sad, happy, surprise and neutral. Example images can be seen in Figure 1. The distribution of these classes in FER-2013 [1] are shown in Table 1.

Table 1. Class distribution of FER-2013 [1].

Expression	Train	Validation	Frequency
Anger	3995	467	14.08%
Disgust	436	56	1.68%
Fear	4097	496	14.43%
Happy	7215	895	25.29%
Sad	4830	653	16.99%
Surprise	3171	415	11.21%
Neutral	4965	607	17.46%

RELATED WORK

All of the top three teams inKagglechallange used convolutional neural networks [6]. Tang et al. have used CNN with L2-SVM loss function which gave them 2% higher accuracy when compared to the second-best team. Their winning solution obtained 71.2% test accuracy.

In 2016, Connie et al. have trained SIFT features together with CNN [7]. With CNN and Dense-SIFT, they have achieved 72.1% test accuracy. Moreover, they used

an aggregator of CNN only, CNN with SIFT, CNN with Dense-SIFT and achieved 73.4% test accuracy which is 2.2% higher than the winning solution.

In 2017, Kacem et al. have used temporal trajectories of facial landmarks to classify expressions in videos [8]. They mapped facial landmarks to Riemann manifold of positive semi-definite matrices of rank 2. Then, they did a temporal alignment to measure dissimilarities between them and finally trained pairwise proximity function SVM. Their evaluation was provided on video FER datasets and the reported accuracies were 83.13% and 13.94% on the Oulu-CASIA and AFEW datasets respectively.

Acharya et al. have used manifold networks in conjunction with convolutional neural networks [9]. They obtained 58.14% and 87.0% on Static Facial Expressions in the Wild (SFEW) and Real-World Affective Faces (RAF) datasets respectively.

In 2018, Georgescu et al. have combined the features of CNN and Bag of Visual Words (BOVW) [10]. They have fine-tuned VGG-face model to obtain automatic features. Furthermore, they have used a local learning framework for prediction which is based on K-Nearest Neighbors and one-versus-all Support Vector Machines. They achieved the current best accuracy which is 75.42% on the test set of FER-2013 dataset.

TRAINING

In this section, the details of our training algorithms with deep and handcrafted features for facial expression recognition are provided. Throughout the experiments SVM with parameter $C = 1$ and radial basis function kernel is used for handcrafted features. All of the experiments are run on Google Colab which has Intel Xeon 2.30GHz CPU and Tesla K80 GPU.

BASELINE

We started with the one of the simplest features to train a facial expression recognition model. Using eigenfaces is one of the oldest techniques for face recognition. Basically, Principal Component Analysis (PCA) is applied to the grayscale images to obtain most characteristic faces in the dataset. With 35 principal components, a validation accuracy of 44.07% is obtained. Total training time was 120.17 seconds.

Then, we used Histogram of Oriented Gradients with $(12, 12)$ pixels per cell and $(2, 2)$ cells per block. With HOG features, we obtained 48.9% validation accuracy is obtained.

For each image in the dataset, keypoints are created densely with $(12, 12)$ blocks. Then SIFT descriptor is used to extract the descriptors around $(12, 12)$ neighborhoods of these keypoints. By using Dense-SIFT features with SVM, 50.15% validation accuracy is obtained which was similar to HOG as expected. A visualization example of Dense-SIFT keypoints can be seen in Figure 2.

Figure 2. Visualization of Dense-SIFT keypoints.

FACIAL KEYPOINTS AND DESCRIPTORS

Our approach for handcrafted features is based on facial landmarks since they are among the powerful features for characterizations of faces. Emotions can also be inferred from the shapes of eyes, eyebrows and mouth. We used Dlib to extract facial landmarks for each image in the dataset. These landmarks include the points for eyes, eyebrows, nose, mouth and jawline. Eliminating the landmarks for nose and jawline, a total of 42 keypoints are obtained from an image.

First, we take the center of the keypoints for each part in the face. In particular, center of the keypoints for left eyes, right eyes, left eyebrows, right eyebrows and mouths are computed separately. An example for center points for facial landmarks is represented in Figure 3.

Figure 3. Example for center points for facial landmarks. In this figure, yellow and green colors represent eyebrows. Purple and yellow represent eyes. Blue represents the mouth keypoints. Centers are represented as light blue points.

Then SIFT keypoints are created from these facial part centers with size 12. Then SIFT descriptors are computed for each facial part. Training SVM on these descriptors gave 51.37% validation accuracy which is around 1% higher than Dense-SIFT.

