Image Redundancy Reduction for Neural Network Classification using Discrete Cosine Transforms

Zhengjun Pan[1,2], **Alistair G. Rust**[1,2], and **Hamid Bolouri**[1]

[1]Science & Technology Research Centre
[2]Department of Computer Science
Faculty of Engineering and Information Sciences
University of Hertfordshire, Hatfield, Herts, AL10 9AB, UK
{Z.Pan, A.G.Rust, H.Bolouri}@herts.ac.uk

Abstract— High information redundancy and strong correlations in face images result in inefficiencies when such images are used directly in recognition tasks. In this paper, Discrete Cosine Transforms (DCTs) are used to reduce image information redundancy because only a subset of the transform coefficients are necessary to preserve the most important facial features, such as hair outline, eyes and mouth. We demonstrate experimentally that when DCT coefficients are fed into a backpropagation neural network for classification, high recognition rates can be achieved using only a small proportion (0.19%) of available transform components. This makes DCT-based face recognition more than two orders of magnitude faster than other approaches.

Keywords— Face recognition, neural networks, feature extraction, discrete cosine transform, data pre-processing.

1 Introduction

Performing face recognition, identification and classification tasks directly using raw images is an inefficient strategy due to high information redundancy in face images. To overcome this difficulty, a computational model is typically designed to transform pixel images into face features, and these features are then used for analysis and recognition[3]. One exemplar neural network approach was developed by Lawrence *et al.* in [9] who used the self-organising map (SOM) as a feature extractor. The self-organised features were exploited as the input to a convolutional neural network for recognition, whose architecture was similar to that of neocognitron[10]. A weakness of this approach is that training either SOMs/convolutional neural networks can be tremendously computationally expensive.

Most facial feature extraction approaches however rely only on localised features, which can be ill-posed and may be brittle subject to variations of illumination, scale and orientation. Due to the difficulty of selecting a representation that can robustly capture features, some researchers tend to avoid the feature extraction procedure by passing the pixel images directly to neural networks, using neural networks as information processing tools[10, 14]. In principle, neural networks can be used to map the pixel face images directly onto the target output values. However, in practice, such an approach will typically generate poor results due to high information redundancy and strong correlations present in face images. These problems can be accommodated by pre-processing the raw face images, reducing the dimensionality of the working space[1]. The choice of pre-processing algorithms can therefore be one of the most significant factors in determining the performance of the final recognition system.

A well-known and widely used statistical technique for dimensionality reduction is Principal Component Analysis (PCA), or the equivalent Karhunen-Loève or Hotelling transforms[8]. PCA is the optimal linear transform in an information packing sense[1], which can combine inputs in high dimensional space and generate a smaller set of features. Due to the fact that PCA is a linear technique, it may however be inappropriate to use it for modelling nonlinear deformations and correlations, such as bending[2]. To overcome such limitations of linear PCA, several variants have been proposed[2, 6, 7]. However, in general, PCA is data dependent and obtaining the principal components is a nontrivial task. Other limitations of PCA-based face recognition techniques are that large memory resources are required to store the components of images and an exhaustive search is needed to identify the closest match to an unknown face. Alternatively nonlinear dimensionality reduction can also be performed by multi-layer neural networks[1, 13, 14]. Again training the multi-layer neural networks can be computationally expensive[1].

0-7695-0619-4/00 $10.00 © 2000 IEEE

In this paper we present a new approach which addresses the two issues of feature extraction and information packing of face images using the Discrete Cosine Transform (DCT) for neural network processing. The DCT is investigated here since its basis functions are input independent and its information packing ability closely approximates the optimal PCA[5]. Furthermore computationally efficient algorithms exist to compute 2D DCTs[4]. The DCT hence provides a good compromise between information packing ability and computational complexity.

Another advantage of the DCT is that most DCT components from real world images are typically very small in magnitude because most of the salient information exists in the coefficients with low frequencies. Truncating, or removing these small coefficients from the representation thereby introduces only small errors in the reconstructed images. Hence, a limited number of DCT components are sufficient to preserve the most important facial features such as hair outline, eyes and mouth[11]. By presenting these components to a classifier, here we use a feed forward multi-layer neural network, high recognition rates can be achieved, whilst the training and recognition speeds of the system are dramatically faster than other comparable approaches.

