Wavelet Transform And Principal Component Analysis Based Feature Extraction

Keyun Tong

June 3, 2010

As the amount of information grows rapidly and widely, feature extraction become an indispensable technique to extract the relatively most significant information from the given data. One benefits of this is that it allows people have better and faster understanding about the main property of a set of data which may contain noises. As long as the feature of the information is found, it can also used to compress, to compare or to identify some related information. One major application of feature extraction is in image analysis. With the increasing amount of images and videos in our life, it is significant to use some intelligent machines to identify the object shown in a picture, or to detect the occurrence of a specific object in the video.

One approach to extracting feature is using wavelet analysis introduced by MIT.[1] The wavelet transform used here is a efficiently computable Haar wavelet transform. Images are mapped from space of pixels to that of Haar wavelet features that contains rich descriptions of the pattern.

Another commonly used technique for feature extraction is the Principal Component Analysis (PCA), especially in face detection. What PCA do is project a set of images from a high dimensional space of pixels to a lower dimensional space which has the set of images as its main component. However, this method has a limitation that it cannot eliminate out noise well enough. Therefore, a better solution is to combine wavelet analysis with PCA, called Wavelet PCA, which can improve the result of feature extraction.[2]

1 The Haar System in Wavelet Transform

This section will first introduce an example of an orthonormal system on $[0, 1)$ known as the *Haar system*. The Haar basis is the simplest example of an orthonormal wavelet basis. Then we will develop a way to express a function as a composition of the *Haar scaling function* and the *Haar function*. Finally, a discrete version Haar transform is given to make it easy to compute.

1.1 The Haar System

Definition 1.1. *For each pair of integers j, $k \in Z$, define the interval $I_{j,k}$ by*

$$I_{j,k} = [2^{-j}k, 2^{-j}(k+1)).$$

The collection of all such intervals is called the collection of dyadic subintervals of R.

Remark. For any two dyadic intervals, either they do not overlap or one contains the other one.[3]

Definition 1.2. *Let $p(x) = \chi_{[0,1)}(x)$, and for each $j, k \in Z$, define the* ***scale j Haar scaling functions*** *to be*

$$p_{j,k}(x) = 2^{j/2}p(2^j x - k) = 2^{j/2}\chi_{I_{j,k}}(x).$$

Remark. For each $j, k \in Z$,

$$\int_R p_{j,k}(x)dx = \int_{I_{j,k}} p_{j,k}(x)dx = 2^{-j/2}$$

and

$$\int_R |p_{j,k}(x)|^2 dx = \int_{I_{j,k}} |p_{j,k}(x)|^2 dx = 1.$$

Definition 1.3. *Let $h(x) = \chi_{[0,1/2)}(x) - \chi_{[1/2,1)}(x)$, and for each $j, k \in Z$, define the* ***scale j Haar functions*** *to be*

$$h_{j,k}(x) = 2^{j/2}h(2^j x - k).$$

Remark. (a) It is obvious that $h_{j,k}$ is associated with the interval $I_{j,k}$, for

$$h_{j,k} = 2^{j/2}(\chi_{I_{j+1,2k}}(x) - \chi_{I_{j+1,2k+1}}).$$

(b) For each $j, k \in Z$,

$$\int_R h_{j,k}(x)dx = \int_{I_{j,k}} h_{j,k}(x)dx = 0$$

and

$$\int_R |h_{j,k}(x)|^2 dx = \int_{I_{j,k}} |h_{j,k}(x)|^2 dx = 1.$$

1.2 Orthogonality of the Haar System

Theorem 1.1. *The Haar system on R is an orthonormal system on R.*

Proof. To show that the Haar system is orthonormal, we need to show

$$\langle h_{j,k}, h_{j',k'} \rangle = \begin{cases} 0 & \text{if } j \neq j' \text{ or } k \neq k' \\ 1 & \text{if } j = j' \text{ and } k = k' \end{cases}$$

