

Visual Perception in Face Recognition

Research work by:

- Carlos Thomaz and Vagner Amaral
Centro Universitario FEI, Sao Paulo
- Gilson Giraldi
Laboratorio Nacional de Computacao Cientifica
Rio de Janeiro
- Duncan Gillies and Daniel Rueckert
Imperial College London

History of Face Recognition

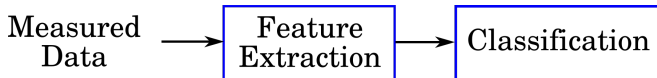
Karl Pearson (1857-1936) - study of biometrics.

Biometrika Journal founded 1901

Biometrika shall serve as a means of publishing biological data not systematically collected elsewhere, and also of spreading a knowledge of such statistical theory as may be requisite for their scientific treatment.

History of Face Recognition

Biometrics, like modern computer vision was conceived as a two part process:

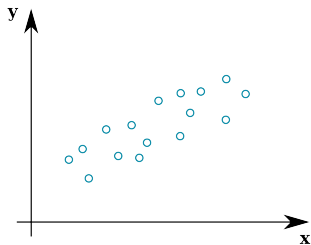


Pearson's work focused on taking physical measurements and modelling them using Gaussian (or higher order) distributions.

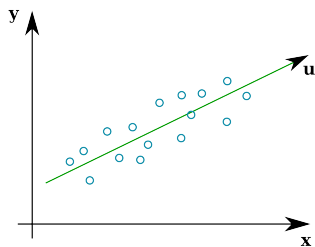
He was one of the inventors of **Principal Component Analysis (PCA)** which is used to find the most compact representation of a set of independent variables.

PCA in Outline

For a given data set



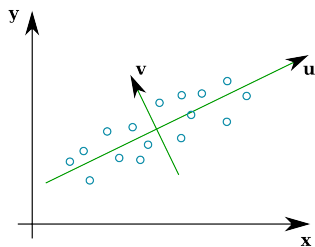
PCA in Outline



For a given data set

We determine the direction where the variation is the greatest

PCA in Outline

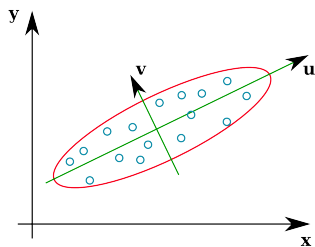


For a given data set

We determine the direction where the variation is the greatest

Then we find the direction where the remaining variation is the greatest

PCA in Outline



For a given data set

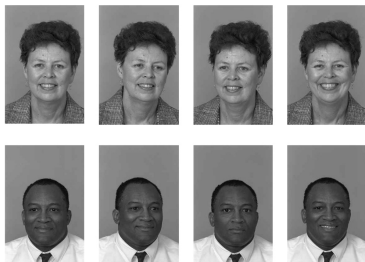
We determine the direction where the variation is the greatest

Then we find the direction where the remaining variation is the greatest

We continue the process and thus find the coordinate system that most compactly represents the data.

Face recognition from 2D Images

PCA can be used in face recognition for **feature selection** and **data reduction**.



Face images are highly redundant:

- All the background pixels are the same
- Each subject has the same facial features

Sirovich and Kirby's Approach (1987)

Pearson could only work on small problems because he did not have any significant computing resources.

However by 1987 it was possible to work on problems with up to 20,000 variables.


In Sirovich and Kirby's method an input image with n pixels is considered a point in an n -dimensional space called the image space.

$$\mathbf{p}_x = (i_1, i_2, i_3, \dots, i_n)$$

Each **pixel** is considered a **variable** with a value for each image in the data base.

Converting an Image to a vector

Given a greyscale image of, for example, 128 by 128 pixels:


$$= \begin{bmatrix} 150 & 152 & \cdot & 151 \\ 131 & 133 & \cdot & 72 \\ \cdot & \cdot & \cdot & \cdot \\ 144 & 171 & \cdot & 67 \end{bmatrix} 128 \times 128$$

We concatenate each row to make a 16384 vector

$$[150, 152, \dots, 151, 131, 133, \dots, 72, \dots, 144, 171, \dots, 67]_{16K}$$

Dimension Reduction

In the data space each pixel is a variable, so the dimension of the space is very high (min 16K).

