A Highly Versatile Facial Expression Recognition System

Wei Cui
University of Toronto
1004536479
w.cui@mail.utoronto.ca

Kewei Qiu
University of Toronto
1003798116
kewei.qiu@mail.utoronto.ca

Chenhao Gong
University of Toronto
1004144598
chenhao.gong@mail.utoronto.ca

Abstract

Facial expression recognition is the operation that analyzes facial expressions on human faces, and connect them to the corresponding emotional states. This report is regarding to a project, conducted by the authors, that suggests a new highly versatile facial expression recognition system, which is based on Fisherfaces algorithm and a Convolutional Neural Network learning model, while supported by the Viola-Jones Feature Detection framework to enable usage in more general scenarios. The report will also compare the system with a number of other popular algorithms and learning models to see how well the system will perform in compasion.

1. Introduction

Over the last couple of decades, facial expression analysis has become an overwhelmingly popular topic in computer vision related researches. Facial expressions are the facial changes in response to a person's internal emotions states, intentions, or social communications. Humans can easily tell a person's emotions by looking at their facial expression; whilst, computers need the assistance of a system to analyze the facial expression shown by a human, and connect it to an emotional state known to humans. Such a system is regarded as a facial expression recognition system.

Prior studies led to the development of a 3-step traditional facial expression analysis thesis, including face localization, feature extraction and expression classification. [3] Face localization refers to the step to identify the face in the image so as to enable further analysis on the expressions it is exposing. The Viola-Jones object detection framework is one of the most popular techniques applied in face detection and localization, which features a rapid detection algorithm after the model is trained. [8] Feature extration is the step that extracts the significant features that may indicate the facial expression. Either geometric or appearance features shall be extracted for classification. The final step, facial expression classification, is to match the facial expression, demonstrated by the extracted features in the previous step, to the emotions humans are familiar with. In practice, this step is highly related to machine learning, while multiple learning classifier models have been used as found in researches. These include Neural Networks (NN) [9], Support Vector Machines (SVM) [6], Adaboost etc.

Our highly versatile facial expression recognition system was inspired by and modified from a feature extraction algorithm called "Fisherfaces", developed by Peter N. Belhumeur, et al.[2] We integrate the algorithm with a convolutional neural network classifier, and implement the Viola-Jones framework [8] ahead of it to enable its versatility to operate on any image including a human face.

2. Methodology

2.1. Viola-Jones Object Detection Framework

Viola-Jones object detection framework, proposed by Paul Viola and Michael Jones, is a rapid object detection algorithm realized using boosted cascade classifiers. [8] This is what we used to detect and align faces from an arbitrary photo.

Viola and Jones' algorithm is a machine learning based algorithm. To train the classifier, a dataset with plenty of positive images (images of faces) and negative images (images without faces) is needed. The first step is to extract features from these images using rectangle-shaped Haar feature filters. Each feature is a single value obtained by subtracting sum of pixels under white rectangle from sum of pixels under black rectangle. Figure 1 provides some examples of Haar feature filters.

To calculate Haar features rapidly, the algorithm uses an intermediate representation for the image called integral image. The integral image at each location (x, y) contains the sum of the pixels above and to the left of x, y, inclusive is:

$$ii(x, y) = \sum_{x' \leq x, y' \leq y} i(x', y')$$

where $ii(x, y)$ is the integral image and $i(x, y)$ is the origi-

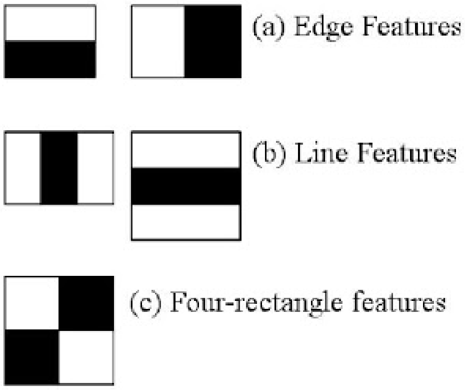

Figure 1. Example Haar feature filters

nal image. Using the following pair of recurrences:

$$s(x,y) = s(x,y-1) + i(x,y)$$
$$ii(x,y) = s(x-1,y) + s(x,y)$$

($s(x,y)$ is the cumulative row sum, $s(x,-1) = 0$, and $ii(-1,y) = 0$) the integral image can be computed within one iteration over pixels of the original image. [8]

After extract features from images, the alogrithm uses AdaBoost to select the best features out of the large number of extracted feathers. The final classifier is a weighted sum of weak classifiers. According to the paper, even 200 features provide detection with 95% accuracy. The final classifier has around 6000 features which gives us a precise object detection classifier.