Suppose $j \neq j'$ or $k \neq k'$, then the two Dyadic intervals $I_{j,k}$ and $I_{j',k'}$ do not intersect. So $h_{j,k}(x)h_{j',k'}(x) = 0$, for the functions do not have positive values on the same point. Then, it is easy to see that

$$\langle h_{j,k}, h_{j',k'} \rangle = \int_R h_{j,k} h_{j',k'} dx = 0.$$

If $j = j'$ and $k = k'$, by the previous remark,

$$\langle h_{j,k}, h_{j',k'} = \int_R h_{j,k} h_{j',k'} dx = \int_{I_{j,k}} |h_{j,k}(x)|^2 dx = 1.$$

□

Remark. With similar proof, we can show that the Haar scaling functions $p_{j,k}$ is also orthonormal.

1.3 The Approximation and Detail Operators

Definition 1.4. *For each $j \in \mathbf{Z}$, define the approximation operator A_j on functions $f(x), L^2$ on $\mathbf{R}$, by*

$$A_j f(x) = \sum_k \langle f, p_{j,k} \rangle p_{j,k}(x).$$

Since $p_{j,k}$ are orthonormal, we can prove the following nice fact of the operator of A_j. To prove this lemma, see David F. Walnut[3].

Lemma 1.1. *(a) Given $j \in \mathbf{Z}$, and $f(x), L^2 on R$, $\|A_j f\|_2 \leq \|f\|_2$.*

(b) Given $f(x), C^0$ on $\mathbf{R}$, $\lim_{j \to \infty} \|A_j f - f\|_2 = 0$.

(c) Given $f(x), C^0$ on $\mathbf{R}$, $\lim_{j \to \infty} \|A_j f\|_2 = 0$.

Definition 1.5. *For each $j \in \mathbf{Z}$, define the detail operator D_j on functions $f(x), L^2$ on $\mathbf{R}$, by*

$$D_j f(x) = A_{j+1} f(x) - A_j f(x).$$

Remark. With some manipulation of the operator $A_j f$, we can show that

$$D_j f(x) = \sum_k \langle f, h_{j,k} \rangle h_{j,k}(x).$$

See [3].

1.4 Expansion in Term of Haar function

With all the previous definitions and lemmas set, we are ready to prove the major theorem.

Theorem 1.2. *Given $f(x), C^0$ on $\mathbf{R}$ and $J \in \mathbf{Z}$, f can be expanded as*

$$f(x) = \sum_{j=J}^{\infty} \sum_k \langle f, h_{j,k} \rangle h_{j,k}(x) + \sum_k \langle f, p_{J,k} \rangle p_{J,k}(x)$$

Proof. Given $\epsilon > 0$, by lemma 1.1(b), there is an integer $N > J$ such that $\|A_N f - f\|_2 < \epsilon$. By definition,

$$\begin{aligned} \sum_{j=J}^{N-1} D_j f(x) &= \sum_{j=J}^{N-1} [A_{j+1} f(x) - A_j f(x)] \\ &= \sum_{j=J+1}^{N} A_j f(x) - \sum_{j=J}^{N-1} A_j f(x) \\ &= A_N f(x) - A_J f(x). \end{aligned}$$

Therefore,

$$A_N f(x) = \sum_{j=J}^{N-1} \sum_k \langle f, h_{j,k} \rangle h_{j,k}(x) + \sum_k \langle f, p_{J,k} \rangle p_{J,k}(x)$$

Since $\|A_N f - f\|_2 < \epsilon$, we proved

$$f(x) = \sum_{j=J}^{\infty} \sum_k \langle f, h_{j,k} \rangle h_{j,k}(x) + \sum_k \langle f, p_{J,k} \rangle p_{J,k}(x)$$

□

Remark. What we have just proved can be also written as $f(x) \in \text{span}\{p_{J,k}(x), h_{j,k}(x)\}_{j \geq J, k \in \mathbf{Z}}$. In fact $f(x) \in \text{span}\{h_{j,k}(x)\}_{j,k \in \mathbf{Z}}$ also holds. See [3].