Dimension reduction is achieved by PCA.

Let an $N \times n$ data matrix D be composed of N input face images with n pixels. Each row is one image of our data set.

$$D = \begin{bmatrix} 150 & 152 & \cdots & 254 & 255 & \cdots & 252 \\ 131 & 133 & \cdots & 221 & 223 & \cdots & 241 \\ \cdot & \cdot & & \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot & \cdot & & \cdot \\ 144 & 171 & \cdots & 244 & 245 & \cdots & 223 \end{bmatrix} N \times n$$

Mean Centring the data

Suppose the mean of the columns of D (the average image) is:

$$[120 \ 140 \ \dots \ 230 \ 230 \ \dots \ 240]$$

The origin is moved to the mean of the data by subtracting this average image from each row. This creates the mean centred data matrix:

$$U = \begin{bmatrix} 30 & 12 & \dots & 24 & 25 & \dots & 12 \\ 11 & -7 & \dots & -9 & -7 & \dots & 1 \\ \cdot & \cdot & & \cdot & \cdot & & \cdot \\ 24 & 31 & \dots & 14 & 15 & \dots & -17 \end{bmatrix} N \times n$$

Calculating the covariance matrix

The covariance matrix Σ can be calculated easily from the mean centered data matrix:

$$\Sigma = U^T U / (N - 1)$$

N is the number of data points (images) and the covariance matrix has dimension $n \times n$.

Finding the PCA space

The axes of the PCA space are the eigenvectors of the covariance matrix.

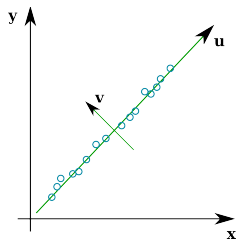
We use linear algebra to find Φ and Λ that satisfy:

$$\Sigma = \Phi\Lambda\Phi^T$$

The eigenvectors ϕ_i form an orthonormal basis:

$$\forall \phi_i, \phi_j \in \Phi, \phi_i \cdot \phi_j = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

Data Reduction



Eigenvectors with low eigenvalues contribute little information in the data representation. Data reduction is achieved by ignoring the eigenvectors with low eigenvalues.

The set of m ($m < n$) eigenvectors of Σ which have the m largest eigenvalues, minimises the mean square reconstruction error over all choices of orthonormal basis of size m .

Data Reduction

Although n variables are required to reproduce an original sample X exactly, much of the significant variability in the data can be accounted for by a smaller number m of principal components.

Thus, the original data set consisting of N examples on n variables can be reduced to a data set consisting of N examples on m principal components (eigenvectors).

In face recognition the eigenvectors are often called eigenfaces.

Practical Face recognition

As an example we will find the eigenface basis of a set of fourteen faces images of resolution 384×256 . This initial data set D is sometimes called the training data set.



What do the eigenfaces look like?

Mean:



The four eigenfaces with the largest eigenvalues:



Reconstructing the face images: example 1

Original:



3 PCs



5



8



11



13



Reconstructing the face images: example 2

Original:



3 PCs



5



8



11



13



Correspondence in PCA

In 2D face images the pixels are not usually in correspondence. That is to say a given pixel $[x_i, y_i]$ may be part of the cheek in one image, part of the hair in another and so on.

This means that any linear combination of eigenvectors does not represent a true face but a composition of face parts.



Face Recognition

PCA extracts a small number of features from each high dimensional images. This makes the classification task easy.

In practice we can use any classifiers to find the best match in the data base for a test face:

- Linear Discriminant Analysis
- Support Vector Machines
- k-nearest neighbours
- Neural Network

Human vision

The above method is very effective and is used in practice in many places - **but!**

Is this how human vision works?

Human vision

The above method is very effective and is used in practice in many places - **but!**

Apparently not!

The Yarbus Experiment (1967)



Alfred Yarbus studied how humans look at images using eye tracking for the first time.

He used “Unexpected Visitors” by Ilya Repin as a stimulus.

The Yarbus Experiment (1967)



He found that the way we look at an image is guided by what we want to find out.

These eye fixations were in response to the task:

Examine the picture freely.

The Yarbus Experiment (1967)



These eye fixations were in response to the question:

What are the ages of the family?

The Yarbus Experiment (1967)



These eye fixations
in response to the
question:

What were the family
doing when the visitor
arrived?