2.2. Fisherfaces vs. Eigenfaces

Fisherfaces [2] and Eigenfaces [7] are two algorithms that were suggested and tested for face recognition, a slightly different problem than what we are interested in.

Over the development of face recognition techniques, Eigenfaces has become one of the popular algorithms with quality outcomes. The essence of the Eigenfaces approach is the use of principal components analysis (PCA) for feature extraction. The algorithm reconstructs a multitude of face images in from the training set by reconstructing them using weighted sums of a small collection of characteristic features (*eigenpictures*), namely *eigenfaces*. After the calculation, it keeps only the images that correspond to the highest eigenvalues in each epoch. [7] This is useful to lower the dimensions of the feature space, which is to be fed to the classifier model that follows. However, it yields projection directions that maximize the total scatter across all classes, which contains unwanted variations that are caused by exterior information, i.e. lighting. [2]

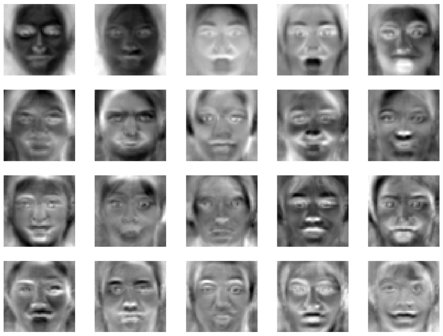

Figure 2. Eigenfaces for CK+48 Face Database.

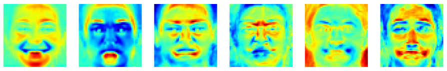

Figure 3. Fisherfaces (6 centroids) for 7 emotion labels of CK+48 Face Database.

Fisherfaces takes the advantage of the fact the classes are actually linearly separable. It employs Fisher's Linear Discriminant (FLD), an class specific method that selects feature spaces in such a way that the ratio of the between-class scatter and the within-class scatter is maximized. Figure 4 is a comparison of PCA and FLD for a two-class problem. One can see that while both PCA and FLD project the points from 2D down to 1D, PCA actually smears the classes together so that they are no longer linearly separable in the projected space, while FLD achieves greater between-class scatter, consequently simplifying the classification. [2]

In actual practice, however, using FLD alone in face detection may lead to problems. Formally, let us consider a set of N sample images $\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\}$ taking values in an n-dimension image space, and assume that each image belongs to one of c classes $\{X_1, X_2, \cdots, X_c\}$. Now let the between-class scatter matrix be defined as

$$S_B = \sum_{i=1}^{c} N_i(\mu_i - \mu)(\mu_i - \mu)^T$$

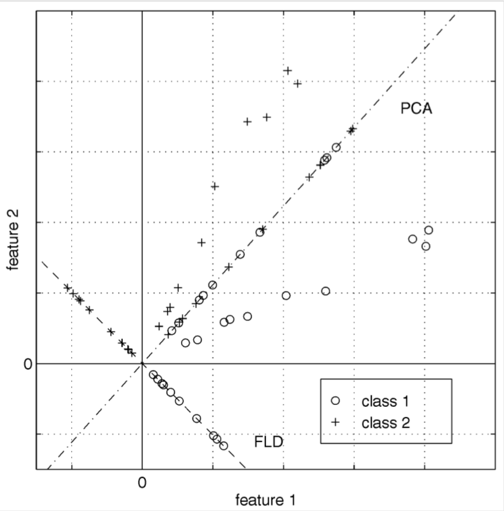

Figure 4. A comparison of principal component analysis (PCA) and Fisher's linear discriminant (FLD) for a two class problem where data for each class lies near a linear subspace. [2]

and the within-class scatter matrix be defined as

$$S_W = \sum_{i=1}^{c} \sum_{\mathbf{x}_k \in X_i} (\mathbf{x}_k - \mu_i)(\mathbf{x}_k - \mu_i)^T$$

where μ_i is the mean image of class X_i, and N_i is the number of samples in class X_i. In FLD, we choose an optimal projection W_{opt} such that it maximazes the ratio of the determinant of the between-class scatter matrix of the projected samples to the determinant of the within-class scatter matrix of the projected samples, i.e.