An observation we made is that expressions can be recognized directly from the relative positions of facial landmarks. For instance, consider the mouth of a surprised face image. Distances between mouth landmarks of surprised faces are usually larger than the mouth landmark distances of other facial expressions (see Figure 4 which represents an example for pairwise distances between mouth landmarks for a neutral and a surprised face). Furthermore, landmarks between different facial parts can also give useful information. For example, eyebrows are closer to the eyes in sad expressions and eyebrows are usually closer to each other in anger expressions.

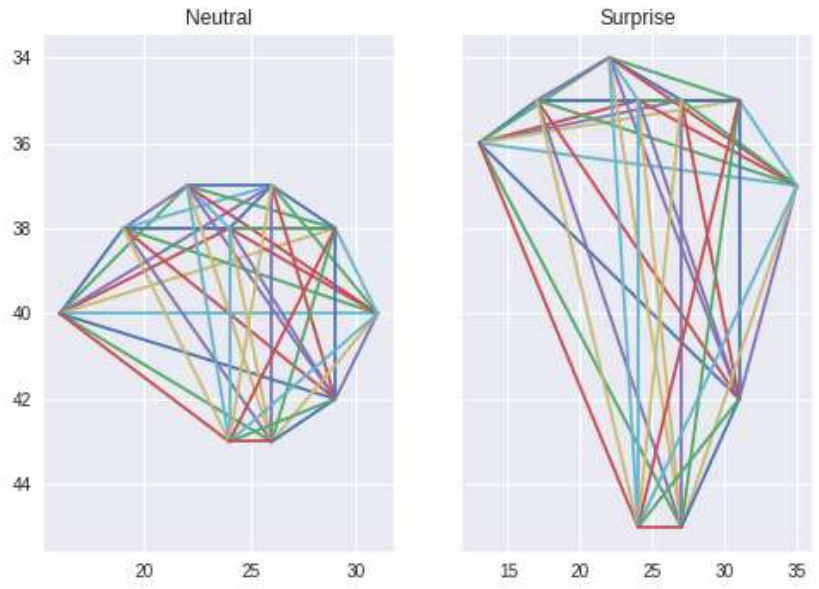

Figure 4. Example for pairwise distances between mouth landmarks for different expressions. Figure on the lefts represents distances for a neutral face. Figure on the right represents distances for a surprised face.

With this intuition, we computed the pairwise distances between each keypoints. Manhattan distance metric was used when calculating the distances. A total of $\frac{42 \cdot 41}{2} = 861$ distances are obtained for each image. Then, we applied PCA to these features and obtained 28 features for distances. By using only the distance features, we obtained 50.54% validation accuracy.

Furthermore, combination of pairwise distance features with the descriptors around facial landmarks gave the largest accuracy obtained from handcrafted features, which was 57.03%. Although this was not a high accuracy when compared to the other works, obtaining a 6.88% higher accuracy than Dense-SIFT may be a significant improvement in the future. This combination will be called

Face-SIFT in the later parts of this paper.

CNN

A CNN model was trained to learn features automatically from the data. Data augmentation is applied to the data which randomly flips the images horizontally, rotates in a range of 30 degrees and 0.2 zoom, shift and shear range. First, a similar architecture to the winner of FER-2013 was used to train a CNN. This architecture starts with 2 convolutional layers with 32 filters, followed by 2 convolutional layers with 64 filters, and then 2 more convolutional layers with 128 filters. (2, 2) max pooling layers were used after each two convolutional layers. Then, after flattening the last convolutional layer, 2 fully connected layers with 2048 neurons were used. Dropout layers with 0.5 probability were used between fully connected layers. ReLU activation function was used for all the convolutional layers. As loss and final activation function, softmax with cross entropy loss were used. Adam optimizer and 128 batch size were used to train the model. With this architecture, 65% validation accuracy is obtained after 50 epochs.