The Yarbus Experiment (1967)



These eye fixations in response to the task:

Memorise the cloths that the characters are wearing.

The Yarbus Experiment (1967)

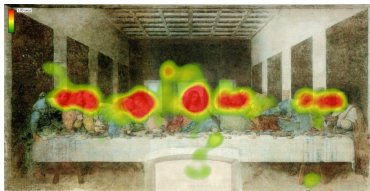
Yarbus' results were very influential and sparked off a wave of new research into human vision, much of it directed towards commercial ends.

It is tempting to think of fixation points as similar to feature points such as SIFT or SURF, but experiments to use them in this way have not been very successful.

Heatmaps



Most humans look at pictures in the similar ways. For Da Vinci's last supper observers will usually pay attention to the faces.



By collecting fixations from a large number of observers we can compute a density of fixations for each pixel of the image used as a stimulus.

This is conveniently displayed as a **heatmap**.

The heatmap assigns a “perceptual importance” to each pixel.

Making use of perceptual importance

How can we we make use of the perceptual information found in heatmaps?

Extracting specific information about the exact positions of fixations has not been very successful. - too little is known about the mechanisms of human perception.

However for face recognition (or similar tasks) we could incorporate perceptual information in the way we define the face space for classification.

The idea is to direct a classifier to examine the most significant pixels in the image.

Two class problems

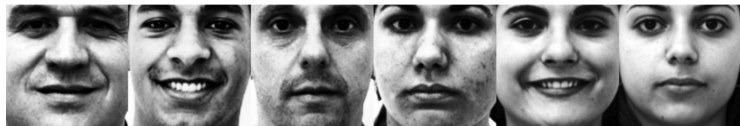
(Experiments by Carlos Thomaz at FEI)

Two class experiments were conducted on a set of faces including males and females with both smiling and with neutral expressions.



The images were rigidly registered so that the face parts were in corresponding positions in each image.

Data Gathering



The data gathering followed the Yarbus protocol.

Participants were asked to answer one of two questions:

1. Male or Female?
2. Smiling or neutral expression?

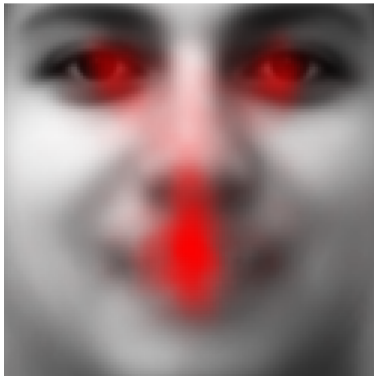
The fixations from the participants over all the faces were used to build a spatial attention map for both questions.

Typical Spatial Attention Maps

Gender



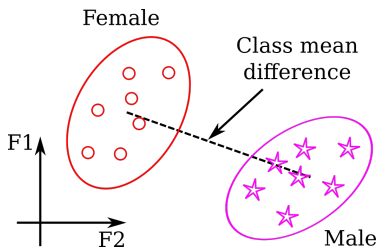
Expression



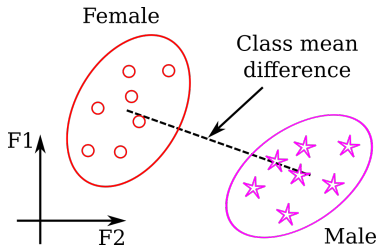
Statistical Feature Maps

It is interesting to compare the **spatial attention maps** with **statistical importance maps** that could be found by machine learning.

Since we have labelled data we can do this in a simple way by considering the means of the two classes in our feature space.



Statistical Importance

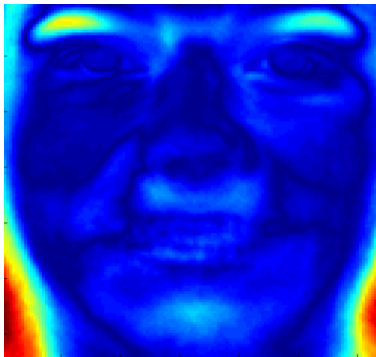


The difference of the class means is the direction in the feature space of the shortest route between the classes.