1.5 Discrete Haar Transform

By the previous lemma, if a function $f(x)$ is defined on $[0,1)$ rather than on $\mathbf{R}$, then given any integer $J \geq 0$, we have

$$f(x) = \sum_{j=J}^{\infty} \sum_{k}^{2^j-1} \langle f, h_{j,k} \rangle h_{j,k}(x) + \sum_{k}^{2^j-1} \langle f, p_{J,k} \rangle p_{J,k}(x)$$

In order to find the Discrete Haar Transform (DHT), assume that we are given a finite sequence of data of length 2^N for some $N \in \mathbf{N}, \{c_0(k)\}_{k=0}^{2^N-1}$. Also assume that for some underlying function $f(x)$, $c_0(k) = \langle f, p_{N,k} \rangle$. Fix $J \in \mathbf{N}, J < N$, and for each $1 \leq j \leq J$, define

$$c_j(k) = \langle f, p_{N-j,k} \rangle \text{ and } d_j(k) = \langle f, h_{N-j,k} \rangle.$$

There exists a convenient recursive algorithm that can be used to compute the coefficients $c_j(k)$ and $d_j(k)$ from $c_{j-1}(k)$.

$$\begin{aligned} c_j(k) &= \langle f, p_{N-j,k} \rangle \\ &= \langle f, p_{N-j+1,2k} \rangle / \sqrt{2} + \langle f, p_{N-j+1,2k+1} \rangle / \sqrt{2} \\ &= c_{j-1}(2k)/\sqrt{2} + c_{j-1}(2k+1)/\sqrt{2}, \end{aligned}$$

and also

$$\begin{aligned} d_j(k) &= \langle f, h_{N-j,k} \rangle \\ &= \langle f, h_{N-j+1,2k} \rangle / \sqrt{2} - \langle f, p_{N-j+1,2k+1} \rangle / \sqrt{2} \\ &= c_{j-1}(2k)/\sqrt{2} - c_{j-1}(2k+1)/\sqrt{2}. \end{aligned}$$

Therefore, we can now define the Discrete Haar Fransform in matrix form, which makes the calculation clear.

Definition 1.6. *Given $J, N \in \mathbf{N}$ with $J < N$ and a finite sequence $\{c_0(k)\}_{k=0}^{2^N-1}$, the DHT of c_0 is defined by*

$$\{c_J(k) : 0 \leq k \leq 2^{N-J} - 1\} \cup \{d_j(k) : 1 \leq j \leq J; 0 \leq k \leq 2^{N-j} - 1\},$$

with

$$\begin{pmatrix} c_j \\ d_j \end{pmatrix} = \begin{pmatrix} H \\ G \end{pmatrix} c_{j-1},$$

where

$$H = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & \cdots & 0 \\ & & & \vdots & & & \\ 0 & 0 & \cdots & 0 & 0 & 1 & 1 \end{pmatrix}$$

$$G = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & -1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & \cdots & 0 \\ & & & \vdots & & & \\ 0 & 0 & \cdots & 0 & 0 & 1 & -1 \end{pmatrix}.$$

Remark. Define matrix

$$W = \begin{pmatrix} H \\ G \end{pmatrix}.$$

Then matrix W is invertible, which allow us to reconstruct c_{j-1} from c_j and d_j by

$$c_{j-1} = W^{-1} \begin{pmatrix} c_j \\ d_j \end{pmatrix}.$$

1.6 The DHT in Two Dimensions

DHT can be widely applied in discrete signal processing. In some case, the signal might be two dimensional, like image data. Therefore, it is necessary to have a DHT for two dimensional case. Actually, two dimensional DHT is just a composition of the one dimensional DHT twice in rows direction and columns respectively.

Let c be a $M \times L$ matrix. Also let H^{row} and G^{row} be the same matrix as H and G but operate on every row of c, and let H^{col} and G^{col} be the same matrix as H and G but operate on every column of c. Now for simplicity, assume c is alway a square matrix that has 2^n rows

Definition 1.7. *Given $J, N \in \mathbf{N}$ with $J < N$ and a matrix $c_0 = \{c(m,n)\}_{n,m=0}^{2^N-1}$. For $1 \leq j \leq J$, define the $2^{N-j} \times 2^{N-j}$ matrices c_j, d_j^v, d_j^h, d_j^d by*

$$c_j = H^{col} H^{row} c_{j-1},$$

$$d_j^v = G^{col} H^{row} c_{j-1},$$

$$d_j^h = H^{col} G^{row} c_{j-1},$$

$$d_j^d = G^{col} G^{row} c_{j-1}.$$

2 Image Analysis with Discrete Haar Transform

2.1 Images in Mathematics

Images in mathematics are just described as two dimensional functions.