$$\begin{aligned} W_{opt} &= \arg\max_{W} \frac{|W^T S_B W|}{|W^T S_W W|} \\ &= \begin{bmatrix} \mathbf{w}_1 & \mathbf{w}_2 & \cdots & \mathbf{w}_m \end{bmatrix} \end{aligned}$$

where $\{\mathbf{w}_i | i = 1, 2, \cdots, m\}$ is the set of generalized eigenvectors of S_B and S_W corresponding to the m largest genrealized eigenvalues $\{\lambda_i | i = 1, 2, \cdots, m\}$. This will work when S_W is non-singular.

However, since the rank of S_W is at most $N - c$ and the number of images in the learning set N is much smaller than the number of pixels in each image n, one can observe that the within-class scatter matrix S_W is always singular. This means that it is possible to choose the matrix W such that the within-class scatter of the projected samples can be made exactly zero.

To overcome this issue aroused by a singular S_W, the Fisherfaces combines FLD and PCA. In specific, it performs PCA to reduce the dimension of the feature space to $N - c$, and then applying the standard FLD to further reduce the dimension to $c - 1$. Formally, the optimal projection W_{opt} in Fisherfaces algorithm is given by

$$W_{opt}^T = W_{fld}^T W_{pca}^T$$

where

$$\begin{aligned} W_{pca} &= \arg\max_{W} |W^T S_T W| \\ W_{fld} &= \arg\max_{W} \frac{|W^T W_{pca}^T S_B W_{pca} W|}{|W^T W_{pca}^T S_W W_{pca} W|} \end{aligned}$$

Here,

$$S_T = \sum_{i=1}^{N} (\mathbf{x}_k - \mu_i)(\mathbf{x}_k - \mu_i)^T$$

where n is the number of sample images, and $\mu \in \mathbb{R}^n$ is the mean image of all samples. Further experiments have indicated that Fisherfaces appears to outperform Eigenfaces in exterior variations including lighting, as well as simultaneously handling variation in lighting and expression. [2]

In our system, we take the same methodology as Fisherfaces, but we rely on it to perform facial expression recognitions instead of face recognitions as it was originally suggested for.

3. Our Approach

The highly versatile facial expression recognition system we are suggesting works in the following approaches. When an image is fed to the system:

1. **Face Localization:** The Viola-Jones object detection framework is used to find a human face that appears in the image; For each face found, crop that area out and use Viola-Jones object detection framework again to find their eyes, then align the face by letting the line connecting two eyes turn to horizontal. At last, crop aligned faces out again, reshape them to 48 × 48 pixels and convert to gray-scale.

2. **Feature Extraction:** After the face is detected, a converted Fisherfaces algorithm is employed to reduce the dimension of features demonstrated on the face; the algorithm incorporates PCA and FLD in the same manner as the original Fisherfaces does, but uses the model to process features for facial expressions instead of human faces.

3. **Facial Expression Classfication:** After the features on the facial expression have been processed by the converted Fisherfaces algorithm, it is fed to a 1-Dimension Convolutional Neural Network Classifier for expression classification. The neural network was trained for 100 epochs using hyperparameters as shown in table 1 below.

Layer	Output Shape	# of Parameters
Convolution 1D	1×32	60032
Leaky ReLU	1×32	0
Max Pooling 1D	1×32	0
Dropout 25%	1×32	0
Convolution 1D	1×64	6208
Leaky ReLU	1×64	0
Max Pooling 1D	1×64	0
Dropout 25%	1×64	0
Convolution 1D	1×128	24704
Leaky ReLU	1×128	0
Max Pooling 1D	1×128	0
Dropout 40%	1×128	0
Flatten	128	0
Dense	128	16512
Leaky ReLU	128	0
Dropout 30%	128	0
Dense	7	903

Table 1. The Hyperparameters used to train the 1-Dimension Convolutional Neural Network Classifier.

