A Deep-Learning Approach to Facial Expression Recognition with Candid Images

Wei Li
CUNY City College
lwei000@citymail.cuny.edu

Min Li
Alibaba. Inc
mushi.lm@alibaba.inc

Zhong Su
IBM China Research Lab
suzhong@cn.ibm.com

Zhigang Zhu
CUNY Graduate Center and City College
zhu@cs.ccny.cuny.edu

Abstract

To recognize facial expression from candid, non-posed images, we propose a deep-learning based approach using convolutional neural networks (CNNs). In order to evaluate the performance in real-time candid facial expression recognition, we have created a candid image facial expression (CIFE) dataset, with seven types of expression in more than 10,000 images gathered from the Web. As baselines, two feature-based approaches (LBP+SVM, SIFT+SVM) are tested on the dataset. The structure of our proposed CNN-based approach is described, and a data augmentation technique is provided in order to generate sufficient number of training samples. The performance using the feature-based approaches is close to the state of the art when tested with standard datasets, but fails to function well when dealing with candid images. Our experiments show that the CNN-based approach is very effective in candid image expression recognition, significantly outperforming the baseline approaches, by a 20% margin.

1 Introductions

Facial expression is important in social interaction. For example, for visually impaired people, without being able to see the expression of other people, the social interaction would be significantly limited. In some other applications such as in TV or show businesses, knowing their audience's real-time expressions when watching the TV programs or shows may obtain more meaningful feedback for their producers. All these require the recognition of facial expressions in real-time using candid (non-posed) images.

To achieve the goal of recognizing facial expression from candid images, we propose to use deep learning methods. For comparison, two baseline approaches are designed, using two popular feature extraction methods, LBP and SIFT, respectively. In our learning-based approach, we propose to use convolutional neural networks (CNNs), which in the past have been proved to be effective in image classification. Compared with the feature based approach, since the features are automatically learned from images, we expect that the CNN-based approach would be much more effective in recognition. As a summary, this paper has the following contributions: 1) A Candid Images Facial Expression dataset, CIFE, is created from online images. 2) A convolutional neural network based approach was proposed for facial expression recognition, which has been shown to be effective in recognizing facial expression of candid images.

Table 1. State-of-the-art approaches on some common used datasets. Please see text for test "Setting"

Approaches	CK	CK+	MMI	Setting
Manifold LBP[13]	96.75	–	–	(1)
Manifold Gabor[13]	95.38	–	–	(1)
ITBN [12]	–	86.3	59.7	(2)
HMM [12]	–	83.5	51.5	(2)
Lucey [12]	–	83.3	–	–
CPL [14]	88.42	–	49.4	(3)
CSPL [14]	89.89	–	73.5	(3)
ADL [6]	82.26	–	47.8	(3)
AFL [6]	86.94	–	–	(3)
MCF [1]	–	70.1	–	(4)
SVM+LBP [6]	–	89.4	–	(4)

The rest of the paper is organized as the following. Section 2 discusses related work. Section 3 introduces our candid facial expression dataset, and test the two feature-based approaches with both standard facial expression datasets and our candid dataset. In Section 4, our proposed CNN-based approach is described, and important issues such as the structure of the model, and the augmentation of the training samples are discussed. Section 5 provides experimental results. Conclusion and future work are presented in Section 6.

2 Related Work

Various approaches have been proposed to recognize facial expression, and most of them use standard datasets. A popular framework was proposed in [9], with three steps: 1) face acquisition, 2) face feature extraction and representation, and 3) face expression recognition. A widely used expression dataset CK+ was proposed in [5]. In this paper we also test our feature-based approaches on this dataset.

Most of the expression approaches focused on recognizing expression of frontal faces, such as the images in CK+. A very straightforward approach was proposed by Shan, et al [6], in which a LBP-based feature extractor was combined with an SVM for classification. Zhong, et al [14] proposed an approach to analyze both the so-called common and specific facial expression patches when dealing with the facial recognition problem. Xiao, et al [13] proposed an approach by training separate model for each expression instead

of training one model for all expressions in learning, and the performance of the recognition becomes better. Chew et al [1] proposed a modified correlation filter (MCF) based expression recognition approach, which is similar to traditional SVM plus correlation filter approaches, but with the difference of optimizing only one hyper-plane so the large margin losses are decreased. Wang, et al [12] modeled the facial expression problem as a complex activity that consisted of temporally overlapping or sequence of face events, then an interval Temporal Bayesian Network (ITBN) was used to capture the complex temporal information

The above approaches used different features and frameworks to accomplish facial expression recognition, and most of them are tested on the datasets with frontal and well-posed images: CK [3], CK+ [5], and MMI [10]. We have assembled the performance results of some of the approaches on the three datasets and list their recognition rates in Table 1. Note that the approaches listed in Table 1 were not tested on all the the three datasets so some of the data items are missing. The recognition results were also affected by testing settings: (1) 50% train, 25%validation and test; (2) cross validation (C-V) - CK+ 15 folds C-V MMI 20 folds, (3)10 folds C-V, and (4) leave-1-out testing, which means training a model with one sample out, then testing it, which may leads to high recognition rates e.g. [6].