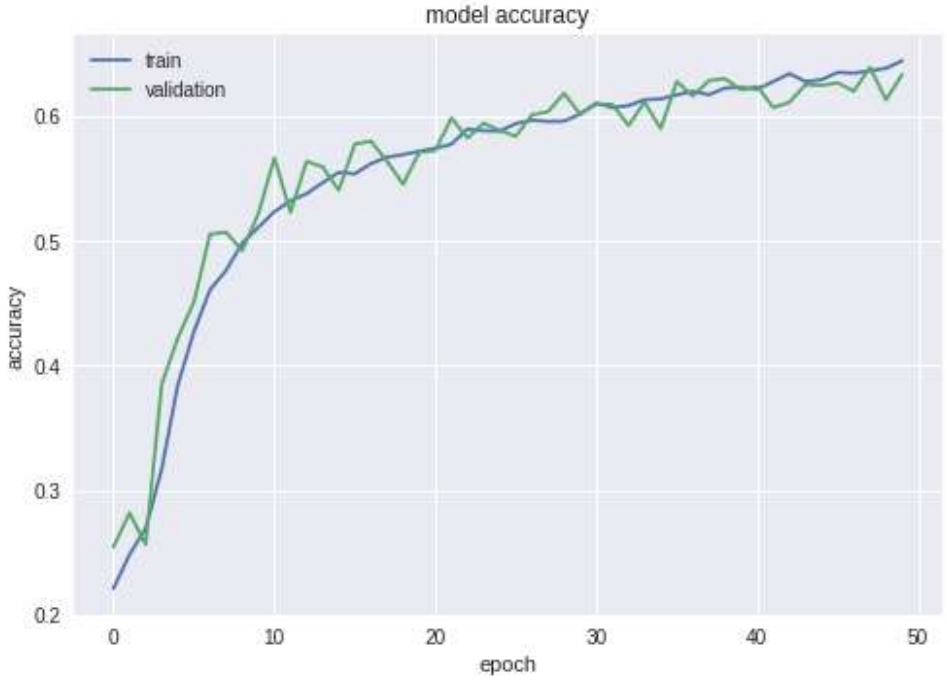

Figure 5. CNN validation accuracy

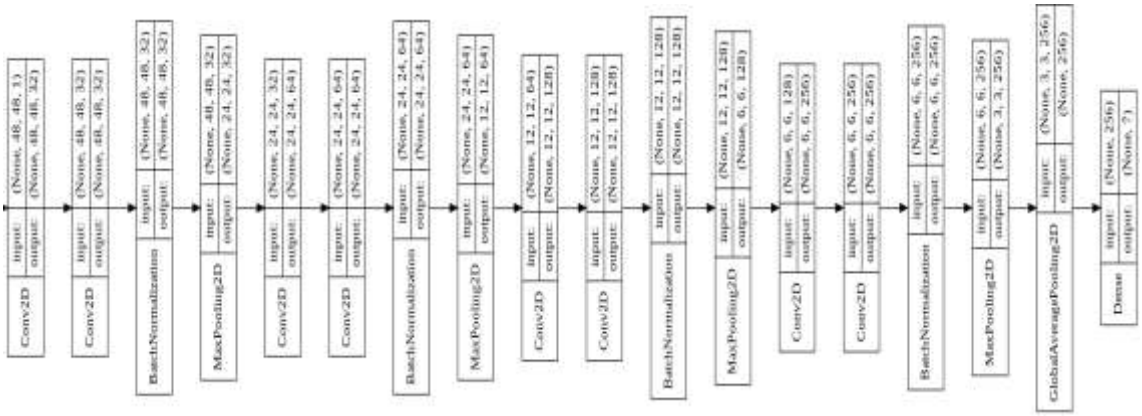

Figure 6. CNN Architecture.

Then, we removed all the fully connected layers except the output layer. We also removed the dropout layers. Batch normalization layers were added before the max pooling layers and two more convolutional layers were added with 256 filters. We also replaced the flattening layer with global average pooling layer which results in less number of features. Complete architecture is shown in Figure 6. With this architecture, we achieved a 67% validation accuracy which can be seen in Figure 5.

Furthermore, a total of 256 features are extracted from the global average pooling layer. We trained an SVM on these features and obtained 68.12% validation accuracy. Finally, combining all the handcrafted features with CNN features resulted in 69.82% validation accuracy. This accuracy is very slightly higher than the validation accuracy of the winner of FER-2013 according to the public leaderboard of the Kaggle competition.

In the following section, evaluation of the model is provided.

EVALUATION

All the development and optimization process made on the validation set. We tested the model once for each feature set.

By using SVM on the extracted CNN features, we obtained 68.34% test accuracy. In Figure 7, the confusion matrix of this model is shown. Anger, fear and sad expressions had relatively low accuracies when compared to the other expressions in the confusion matrix.

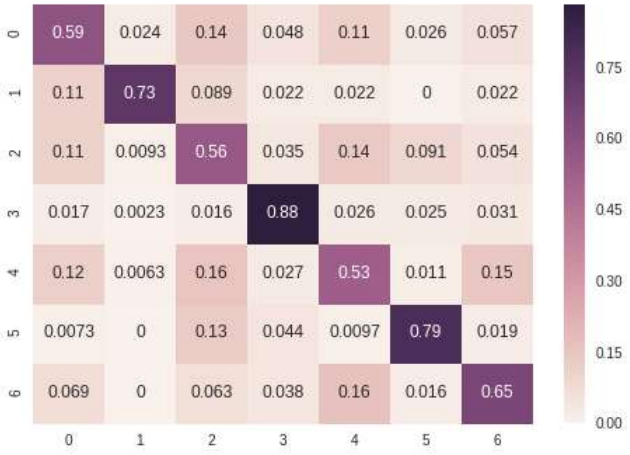

Figure 7. CNN confusion matrix.