Definition 2.1. *A $N \times M$* ***gray image*** *is a function $f(n, m)$ maps $n, m \in \mathbf{Z}^+$ with $1 \leq n \leq N, 1 \leq m \leq M$ to $\mathbf{R}^+$.*

Definition 2.2. *A $N \times M$* ***color image*** *is a multi-value function $\mathbf{f}(n, m) = (R(n, m), G(n, m), B(n, m))$ maps $n, m \in \mathbf{Z}^+$ with $1 \leq n \leq N, 1 \leq m \leq M$ to $\mathbf{R}^{\mathbf{3}+}$, where function R, G, B represent the color decomposition on red, green and blue respectively.*

Although color image contains richer information, it is more complicated to deal with. Here we mainly focus on the analysis of gray image which is just a two-dimensional single valued function.

2.2 DHT on Image

Two dimensional DHT will decompose a image into four components. c_j represents the approximated image. In the DHT case, c_j is just averaging the values of every continuous 2^j pixels. d_j^h will give the horizontal detail information of the given image while d_j^v gives the detail information vertically. d_j^d is the main wavelet coefficient we will work on, which shows the detail information of the image in diagonal.

For example, given a image of c car

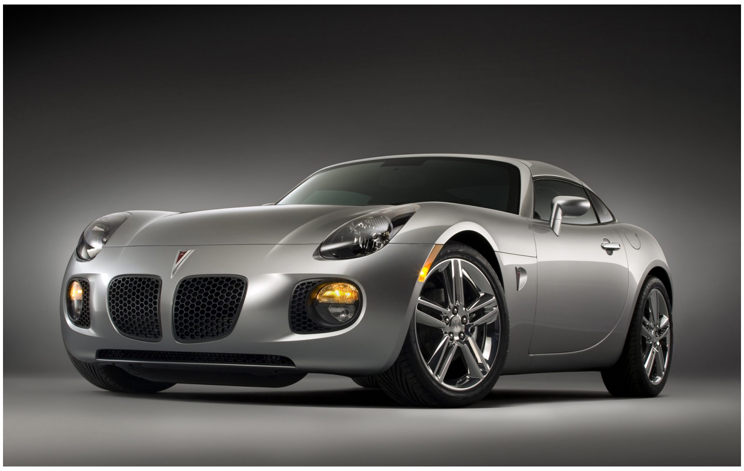

Figure 1: A color image of a car

We apply the two dimensional DHT on the image.

