GROUP 36: VALIDATION ON FEATURE EXTRACTION AND NEURAL NETWORK METHODS FOR DEEP LEARNING FACIAL RECOGNITION

Ivan Ferrier[⋆]*, Rom Tendencia*[⋆]*, and Yunzhang Liu*[⋆]

[⋆]University of California San Diego, La Jolla, CA 92093-0238

ABSTRACT

Facial recognition and identification has been applied to a variety of areas in the modern world. Yet, a challenge to solving a realistic problem is to appropriately identify an efficient neural network model which can work reliably.
This project is aimed to construct an optimal machine learning model on the facial images from the ORL Database that can exceed human performance in executing facial identification tasks. For feature extraction, we implemented three different methods: principal component analysis(PCA), linear discriminant analysis(LDA) and independent component analysis(ICA). We applied support vector machine(SVM), linear retgression(LR) and LDA classifier on the transformed dataset and the highest accuracy we get is 98.00%. We also used convoluted neural network(CNN) which gives an average 97.75% accuracy.

Index Terms— Facial recognition, feature extraction, PCA, ICA, LDA, CNN

1. INTRODUCTION

Facial recognition technology (FRT) is a visual pattern recognition capable of identifying human faces from video frames or digital images. FRT utilizes a software to map the face of an individual's features then the data is stored as a face template.[1] Facial recognition is a category of biometric security and are key enablers to enhance the safety and security in a wide range of industries, including banks, law enforcement, and healthcare.[2]
Humans have an innate ability to process and detect the perception of human faces. There are regions of the brain that allow pattern identification and storage for thousands of individuals. This perception is dependent on identifying specific features, such as the eyes, nose, and mouth, as well as the perceiving specific spatial arrangement of those features.[3] FRT adopts this salient social perception skill and is the fundamental operation of this technology. The potential of this technology provides a sophisticated surveillance technique to enhance the security and safety capabilities than can be more accurate than the human eye. The application of FRT have garnered attraction as a promising solution to address the needs for identification and verification of identity claims.

1.1. Literature Work

The novel method for face classification originates back to the nineteenth century as (author) proposed to collect the curves of facial profiles, compute their norm, then classify other facial profiles by their deviations from the norm.[4] The basis of this proposal demonstrated rapid progress in data development in real-world settings. The contribution factors that have established an impact to these developments include large databases of facial images that are readily available online, the active development and improvement of algorithms, and the methods of performance metrics of face recognition algorithms. In literature, FRT is approached in a generalized 4 step process. (1) The detection of facial features or patterns. (2) Normalization of facial images to account for geometrical and illumination changes. (3) Identify the faces using appropriate classification algorithms. (4) Verify model algorithm tolerance for logistic feedback. [5]

1.2. Modern Applications

In commercialized applications, FTC implements still (static) matching that range from photos on credit cards, passport, driver's license, and high security work environments that require access control. The FRT is predominately utilized in law enforcement that implement both static matching and real-time matching which include mug shots or video surveillance tapes.
Despite the recent accomplishments made thus far, challenges, determination of the method and technique, can potentially hamper the performance and accuracy level of facial recognition. This project is aimed to construct an optimal machine learning model on the facial images from the ORL Database that can exceed human performance in executing facial identification tasks. The methodology of approach will be to implement three feature extraction methods: principal component analysis(PCA), independent component analysis(ICA), linear discriminant anal-

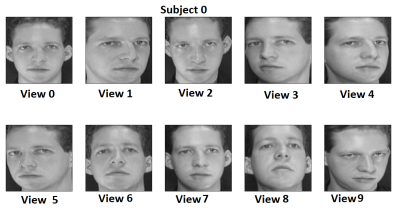

Fig. 1. 10 images of the first person

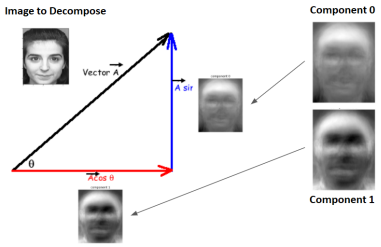

Fig. 2. Decomposing vectors vs decomposing images

ysis(LDA). In addition to a neural network method with a convolutional neural neural (CNN) architecture. This work provides a basis and guidance for future research on facial recognition.