2 From Pixel to Digital

2.1 Discrete Cosine Transform

The DCT of an $N \times M$ image $f(x,y)$ is defined by

$$C(u,v) = \frac{2}{\sqrt{MN}} \cdot \alpha(u)\alpha(v) \sum_{x=0}^{N-1} \sum_{y=0}^{M-1} f(x,y) \cdot \cos\left[\frac{(2x+1)u\pi}{2N}\right] \cos\left[\frac{(2y+1)v\pi}{2M}\right] \tag{1}$$

for $u = 0,1,2,\cdots,N-1, v = 0,1,2,\cdots,M-1$, and the inverse transform is defined by

$$f(x,y) = \frac{2}{\sqrt{MN}} \cdot \sum_{u=0}^{N-1} \sum_{v=0}^{M-1} \alpha(u)\alpha(v)C(u,v) \cdot \cos\left[\frac{(2x+1)u\pi}{2N}\right] \cos\left[\frac{(2y+1)v\pi}{2M}\right] \tag{2}$$

for $x = 0,1,2,\cdots,N-1, y = 0,1,2,\cdots,M-1$; where $\alpha(w) = \frac{1}{\sqrt{2}}$ for $w = 0$ and otherwise $\alpha(w) = 1$.

The role of the DCT used in this paper is to reduce the dimensionality of the working space. The dimension of the original space is determined by the maximum number of DCT coefficients (see (1)). Hence, in order to reduce the dimensionality, a selection of coefficients should be omitted.

To decide which components should be retained for the classification task in hand, let us denote the kernels of the 1-D DCT in (1) by

$$h(u,x) = \frac{2}{\sqrt{N}} \cdot \alpha(u) \cdot \cos\left[\frac{(2x+1)u\pi}{2N}\right], \quad g(y,v) = \frac{2}{\sqrt{M}} \cdot \alpha(v) \cdot \cos\left[\frac{(2y+1)v\pi}{2M}\right] \tag{3}$$

for $u = 0,1,\cdots,N-1, x = 0,1,\cdots,N-1$ and $v = 0,1,\cdots,M-1, y = 0,1,\cdots,M-1$.

Note the kernels $h(u,x)$ and $g(y,v)$ in (3) depend only on the locations x,y,u,v and not on the image (i.e., the values of $f(x,y)$). Therefore, they can be viewed as a set of basis functions of the DCT. Based on this interpretation, the DCT $\mathbf{C} = \big(C(u,v)\big)_{N\times M}$ of image $\mathbf{F} = \big(f(x,y)\big)_{N\times M}$ can be expressed in matrix form as

$$\mathbf{C} = \mathbf{H} \cdot \mathbf{F} \cdot \mathbf{G} \tag{4}$$

and its inverse transform can be expressed as

$$\mathbf{F} = \mathbf{H}^T \cdot \mathbf{C} \cdot \mathbf{G}^T \tag{5}$$

where $\mathbf{H} = \big(h(u,x)\big)_{N\times N}$ is the 1-D transform along the column of image $\mathbf{F}$ and $\mathbf{G} = \big(g(y,v)\big)_{N\times N}$ is the row transform; $\mathbf{H}^T$ and $\mathbf{G}^T$ are the transposes of unitary matrices $\mathbf{H}$ and $\mathbf{G}$, respectively.

If we now define a masking function

$$m(u,v) = \begin{cases} 0, & \text{if } C(u,v) \text{ is omitted;} \\ 1, & \text{otherwise.} \end{cases} \tag{6}$$

for $u = 0,1,\cdots,N-1, v = 0,1,\cdots,M-1$. $m(u,v)$ is employed to eliminate DCT components, i.e., coefficients not selected to reconstruct the image. The reconstructed image $\hat{\mathbf{F}}$ can be obtained from

$$\hat{\mathbf{F}} = \mathbf{H}^T \cdot \hat{\mathbf{C}} \cdot \mathbf{G}^T \tag{7}$$

where $\hat{\mathbf{C}} = \big(C(u,v) \cdot m(u,v)\big)_{N\times M}$.