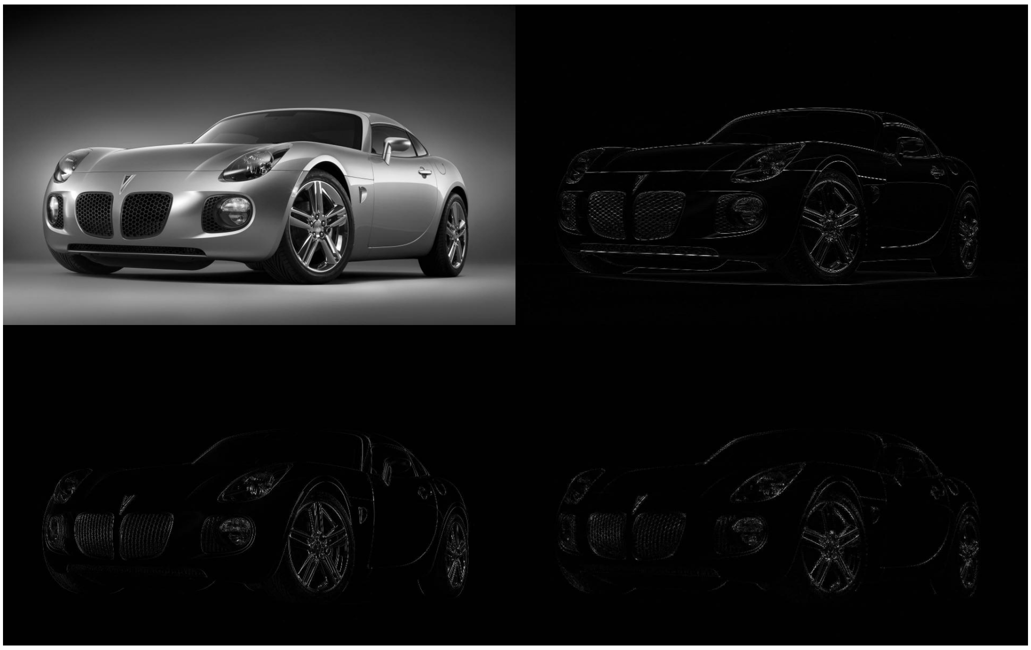

Figure 2: The coefficients of the scale 1 (i.e. $j = 1$) decomposing of the original image. Going clockwise from up-left corner, they are c_1, d_1^v, d_1^h, d_1^d respectively.

We can see, the vertically decomposing coefficients d_1^v show more clearly the horizontal profile line of the car and the horizontally decomposing coefficients d_1^h show more vertical profile line. d_1^d is more balanced. We call coefficients d_1^d the extracted feature of the image.

To see more concise extracted feature, we can apply DHT on the resulting coefficient c_1 to get a deeper decomposing of the original image.

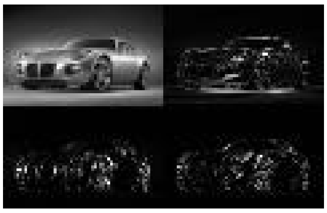

Figure 3: The coefficients of the scale 5 (i.e. $j = 5$)decomposing of the original image. Going clockwise from up-left corner, they are c_5, d_5^v, d_5^h, d_5^d respectively.

Remark. Wavelet transform can extract the detail of data pretty well. However, if given a set of similar data, for example a set images of different cars, how do we get the major feature which describes all the car images. Here we will take the advantage of the Principal Component Analysis.

3 Principal Component Analysis

3.1 Statistics

Give a list of data of size n, $X = (X_1, X_2, \cdots, X_n)$, we have mean value $E[X] = (\sum X_i)/n$ to describe the average information, and the Standard Deviation $s = \sqrt{\frac{\sum(X_i - E[X])^2}{n-1}}$ illustrate how spread out the data is. Another common way to describe the spreading out of data is **variance**.

Definition 3.1. *Define the variance of data X to be*

$$Var(x) = E[(X - E[X])^2].$$

However, the variance only describe how X is spread out on the axis. Suppose another list of data Y is given, it will be use to see how the points (X_i, Y_i) is spread out on the plane.

Definition 3.2. *Define the covariance to be*

$$Cov(X, Y) = E[(X - E[X])(Y - E[Y])].$$

Remark. The covariance of X with itself is just the variance of X.

Consider a higher dimensional case, that is give more data Z, W and so on. We want to study how the data are related with each other. It will be convenient if all the covariance between every two list of data.

Definition 3.3. *Define the covariance matrix to be*

$$C^{n\times n} = (c_{i,j}, c_{i,j} = Cov(Dim_i, Dim_j)),$$

where $C^{n\times n}$ is a $n \times n$ matrix, and Dim_i is the ith dimension.