2. METHODS

2.1. Dataset

We used the ORL Database of Faces as our dataset. The ORL dataset consists of images of 40 different people with 10 distinct images for each person. The images of one person are taken at different times, different angles and with different facial expressions as in Fig. 1. Each image is of size 112x92 pixels with 256 grey levels per pixel.

We used 300 images for training and used 100 images for testing/validation.

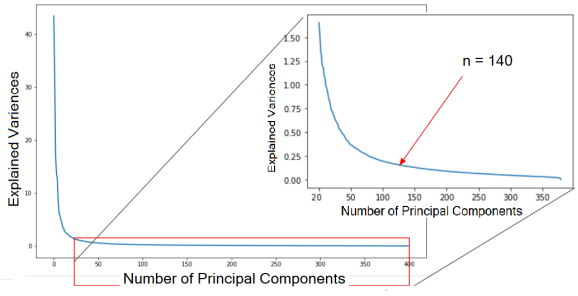

Fig. 3. Variance vs number of principal components

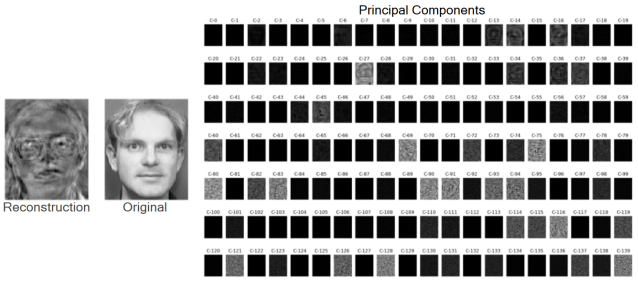

Fig. 4. Original images vs reconstruction 140 components
The values used to construct the image (left) are multiplied by the principal component photos, (so the lighter the component appears, the more prevalent that component was in reconstructing this image)

2.2. Feature Extraction

2.2.1. Principal Component Analysis

Principal component analysis is a method used to project higher dimensional data onto a lower dimensional space in such a way that the resulting lower dimensional data will be easier to classify than the original. The first step to PCA is to choose the number of principal components(n) desired. From there, n matrices will be found such that the sum of the error after reconstruction will be minimized. These matrices are the principal components. This is similar to breaking a vector up into its x and y components. Fig. 2 Instead of Va = (10x,7y) and Vb = (3x,9y) we can shorten things by listing Va = (10,7) and Vb = (3,9) with principal components x and y. note, that if we were trying to reconstruct Vc = (3x, 8y, 2z) only using principal components x and y, the closest we could come is Vc = (3,8) and the z component is lost. In this example the x and y vectors are trivial $x = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$, $y = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$ and there is only a marginal benefit to using PCA. However, the same example can be used where x and y are much larger. Lets say we have two (m*n) photos and we find that we can represent their flattened versions as (10x,7y) where x and y are (mxn*1) vectors. In this scenario, PCA can dramatically reduce the dimension space of the data, but it is likely that a lot of the data will be lost. This will make reconstruction more difficult, thereby making the PCA reconstruction less useful. Increasing the number of principal components solves this problem, but there is a trade-off. The more components, the closer the reconstruction, but the less the dimensionality is reduced. For this reason, we want to increase the number of principal components (n) until there is negligible change in the explained variance between n and n-1. In the case of our data set, by plotting the explained variance vs the number of components Fig. 3, we found 140 principal components have a small enough change in explained variance, while still significantly reducing the dimension of the data. After

the principal components have been found for the training data, new images can be projected onto the principal components and the resulting representation can be compared with other, labeled representations in order to classify the new image. An example of an image being reconstructed using 140 components can be seen in Fig. 4.