The expression recognition approaches we mentioned above are mostly based on the extracted features. We have noted that a convolutional neural network (CNN) based approach was proposed in [4] to deal with the image classification, which gained a huge improvement. In [8] and [7], face recognition were accomplished by using convolutional neural networks, and the recognition rates on the Labeled Faces in the Wild (LFW) dataset were 97% and 99%, respectively.

There are some efforts made to accomplish the facial recognition task by using neural network based methods. Fasel [15] used a simple CNN structure to deal with the JAFFE database in 2002. Due to hardware limit, only a small dataset is processed. Liu, et al [16] used combined deep learning models (CNN and Restricted Boltzmann Machines(RBM)) and FACS (Facial Coding System) to recognize basic expressions in different datasets. Besides CNN models, Deep Belief Network (DBN) is also used by researchers to accomplish expression recognition. Liu, et al [17] proposed a Boosted Deep Belief Network to perform feature learning, feature selection and classifier construction for expression recognition. Kim, et al [18] compared different DBN models for unsupervised feature learning in audio-visual emotion recognition. But these deep learning based approaches are mainly based on standard database like CK+ and JAFFE.

If we want to recognize expression in real scenes, it is not enough just relying on the widely used datasets. We need get more candid expression data. One way to obtain large number of data may be grabbing images from the web. Richter, et al [19] obtained 4761 images from web based on key words searching, then represented the images with DCT, LBP and Gabor filters. Instead of extracting those features after gathering candid images, we will use deep learning based method to accomplish both feature extraction and expression recognition.

Table 2. Evaluating feature based approaches

SVM (Soft margin c)	LBP (%)	SIFT (%)
c=0.1	79.52	86.69
c=1	79.86	85.67
c=3	79.86	86.01
c=5	79.5	83.28

3 Our Candid Dataset and Baselines

In order to evaluate the performance of our proposed learning-based approach for candid expression recognition, we have done two preparations: developing baseline approaches and collecting a candid facial expression dataset. To make sure the performance of our baseline approaches are comparable to the state of the art, we first test them on the standard dataset CK+, before we test them on our candid dataset.we follow the framework proposed by Tian, et al [9], and employ two popular features to represent the face images: Local Binary Pattern (LBP) and Scale-Invariant Feature Transform (SIFT). Then SVMs are used to accomplish the recognition.

3.1 Baseline approaches

Images of CK+ are represented by the LBP and SIFT features, then SVMs are employed as the classifiers to obtain expression recognition results with various parameter settings for the SVM soft margin c (Table 2). We take 70% of images for each class of expression as the training data and the rest 30% for testing. The results are shown in Table 2. Comparing the recognition rates of our feature-based approaches (Table 2) with the results reported in literature (Table 1) on the dataset CK+, we can see that the performance of our feature-based approaches is very close to the state of the art, especially the SIFT-based approach. Therefore we use them as our baseline approaches to evaluate our proposed CNN-based approach.

3.2 CIFE dataset and test

We note that most of the expression images on the Web are randomly posed, and most of the expressions are naturally posed. Therefore we use web gathering techniques to acquire candid expression images from the Web, and create our candid image facial expression dataset CIFE (please contact the author to acquire the data). As we have mentioned, we define seven types of expressions: Happiness, Anger, Disgust, Sadness, Surprise, Fear and Neutral. Using related key words to the each of the 7 expressions in addition to the name of the expression (e.g., joy, cheer, smile for Happiness), we have collected a large number of images that belong to the same expression. Finally, we keep 10,595 images (after some post-filtering by humans) for the 7 classes. We used the Viola face detector [11] to detect faces from the candid datasets, and scaled them to 64x64. Figure 1 show a few typical examples of faces with various poses. In our experiments, we use 70% (7417) of the images for training and the rest for testing. The dataset will be made publicly available after

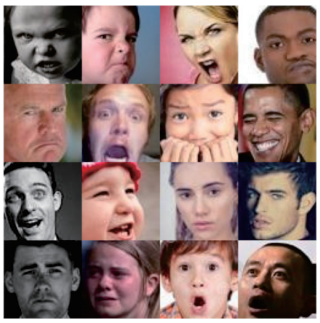

Figure 1. Candid face samples of various poses.