Remark. In three dimensional case, C can be write as

$$C = \begin{pmatrix} cov(X,X) & cov(X,Y) & cov(X,Z) \\ cov(Y,X) & cov(Y,Y) & cov(Y,Z) \\ cov(Z,Y) & cov(X,Y) & cov(Z,Z) \end{pmatrix}.$$

3.2 The Mechanism of PCA

Consider points in n-dimensional space. Suppose we have m points $X_i = (x_1, x_2, \cdots, x_n)^T, 1 \leq i \leq n$ satisfying that $E[X_i] = 0$. What PCA does is to rotate the points together on the space such that the points only spread out along axises. To see a mathematical description, define $X = (X_1, X_2, \cdots, X_m)$ so that X is a $n \times m$ matrix which contains all the information of the given data. By the definition above, the covariance matrix of these points on the space is simply written as

$$C_x = \frac{1}{m} X X^T.$$

Now we want to find a linear transform orthonormal matrix P that transform X to Y by

$$Y = PX,$$

such that the covariance matrix of Y, C_Y, is a diagonal matrix. The rows of P is then called the *principal components* of X.

3.3 Solving PCA Using Eigenvector Decomposition

Let $Y = PX$, where P is an orthonormal matrix. Notice that

$$\begin{aligned} C_Y &= \frac{1}{m} Y Y^T \\ &= \frac{1}{m} (PX)(PX)^T \\ &= \frac{1}{m} P X X^T P^t \\ &= P(\frac{1}{m} X X^T) P^T \\ &= P C_X P^T. \end{aligned}$$

For given matrix $A = XX^t$, A is symmetric, for

$$A^T = (XX^T)^T = XX^T = A.$$

Therefore, C_X is a symmetric matrix. There are some very nice properties about a symmetric matrix.

Definition 3.4. *(Orthogonal Diagonalizable) Let A be a $n \times n$ matrix. A is orthogonal diagonalizable if there is a orthogonal matrix B(i.e. $B^T B = I$) such that $S^{-1} A S$ is diagonal.*

Theorem 3.1. *(Spectral Theorem) Let A be a $n \times n$ matrix. A is orthogonal diagonalizable if and only if A is symmetric.*

See Serge Lang[4] for the proof.

Theorem 3.2. *Let A be a $n \times n$ matrix. A is diagonalized by its matrix of eigenvectors.*

Proof. let $\mathbf{e}_i, 1 \leq i \leq n$ be the independent eigenvectors of A with $\|\mathbf{e}_i\| = 1$. for any two eigenvector $\mathbf{e}_i$ and $\mathbf{e}_j$, $i \neq j$,

$$\begin{aligned}\lambda_i \langle \mathbf{e}_i, \mathbf{e}_j \rangle &= (\lambda_i \mathbf{e}_i)^T \mathbf{e}_j \\ &= (A\mathbf{e}_i)^T \mathbf{e}_j \\ &= \mathbf{e}_i^T (A^T \mathbf{e}_j) \\ &= \mathbf{e}_i^T (A\mathbf{e}_j) \\ &= \lambda_j \langle \mathbf{e}_i, \mathbf{e}_j \rangle\end{aligned}$$

Since $\lambda_i \neq \lambda_j$, we have $\langle \mathbf{e}_i, \mathbf{e}_j \rangle = 0$. Let $E = (\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_n)$. By orthogonality, $EE^T = I$. Therefore, $AE = E\Lambda$ implies

$$A = E\Lambda E^{-1} = E\Lambda E^T.$$

□

Now to find the transform P, the trick is let $P \equiv E^T$, where $C_X = E\Lambda E^T$. Then

$$\begin{aligned}C_Y &= PC_XP^T \\ &= PE\Lambda E^T P^T \\ &= (E^T E)\Lambda(E^T E) \\ &= \Lambda.\end{aligned}$$

3.4 Maximum Principal Components of Data

We have found the transform P that $Y = PX$ and C_Y is the eigenvalue matrix of C_X.Suppose we are only interested in the subspace in the n-dimensional space where the points are mostly spread out, then we need to choose the C_Y to be the $k \times k$ eigenvalue matrix where the $\lambda_1, \lambda_2, \cdots, \lambda_k$ is the top k maximum eigenvalues of C_X. Let $\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_k$ be the corresponding eigenvectors. P is therefore composed by $(\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_3)^T$. Thus the Y will be the coordinates of the projection of the points on those principal axes.[5]

4 Wavelet PCA based Feature Extraction Example

Now suppose we have a set of images, which represents the same object with some noise in the images. We will use wavelet transform and PCA to find a better way to get the description of the set of images.