2.2.2. *Independent Component Analysis*

Independent component analysis (ICA) is an unsupervised learning technique that is a generalization such that it separates the the high-order moments of the input in addition to the second-order moments. Thus, this machine learning technique separates independent sources from a mixed signal into individual components from the maximum mutual information transfer. Unlike PCA which focuses on maximizing the variance of the data the data points, the ICA focuses on independence. The method of approach can be shown in Fig. 5.
Algebraically, **X**=a**S**.
where **X** is number of components(or observations), a is the unknown mixing matrix, **S** is the vector of independent source components. Or it can be expressed as **S**=w**X**. Where w is the weighted matrix.
Two key assumptions must be met prior to implementing this technique.
(1) The individual components are assumed to be statistically independent of each other – Leading to the joint distribution of two variables x and y: **p(x,y) = p(x)p(y)**
(2) The independent components are non-gaussian. Therefore, each source signal have non-gaussian distribution values.
The algorithm for ICA consist of 7 steps:
(1) Center **X** by subtracting the mean.
(2) Whiten **X**
(3) Choose initial values for **W** (mixing matrix)
(4) Calculate new **W**
(5) Normalize **W**
(6) Check whether the algorithm has converged, if not return to **(4)**
(7) Take the dot product of **X** and **W** to determine independent sources **S**

2.2.3. *Linear Discrimination Analysis*

Similar to PCA, linear discrimination analysis(LDA) also projects the higher dimensional data points onto a lower dimensional space. The difference between these two projections as shown in Fig. 6, is that the lower dimension of LDA is constructed by the directions that maximizes the separation between different classes. This means that LDA has to be supervised with class labels specified.
Computing the LDA transformation consists of 5 steps.
(1) Compute the main vectors m_i for each class of X

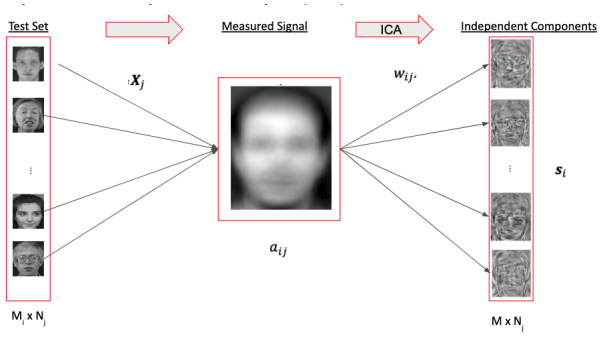

Fig. 5. Independent component analysis method approach

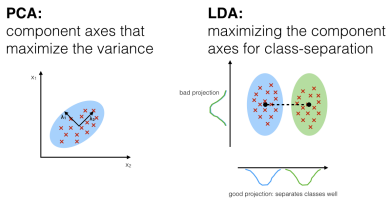

Fig. 6. Difference between PCA and LDA. [6]

(2) Compute the scatter matrices.(Within class scatter: $S_W = \sum S_i$) (Between class scatter: $S_B = \sum N_i(m_i - m)(m_i - m)^T$) where $S_i = \sum_{\text{x in class i}}(x - m_i)(x - m_i)^T$, N_i is the sample size of each class and m is the overall mean
(3) Find the eigenvalues and corresponding eigenvectors of the matrix ${S_W}^{-1}S_B$
(4) Choose k eigenvectors with the largest eigenvalues and form the transformation matrix $W = [\upsilon_1\ \upsilon_2\ \ldots\ \upsilon_k]$
(5) Transformed dataset $Y = X * W$

2.3. Neural Network

2.3.1. *Convolutional Neural Network*

The idea of Convolutional Neural Network on image classification comes from the human visual organization called Visual Cortex[7]. Human neurons can only respond to a restricted region of visual field and the collection of these fields from neurons overlap to cover the entire spectrum of sight.
Inspired by this idea, CNN models are constructed by different layers as shown in Fig. 7.The tasks of each convoluted layer is to extract the 'high-level features' [7] so that the model can have a better understanding of the image. The pooling layers are used to reduce the spatial size of the features extracted while maintain those that have higher weights.