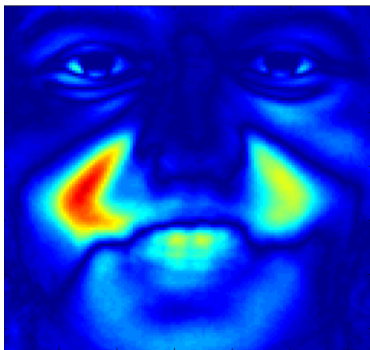
It provides us with an importance ranking for each pixel (dimension) - **just like a heat map!**

Typical Statistical Importance Maps

Gender



Expression



These are strikingly different from the spatial attention maps.

Making Use of Our Prior Knowledge

We have now found two different pieces of information which might be able to help us in building a better face recognition system.

1. Spatial Attention Maps
2. Statistical Importance Maps

Both take the form of a unit vector with an entry ranking the importance of each pixel in the image:

$$\mathbf{w} = [w_1, w_2, \dots, w_n]$$

So how can we make use of this information?

Weighting the individual pixels

PCA makes use of the covariance between each pair of pixels to find the directions where the variance is greatest.

If we multiply each pixel value x_j in the data set by its importance $\sqrt{w_j}$ then we change the face space as follows:

- Pixels with significant discriminative properties but low variance contribute more to the recognition process.
- Pixels with high variance but low discriminative properties contribute less.

Weighting the individual pixels

In terms of the covariance, the standard formula for covariance between pixels j and k :

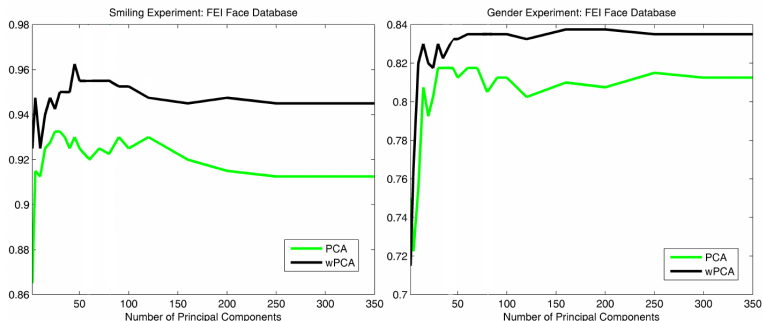
$$s_{jk} = \frac{1}{N-1} \sum_{i=1}^N (x_j^i - \bar{x}_j)(x_k^i - \bar{x}_k)$$

is replaced by:

$$s_{jk}^* = \frac{1}{N-1} \sum_{i=1}^N \sqrt{w_j}(x_j^i - \bar{x}_j)\sqrt{w_k}(x_k^i - \bar{x}_k)$$

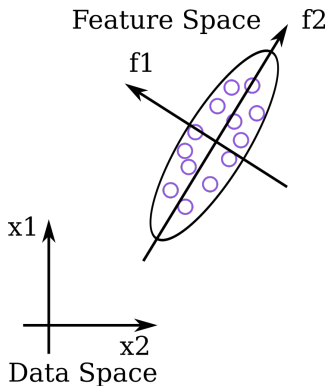
Experimental Results - Statistical Maps

Weighting the pixels using statistical importance maps based on the difference between the class means produced a significant improvement in recognition accuracy.



Ranking the PCA components

An alternative to weighting the pixels is ranking the standard principal components.



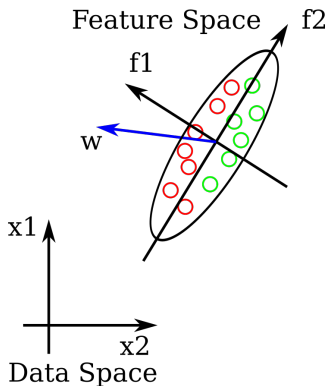
In a standard face space features are ranked according to their variance.

Here f_2 is considered more important than f_1 .

We discard features with very low variance.

Ranking the PCA components

An alternative to weighting the pixels is ranking the standard principal components.



As an alternative we can rank features according to how well they align with our prior knowledge.

Here f_1 is considered more important than f_2 .

We discard features that are out of line with our prior knowledge.

Experimental Results - Data Gathering

- 43 observers carried out the data gathering
- the stimulus images were shown for 3 seconds
- for gender recognition they were shown 30 male and 30 female images, all with neutral expressions
- for expression recognition they were shown 30 smiling images and 30 neutral images with equal proportions of male and female
- all observers classified the images with accuracy well above the random level.