To illustrate the method more clearly, we choose a set of 24 face images which is token from the same person.

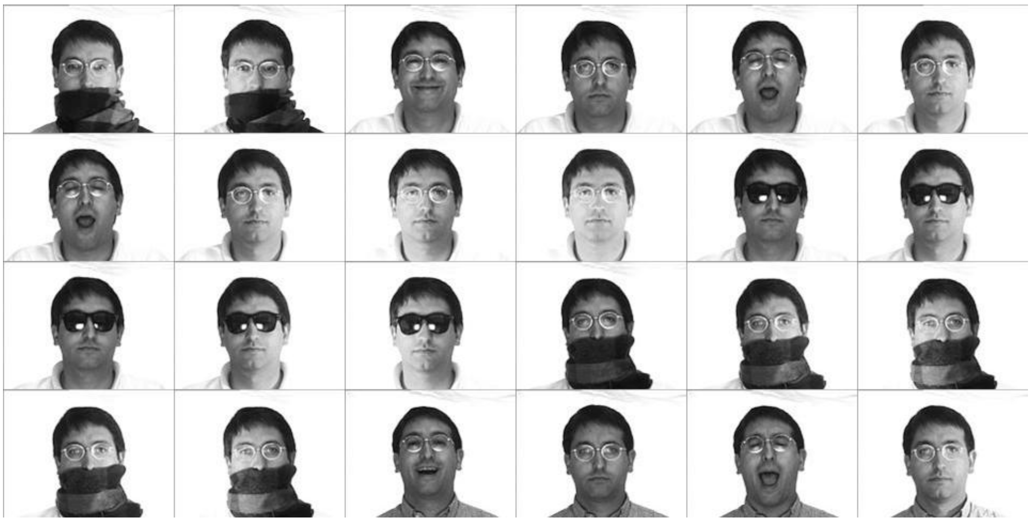

Figure 4: A set of black and white image of faces of the same person

Next apply DHT of scale 2 to each image of face, we obtain

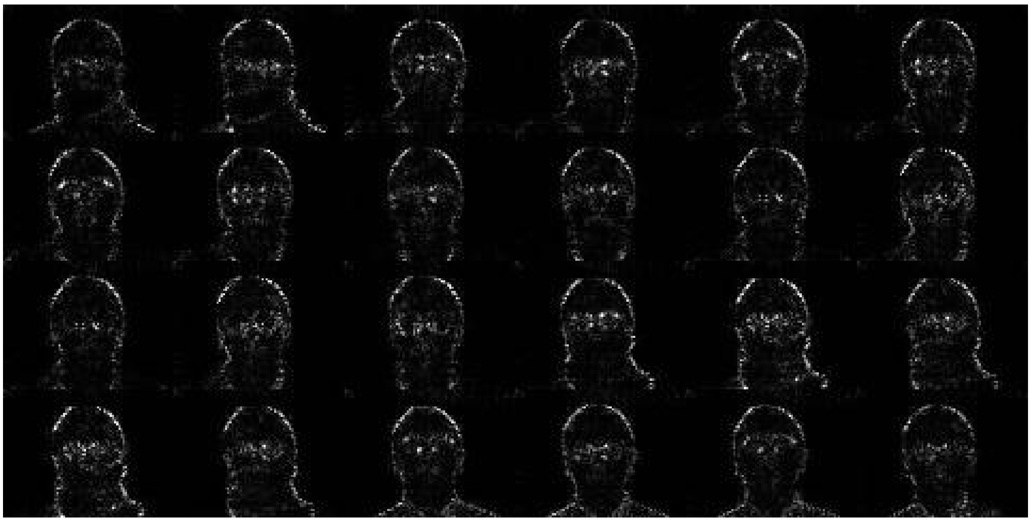

Figure 5: A set of detail of face images after DHT.

The average the these images are

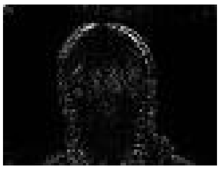

Figure 6: The average of the detail of the face images after DHT.

Apply PCA on this set of 24 images to find the a better feature which puts more emphasis on the common of the faces.