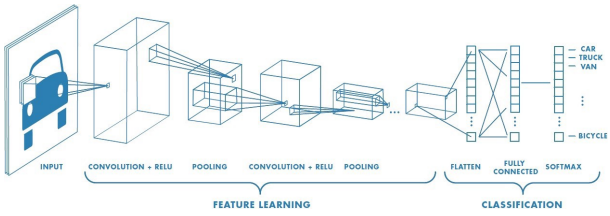

Fig. 7. Sample Structure of the CNN. [7]

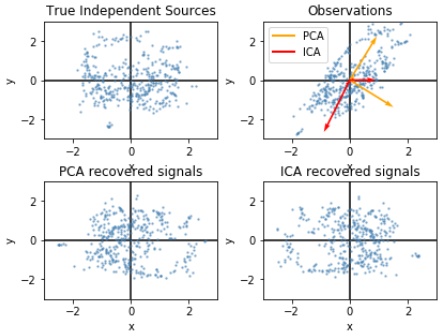

Fig. 8. Independent component analysis Statistics

3. RESULTS

3.1. Model Structures and Parameters

3.1.1. Models with Feature Extraction

We transformed our flattened training and testing images by PCA with n-component = 140. Then we applied LDA classifier.
We transformed our flattened training and testing images by ICA with n-component = 140. Then we applied LDA classifier.
We used the maximum $k = n_{classes} - 1 = 39$ as parameter for LDA transformation and transformed our flattened training images and testing images.
Then we used SVM classifier with C = 1e7 and linear regression(LR) classifier with C = 1.0.

For simplicity, the statistics for the ICA are shown in Fig. 8 with n-components(or observations) equal to 2. The true independent sources with highly non-Gaussian process (top left). The test data is mixed with the observations (top right). The PCA displays directions that are represented by orange vectors. The signal in the PCA space, is represented after the whitened process by the variance which correspond to the PCA vectors (bottom left). The ICA corresponds to finding a rotation in the raw space to correctly identify the largest non-Gaussian vector directions, which are non-orthogonal(bottom right).

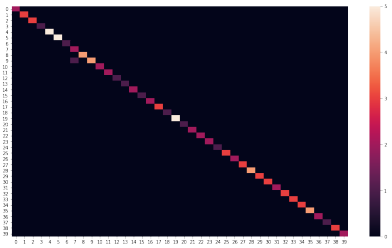

Fig. 9. Confusion matrix for CNN-final

3.1.2. Neural Network

We used a base cnn model and a final cnn model. Base model contained one Conv2D-Pooling-Normalization(CPN) combination and final model contains two CPN combinations. Both models have a drop(0.25) layer after each combination. A Flatten layer and a dense(256, relu) layer is added after the convolution layers. We used a dense layer with softmax as the output layer. For compiling we used adam as optimizer, categorical cross-entropy as loss function and accuracy as metrics, as shown in Fig. 9.

3.2. Experiment Result

Feature Extraction	Classifier	Accuracy(%)
PCA(n-comp=140)	LDA	95.00
ICA(n-comp=140)	LDA	97.00
LDA(n-comp=39)	SVM(C=1e7)	97.20
LDA(n-comp=39)	LR(C=1.0)	98.00
-	CNN-base	94.00
-	CNN-final	97.75

Table 1. Accuracy for each algorithm

4. CONCLUSION

In this project dimensionality-reduction and neural networking were implemented as methods for deep learning facial recognition to investigate the validity of various techniques. For dimensionality-reduction, or feature extraction, the Principal Component Analysis, and Linear Discriminant Analysis demonstrated an effective accuracy of 95.00%, 97.00%, and 97.20%, respectively and can be shown in Tab. 1. For neural networking, the architecture of base and final Convolutional Neural Networks demonstrated average accuracy of 94.00% and 97.75% respectively as shown in Tab. 1 as well.
From our observations, CNN and a combination of extraction-classifiers yield the best results. We believe that additional algorithms in combination could enhance the efficiency of the accuracy overall.