In Table 2, test accuracies for each separate features are provided. Test accuracies were similar to the validation accuracies. By combining all handcrafted features with CNN features, 69.54% validation accuracy was obtained.

Table 2. Test accuracies.

Features	Accuracy
Eigenfaces	44.74%
HOG	49.23%
Dense-SIFT	49.65%
Face-SIFT	56.14%
CNN	68.34%
All features	69.54%

In Figure 8, the confusion matrix for combination of all features is plotted. Scores for all classes were improved slightly.

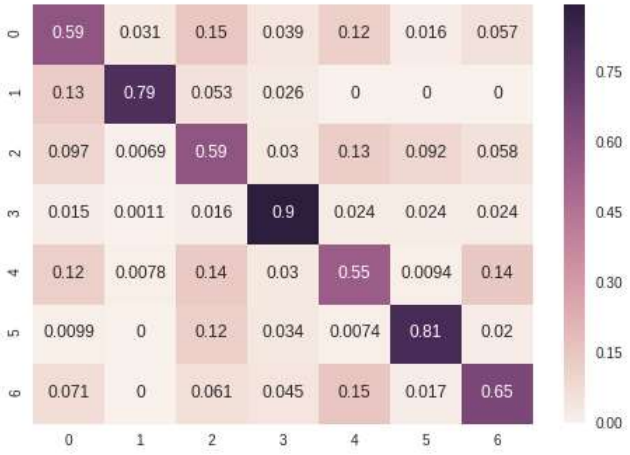

Figure 8. Confusion matrix for combination of all features.

2. CONCLUSION AND OUTLOOK

Main concentration in this work was to improve the performance of CNN with handcrafted features. This required us to develop additional features in addition to the existing feature descriptors. We achieved 1.20% percent higher accuracy than CNN with the help of handcrafted features. Almost no parameter tuning was made on both CNN and SVM. In the future, more advanced CNN architectures with proper tuning and more careful design of handcrafted features can be done to improve the overall accuracy.

3. REFERENCES

[1] S.Liand W.Deng."DeepFacial Expression Recognition: A Survey,"*CoRR*, 2018.

[2] 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,"*2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, 2010.

[3] M.Lyons,S.Akamatsu,M.Kamachi, and J. Gyoba "Coding facial expressions with Gabor wavelets,"*Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*, 1998.

[4] A. Dhall, O. R. Murthy, R. Goecke, J. Joshi, and T. Gedeon, "Video and Image based Emotion Recognition Challenges in the Wild,"*Proceedings of the 2015 ACM on International Conference on Multimodal Interaction - ICMI 15,* 2015.

[5] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D. H. Lee, Y. Zhou, C. Ramaiah, F. Feng, R. Li, X. Wang, D. Athanasakis, J. Shawe-Taylor, M. Milakov, J. Park, R. Ionescu, M.

Popescu, C. Grozea, J. Bergstra, J. Xie, L. Romaszko, B. Xu, Z. Chuang, and Y. Bengio, “Challenges in representation learning: A report on three machine learning contests,” *Neural Networks*, vol. 64, pp. 59-63, 2015.

[6] Y. Tang. “Deep Learning using Linear Support Vector Machines,” arXiv preprint arXiv:1306.0239, 2013.

[7] T. Connie, M. Al-Shabi, W. P. Cheah, and M. Goh, “Facial Expression Recognition Using a Hybrid CNN-SIFT Aggregator,” *Lecture Notes in Computer Science Multi-disciplinary Trends in Artificial Intelligence*, pp. 139-149, 2017.

[8] A. Kacem, M. Daoudi, B. B. Amor, and J. C. Alvarez Paiva, "A Novel Space-Time Representation on the Positive Semidefinite Cone for Facial Expression Recognition,"*2017 IEEE International Conference on Computer Vision (ICCV)*, 2017.

[9] D. Acharya, Z. Huang, D. P. Paudel, and L. V. Gool, "Covariance Pooling for Facial Expression Recognition,"*2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2018.

[10] M. Georgescu, I. Georgescu, R. T. Ionescu and M. Popescu. "Local Learning with Deep and Handcrafted Features for Facial Expression Recognition,"*CoRR*, 2018.