Spatial attention maps were then constructed for each recognition task.

Experimental Results - Observation time

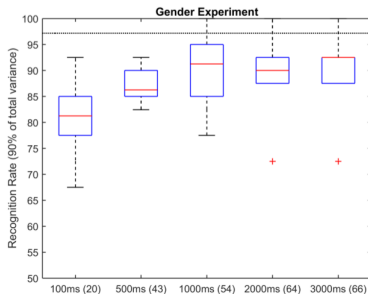
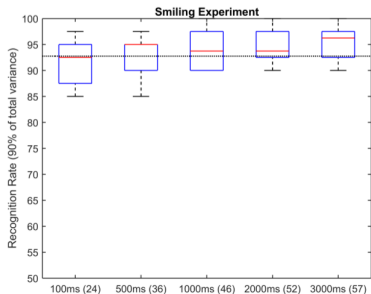
Spatial attention maps were collected at 0.1, 0.5, 1, 2 and 3 seconds.



The upper trace is for expression recognition and the lower trace for gender.

Experimental Results - Observation time

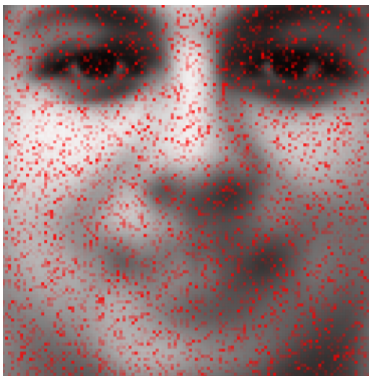
As the detail in in spatial cognitive maps increases we see a trend towards increasing accuracy



These results were obtained by weighting the individual variables.

Randomly generated map

To check the utility of the spatial cognitive maps a comparison was made with a map created by random sampling.



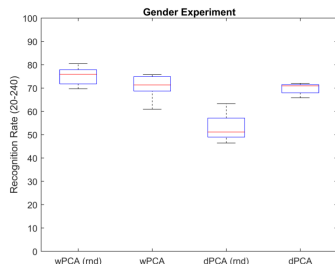
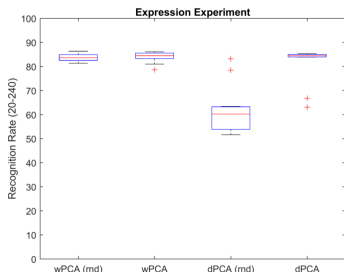
Experimental Results - Classification

To investigate the use of data from eye tracking four different feature spaces were tested:

- wPCA - the PCA face space using spatial attention maps to weight the pixels.
- wPCA (rnd) - the PCA face space weighted by a randomly generated map.
- dPCA - the PCA faces space with the features ranked by their agreement with the spatial attention direction.
- dPCA (rnd) - the PCA faces space with the features ranked by their agreement with a randomly generated map direction.

Expression Results

The results shown are the highest accuracies that were found using 20-240 principal components.



There is no significant difference between the wPCA, wPCA (rnd) and dPCA results. However the dPCA (rnd) results are significantly worse. This demonstrates the importance of the ranking of features in recognition.

Conclusions

We know that the technique of weighting the individual pixels works well when the weights have significant discriminant information. We saw this in practice with the statistical importance maps.

However, it does not work using the spatial attention maps obtained from eye tracking. Random weights can even out perform the attention maps.

Overall we must conclude that the fixation points in human vision are not specific feature points chosen for their discriminant information.

Conclusions

We know that the technique of weighting the individual pixels works well when the weights have significant discriminant information. We saw this in practice with the statistical importance maps.

However, it does not work using the spatial attention maps obtained from eye tracking. Random weights can even out perform the attention maps.

Overall we must conclude that the fixation points in human vision are not specific feature points chosen for their discriminant information.

Conclusions

The two methods for incorporation of the eye tracking information are radically different.

- Weighting the individual variables can be considered a **feature based** approach. We are trying to improve the feature space using the new knowledge.
- Ranking the features can be considered a **pattern based** approach. We are seeking the best directions to use.

The results suggest that human recognition is pattern based by nature.