To use PCA on images. First, concatenate each image as an vector Z_i. In this case, each image is 60 by 80, so each vector Z_i has dimension 4800. Suppose $\bar{Z}$ is the average of Z_i, let $X_i = Z_i - \bar{Z}$. As described earlier, this will centralize the data in the space. So we have a matrix form

$$X = (X_1, X_2, \cdots, X_{24}).$$

which is better for computing. Again let

$$Y = PX.$$

Then we can solve for P and Y by the method introduced previously, where P is the transform matrix and each column Y_i of Y is the projection of image X_i.

Particularly, let P be consisted by the eigenvectors with the top 3 greatest eigenvalues so that we can eliminate some noise of the 24 image and get the principal components. Therefore, P has dimension 3×4800, and Y is a 3×24 matrix. So under the transform of$P$, Y_i has three values and gives a good description of image X_i. The amazing thing is when comparing different faces, projecting face image by P first will gives a better comparison result between images. Therefore, PCA is also a common way used in face detection and recognition.

Finally, to get an intuitive feeling of Y, we can transform back be reconstruct the images which are the results of enhancing the common feature and weakening the noises by the formula

$$Z_i^{'} = P^T Y_i + \bar{Z}.$$

and $Z_i^{'}$ can be shown by the images below.

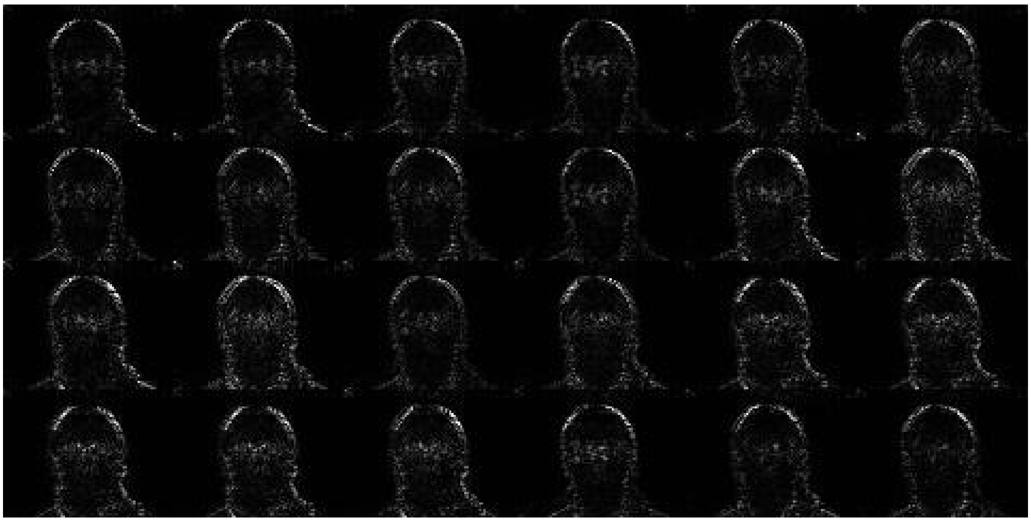

Figure 7: Reconstruct images according to the main feature

This example gives some illustration that wavelet transform and Principal Component Analysis can extract the common feature of a set of data, and gives a concise representation of them. However, this method still has some limitation and does not work well in some case, see [1] and [2]

References

[1] Center for Biological and Computational Learning, Artificial Intelligence Laboratory, MIT. *A Trainable System for Object Detection*, International Journal of Computer Vision 38(1), 15C33, 2

[2] Maya R. Gupta and Nathaniel P. Jacobson. *Wavelet Principal Component Analysis And Its Application To Hyperspectral Images*, Department of Electrical Engineering, University of Washington

[3] David F. Walnut. *An Introduction to Wavelet Analysis*, Department of Mathematical Science, George Mason University, $P_g.116 - 159$.

[4] Serge Lang. *Introduction to Linear Algebra*,Department of Mathematics, Yale University. $Pg.250 - 258$

[5] Jonathon Shlens. *A Tutorial on Principal Component Analysis*,Center for Neural Science, New York University.