Note that in each layer, a portion of the data are dropped out so as to reduce the reliance of the trained model to the training data set, in attempt to prevent overfitting.

4. Experimental Results

4.1. Face Databases

We evaluate the proposed method on well-known publicly available facial expression databases: CK+48 [4], and FER2013 [1]. In this section we briefly review the content of these databases.

CK+48: The Cohn-Kanade (CK) [4] database was released for the purpose of promoting research into automatically detecting individual facial expressions. Since then, the CK database has become one of the most widely used testbeds for algorithm development and evaluation. The Cohn-Kanade (CK) [4] database includes 593 video sequences recorded from 123 subjects ranging from 18 to 30 years old. Subjects displayed different expressions starting from the neutral for all sequences, and some sequences are labeled with basic expressions. [5] And CK+48 is a cropped version of CK+, it extracted the last three frames from each sequence in the CK+ dataset [4], which contains a total of 981 facial expressions.

FER2013: The Facial Expression Recognition 2013 (FER-2013) database was introduced in the ICML 2013 Challenges in Representation Learning. [1] The database was created using the Google image search API and faces have been automatically registered. Faces are labeled as any of the six basic expressions as well as the neutral. The resulting database contains 35,887 images most of them in wild settings. [5]

Model	Database	Eigenfaces	Fisherfaces
CNN	CK+48	78%	99%
	FER2013	32%	35%
SVM	CK+48	32%	99%
	FER2013	58%	59%
Adaboost	CK+48	70%	97%
	FER2013	31%	34%
MLP	CK+48	85%	95%
	FER2013	34%	35%

Table 2: Average accuracy for all models on CK+48 and FER2013 database

4.2. Results

For each database, we split it into three parts: training, validation and test sets with a ratio of 0.7:0.15:0.15. In the evaluation process, we use GridSearchCV to exhaustively search over and tune the hyperparameters of the models and calculate the confusion matrix, mean squared error and evaluate the accuracy for each different models. Table 2 reports the average accuracy when classifying the images into the seven basic expressions. According to the table 2, overall all of the models generate a fairly better result on CK+48 database compared with FER2013. Since in the FER2013 database, there are many tilted faces and also faces covered by the corresponding subject's hands, thus leading to unsatisfactory results. Figures 5, 6 and 7, 8 show the projection result onto the first 2 principal components and linear discriminants of eigenfaces and fisherfaces respectively on database CK+48 and FER2013. We can clearly see that the projection result for CK+48 is much better than that of FER2013, which accords the accuracy result in the table 2. In addition, fisherfaces performs better than eigenfaces on database CK+48 while on the FER2013, the eigenfaces and fisherfaces have the similar projection result according to Figure 7 and 8. In summary, for CK+48 database all models perform better and achieve around 95% with algorithm fisherfaces compared with eigenfaces. But on FER2013 database, except that SVM has a around 60% average accuracy, the rest of models achieve around 31% accuracy for both fisherfaces and eigenfaces.

Please be advised: Since we randomly shuffled our datasets into training, test and validation sets in data preprocessing step, these accuracies might be slightly different in each run.

5. Conclusion and Future Work

In this report, a highly versatile facial expression recognition system is proposed as a solution of face detection and

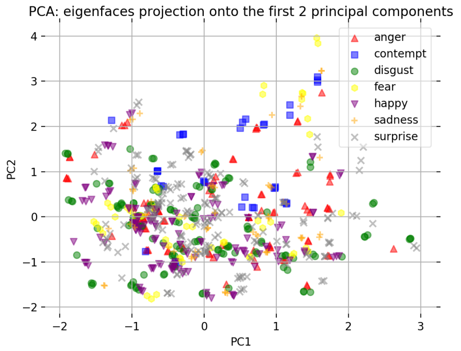

Figure 5. PCA: eigenfaces of database CK+48 projection onto the first 2 principal components.