Table 3. Performance of the baseline approaches on our CIFE dataset vs on CK+

Approaches	CK+ (%)	CIFE (%)
LBP+SVM(c=1)	79.86	62.3
SIFT+SVM(c=3)	86.69	59.7

the publication of the work. We would like to note that images of various expressions are not distributed evenly. The numbers of samples for the seven types of expression are: Anger (1785), Disgust (266), Fear (781), Happiness (3636), Neutral (644), Sadness(2485) and Surprise(997).

We use the parameter settings with the best performance for LBP-based and SIFT-based approaches (Table 1), respectively, to train and test on our candid expression dataset CIFE, and the recognition results are shown in Table 3. Unfortunately but not surprisingly, the recognition rates of the two baseline approaches on the candid images dataset are just around 60%. In this case, LBP-based is slightly better than the SIFT-based. The reason for low performance may be due to the varieties in the face poses, which is a major difference from the well-posed images in CK+.

4 The proposed CNN-based Approach

Convolutional neural networks (CNNs) are one of the most popular deep learning structures. CNNs have been shown to be highly effective in image based classification and regression. Many details about the components and configurations of the convolutional neural networks have been provided in literature, e.g., in [4]. In this paper, we focus on the design of the network structure for our goal of candid image expression recognition.

4.1 Data augmentation for CNN training

Deep learning with CNNs always requires very large number of training images in order to obtain good classification results. Even though our CIFE dataset has 10,595 images for 7 classes, the largest we know of, it is still insufficient for training a deep CNN model. So before training the CNN model, we need to augment the dataset with various transformations for generate various small changes in appearances and poses. We applied five image appearance filters and six affine transform matrices. The five filters are disk, average, Gaussian, unsharp and motion filters, and the six affine transforms are formalized by adding slight geometric transformations to identity matrix, including a horizontally mirror transformation. By doing this augmentation, for each original image in the dataset, we can generate 30 (=5x6) samples, therefore the number of possible training samples would be 222,510 (=7417x30). The competition time will be a big issue if we also use the augmented dataset for the training of our baseline approaches. Therefore only the original dataset is used. One of the example is shown in Figure 2.

Figure 2. Image augmentation: from 1 to 30.

4.2 The CNN model structure

After data augmentation, we now have 222,510 training images, and the model will be tested on 3,178 original testing images (30% of 10,595). The goal is to classify all the images into 7 groups. In face recognition, face alignment and rectification are usually performed to ensure the same features on face like nose or eyes aligned to improve the recognition accuracy. However, in facial expression recognition, an alignment may distort or reduce the expression feature. Therefore we simply use the original images that are cropped with the Violo face detector [11] . We hope the CNN structure will deal with the viewing changes. Figure 3 shows the CNN model structure, with one input layer (the original image), three convolutional layers, and an output layer. This structure is arrived with many experimental tests. We set the initial convolutional filters size to be 7x7. Then we vary the numbers of layers and the number of filter for each layer. After many rounds of test, we finally arrived the 'best' structure with 3 convolutional layers, and the filter numbers for each layer to be 32, 32, 64, respectively. The convolutional filters in the 1st layer are applied to color images, and then the filtered images are combined into intensity images. Filters in the 2nd and 3rd layers are applied to intensity images. For each of the three convolutional layers, we add a 2:1 pooling layer to make the output data less redundant. With this structure, we can easily know the numbers of the parameters to be around 184,000. Comparing to the number of training images (222,510), the structure setting is also appropriate.

5 Experiments

We use the CAFFE lib [2] to implement our CNN model. A Nvidia k20 graphic card is used to train the model for more time-efficient training. We can train the model of about 184, 000 parameters for 40,000 iterations in about 2 hours. The learning rate is changed

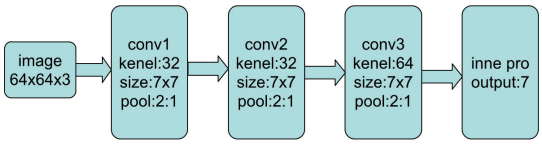

Figure 3. CNN structure design.

Table 4. Comparison: CNN and baselines.