5. REFERENCES

[1] N. Martinez. What are important ethical implications of using facial recognition technology in health care? *AMA J. Ethics*, 2:180–187, 2019.

[2] M. Mann; M. Smith. Automated facial recognition technology: Recent developments and approaches to oversight. *U.N.S.W.L.J.*, 1, 2017.

[3] C.; Ribarsky W.; Chang R. Jeong, D.; Ziemkiewicz. Understanding principal component analysis using a visual analytics tool. 2009.

[4] K. M. Apampa; G. Wills; D. Argles. An approach to presence verification in summative e-assessment security. *2010 International Conference on Information Society; IEEE*, page 647–651, 2010.

[5] R. Chellappa; C. L. Wilson; C. Sirohey. Humain and machine recognition of faces: A survey. *Proc. IEEE*, 2:705–740, 1995.

[6] Sebastian Raschka. Linear Discriminant Analysis -Bit by Bit. Online - Accessed on Jun 11, 2021.

[7] S. A Saha. comprehensive guide to convolutional neural networks — the ELI5 way . Online - Accessed on Jun 11, 2021.

Individual Contributions

As a whole:

- We met twice a week and worked together as a whole.

Ivan Ferrier:

- Uploading data as a .npy file
- Examining principal component analysis and its efficiency when combined with LDA for classifying faces
- Designed graphics representing the underlying ideas behing PCA
- Compared efficiency of PCA with a LDA LR pipeline method

Yunzhang Liu:

- Reading and visualizing the ORL dataset.
- Examining linear discrimination analysis and construction convolutional neural network models.
- Answering review questions regarding CNN and making adjustments.
- Collecting Results.

Rom Tendencia:

- Examined, implemented, and analyzed the methods PCA and ICA's efficiency and accuracy.
- Formatted the LaTex structure for the final Report.
- Designed the graphics for the presentation.

Critical Reviews

Critical Review from **Team 16:**

1. Can you specify what you are planning to do in your further work to make the model have higher accuracy?
2. Based on my knowledge, CNN should have higher accuracy than the traditional machine learning methods, therefore the conclusion that LR has the highest accuracy is unconvincing to me. Did you try different CNN structures to this problem to see whether the accuracy could be improved?

Response:

1. We are going to try some different model parameters or structures. Also we are going to compute average accuracy to have more convincing result.
2. We have tried some other custom structures but the structure we showed in presentation provided the best accuracy. However, we figured out that the accuracy of our model did vary in different runs. So we have it ran 4 times and get the average accuracy 98% on test set.

Critical Review from **Team 19:**

1. For CNN part, it seems that you are making your custom CNN. Could you explain why you use this specific architecture and maybe explain some architectures you have tried but fail to predict.

Response:

1. First, we use drop out between main layers to avoid over-fitting and use dense layer as output layer to calculate loss. The reason we use two Conv2D-Pooling-Normalization(CPN) layer structure is that we get 92% accuracy when we only use one OPN layer and using three CPN layer does not provide noticeable increase on test set accuracy.

Critical Review from **Team 20:**

1. Would be nice to give informative insight into how the problem arises and why this needs to be addressed.
2. Would be nice to give an introduction for other peoples work upon this topic, talk about different approaches developed in order to solve this problem, it helps to further understand the topic.
3. What was the training/validation/test data used during training?
4. What was the further Improvements decided to implement? Would you try different combinations of feature extraction method and classifiers?

Response:

1. We have added discussion of why the problem is important in the Introduction section.
2. We have added literature work as well as modern approaches on this problem in the Introduction section.
3. We used 300 out of 400 images as training set and 100 images as testing set.
4. We are going to try some different model parameters or structures. Also we are going to compute average accuracy to have more convincing result.