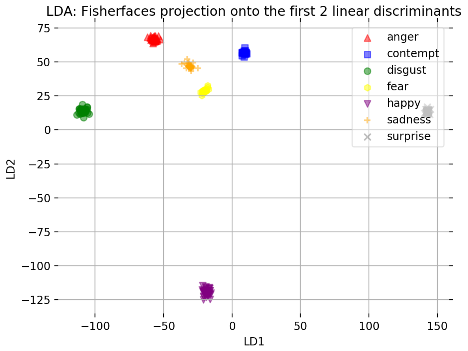

Figure 6. LDA: fisherfaces of database CK+48 projection onto the first 2 linear discriminants.

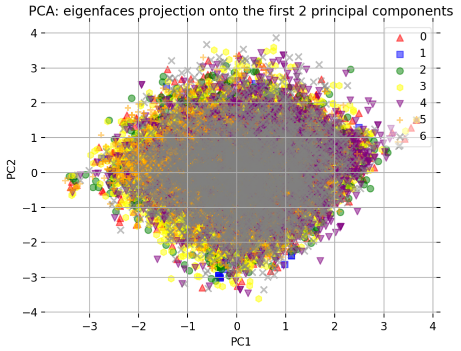

Figure 7. PCA: eigenfaces of database FER2013 projection onto the first 2 principal components. [2]

emotion analysis. And we also combine the eigenfaces and fisherfaces algorithms with several different models to classify the images into some basic expressions and evaluate their accuracy respectively. Due to the computational constraints of training and evaluating, many techniques to improve performance was not employed in this project. In particular, some additional techniques including weighting the loss for class balancing, pre-training a more complex network with more layers using bigger filters such as ALexNet and VGGNet on a large database were left unexplored. In the future work, we would use a bigger and more varied database with equal race distributions to eliminate bias and pre-train other CNN's with more complex architecture. And we would also apply the method of weighting the loss for class balancing to improve the overall performance of the system.

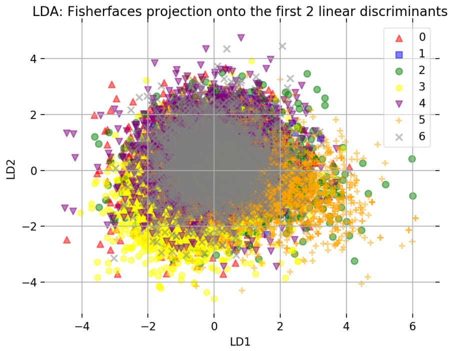

Figure 8. LDA: fisherfaces of database FER2013 projection onto the first 2 linear discriminants. [2]

6. Authors' Contributions

- Wei Cui: Refactoring all the codes, codes relate to data preprocess, feature extraction, model construction and evaluation using CNN, final report.
- Kewei Qiu: Codes relate to face detection and alignment, model construction and evaluation using SVM, AdaBoost and MLP, final report.
- Chenhao Gong: Codes relate to model construction and evaluation using CNN, presentation slides, README.md and final report.

References

[1] Challenges in representation learning: Facial expression recognition challenge. https://www.kaggle.com/c/challenges-in-representation-learning-facial-expression-recognition-challenge. [Online; accessed December 17 2020].

[2] P.N. Belhumeur, J.P. Hespanha, and D.J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):711–720, 1997.

[3] C.A. Corneanu, M.O. Simón, J.F. Cohn, and S.E. Guerrero. Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related

applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38:1548–1568, 2016.

[4] P. Lucey, J. 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. pages 94 – 101, 07 2010.

[5] A. Mollahosseini, D. Chan, and M. H. Mahoor. Going deeper in facial expression recognition using deep neural networks. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–10, 2016.

[6] K. Sung and T. Poggio. Example-based learning for view-based human face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(1):39–51, 1998.

[7] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.

[8] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2001), with CD-ROM, 8-14 December 2001, Kauai, HI, USA*, pages 511–518. IEEE Computer Society, 2001.

[9] Y. Yoshitomi, N. Miyawaki, S. Tomita, and S. Kimura. Facial expression recognition using thermal image processing and neural network. *RO-MAN*, pages 380–385, 1997.