Approaches	CK+ (%)	CIFE (%)
LBP+SVM(c=1)	79.86	62.3
SIFT+SVM(c=3)	86.69	59.7
CNN	83	81.5

from 0.001 to 0.0001 when the number of training iterations reaches 20, 000 and to 0.00001 when the number of iterations reaches 30, 000.

At each round of the iterations in model training, the layer parameters of the network are updated based on the loss. We set a maximum number of iterations and when the training times reaches the number, we obtain a trained model, which is essentially the parameters of all the filters. We then save the model so we can use the model to predict an expression of a candid image.

We have tested the trained CNN model on the 30% test images from our CIFE dataset. By comparing the recognizing results on testing images with saved labels, an average recognition rate is computed. We also compare this result with that of our baselines - the two feature based approaches. The results are summarized in Table 4. From the table, we can draw two conclusions. First, the CNN-based model's performance is comparable with feature-based approaches on well-posed data, even though with the CK+ dataset the number of training samples is apparently not sufficient for the deep learning approach. Second, and more importantly, the CNN-based model significantly outperforms the feature-based approaches, by 20%.

After the training, the recognition procedure is near real-time: for the model with seven types of expression, the computing time varies from 200ms to 300ms in recognizing facial experiment of an input image. A demo system has been built for demonstration of images captured by a built-in camera in a laptop or desktop computer.

6 Conclusion and Future Work

In this paper, a deep learning based approach has been proposed to recognize facial expression from non-posed facial images. The proposed CNN-based approach has been compared with two feature-based approaches (basleines).The feature based approaches are shown to be able to obtain comparable performance as the state of the art approaches, on well-posed face datasets. For candid images, our experiments show that our CNN-based approach significantly outperforms the baselines We have also built a real-time candid image expression recognition system based on the trained model.

There are a number of future research directions. (1) We would like to speed up the online recognition so a dynamic facial expression analysis of live video will be possible. (2) The integration of the engineered features and the learned features would be one of the possibilities for further improving the recognition performance of the integrated approach.

Acknowledgements. This work has been supported by the IBM China Research Lab, and the National Science Foundation (Award EFRI-1137172). Part of the work was performed when the first author was visiting IBM China Research Lab.

References

[1] S. W. Chew, et al. Improved facial expression recognition via uni-hyperplane classification. CVPR 2012.

[2] Y. Jia. Caffe: An open source convolutional architecture for fast feature embedding. http://caffe. berkeleyvision. org, 2013.

[3] T. Kanade, J. F. Cohn, and Y. Tian. Comprehensive database for facial expression analysis. FG 2000.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In Advances in NIPS, 2012.

[5] P. Lucey, et al. The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression. CVPRW, 2010.

[6] C. Shan, S. Gong, and P. W. McOwan. Facial expression recognition based on local binary patterns: A comprehensive study. Image Vision Comput. 27 (6), 2009: 803-816.

[7] Y. Sun, X. Wang, and X. Tang. Deep learning face representation from predicting 10,000 classes. CVPR 2014.

[8] Taigman, et al. Deepface: Closing the gap to human-level performance in face verification. CVPR2014.

[9] Y. Tian, T. Kanade, and J. F. Cohn. Facial expression analysis. In Handbook of Face Recognition, pp. 247-275. Springer, 2005.

[10] M. Valstar and M. Pantic. Induced disgust, happiness and surprise: an addition to the mmi facial expression database. In Proc. Intel Conf. Language Resources and Evaluation, Workshop on EMOTION, pp. 65-70. 2010.

[11] P. Viola, and M. Jones. Rapid object detection using a boosted cascade of simple features. CVPR 2001.

[12] Z. Wang, S. Wang and Q. Ji. Capturing complex spatio-temporal relations among facial muscles for facial expression recognition. CVPR 2013.

[13] R. Xiao, et al. Facial expression recognition on multiple manifolds. Pattern Recognition 44(1), 2011:107-116.

[14] L. Zhong, et al. Learning active facial patches for expression analysis.CVPR 2012.

[15] B. Fase. Head-pose invariant facial expression recognition using convolutional neural networks. In Proc. Multimodal Interfaces, 2002.

[16] M. Liu, et al. Au-aware deep networks for facial expression recognition. FG 2013.

[17] P. Liu, et al. Facial expression recognition via a boosted deep belief network. CVPR 2014.

[18] Y. Kim, H. Lee, and E.M. Provost. Deep learning for robust feature generation in audiovisual emotion recognition. ICASSP 2013.

[19] M. Richter, T. Gehrig and H. K. Ekenel. Facial expression classification on web images. ICPR 2012.