2.2 Feature Selection

Now consider a whole set of n images $\{\mathbf{F}_1, \mathbf{F}_2, \cdots, \mathbf{F}_n\}$. From all of their DCT features (i.e., DCT components), we wish to select the most informative components for our task. To achieve this, we try to minimize the reconstruction error for the selection scheme $m(u,v)$. Using the previous notations, the mean-square reconstruction error between the images $\{\mathbf{F}_1, \mathbf{F}_2, \cdots, \mathbf{F}_n\}$ and their approximations $\{\hat{\mathbf{F}}_1, \hat{\mathbf{F}}_2, \cdots, \hat{\mathbf{F}}_n\}$ can then be estimated as

$$\begin{aligned}\bar{E}_{mse} &= \frac{1}{n}\sum_{i=0}^{n}\Big\{\frac{1}{MN}\sum_{x=0}^{N-1}\sum_{y=0}^{M-1}\left[f_i(x,y) - \hat{f}_i(x,y)\right]^2\Big\} \\ &= \frac{1}{n}\sum_{i=0}^{n}\Big\{\frac{1}{MN}\sum_{u=0}^{N-1}\sum_{v=0}^{M-1}\left[C_i(u,v) - \hat{C}_i(u,v)\right]^2\Big\} \\ &= \frac{1}{MN}\sum_{u=0}^{N-1}\sum_{v=0}^{M-1}\Big\{\frac{1}{n}\sum_{i=0}^{n}C_i^2(u,v)\Big\}\cdot[1 - m(u,v)] \\ &= \frac{1}{MN}\sum_{u=0}^{N-1}\sum_{v=0}^{M-1}E\big\{C^2(u,v)\big\}\cdot[1 - m(u,v)]\end{aligned} \tag{8}$$

where $E\big\{C^2(u,v)\big\}$ is the expectation of DCT components at location (u,v). The first simplification is due to the orthonormal nature of the matrices $\mathbf{H}^T$ and $\mathbf{G}^T$.

If we denote $\bar{\mathbf{F}}$ as the mean of the image set $\{\mathbf{F}_1, \mathbf{F}_2, \cdots, \mathbf{F}_n\}$ and replace $\mathbf{F}_i$ with $\mathbf{F}_i - \bar{\mathbf{F}}$ and $\hat{\mathbf{F}}_i$ with $\hat{\mathbf{F}}_i - \bar{\mathbf{F}}$ in equation (8), then the mean-square reconstruction error between the images $\{\mathbf{F}_1 - \bar{\mathbf{F}}, \mathbf{F}_2 - \bar{\mathbf{F}}, \cdots, \mathbf{F}_n\bar{\mathbf{F}}\}$ and their approximations $\{\hat{\mathbf{F}}_1 - \bar{\mathbf{F}}, \hat{\mathbf{F}}_2 - \bar{\mathbf{F}}, \cdots, \hat{\mathbf{F}}_n - \bar{\mathbf{F}}\}$ can then be decided as

$$\begin{aligned}E_{mse} &= \frac{1}{n}\sum_{i=0}^{n}\Big\{\frac{1}{MN}\sum_{x=0}^{N-1}\sum_{y=0}^{M-1}\left[(f_i(x,y) - \bar{f}(x,y)) - (\hat{f}_i(x,y) - \bar{f}(x,y))\right]^2\Big\} \\ &= \frac{1}{MN}\sum_{u=0}^{N-1}\sum_{v=0}^{M-1}\sigma^2_{C(u,v)}\cdot[1 - m(u,v)]\end{aligned} \tag{9}$$

where $\sigma^2_{C(u,v)}$ is the variance of DCT components at location (u,v). This equation holds because the pixels of the images $\mathbf{F}_i - \bar{\mathbf{F}}$ can be regarded as being generated by a random process with zero mean and known variance. We know from equation (9) that the total mean-square reconstruction error is thus equal to the average of the variances of the discarded transform components (i.e., the components for which $m(u,v) = 0$). In this paper, DCT features are selected to minimize the mean-square reconstruction error in equation (9) (see also [11] in which a zonal selection scheme was used). Furthermore, DCT features (or components) addressed thereafter are the DCT coefficients of the difference $\mathbf{F} - \bar{\mathbf{F}}$ between an image $\mathbf{F}$ and the average image over the training set $\mathbf{F} - \bar{\mathbf{F}}$.

2.3 Normalisation of Features

The feature selection scheme (i.e., the number and locations of transform components) is determined based on analysis known images (i.e., training images) to minimize the mean-square error E_{mse}. The chosen features are then fixed within the recognition system and applied to unknown images. Selected components are arranged in one dimensional format based on the order of magnitude of their variances. Since DCT components in different locations usually have different orders of magnitude, we need to estimate the upper bound and the lower bound of DCT components for all images to convert the components into $[-1,1]$ (bipolar activation functions are used in our classifier[11]).

Suppose $x_1, x_2, \cdots, x_n$ are the components retained from the feature selection procedure and $\{(x_1^{(j)}, x_2^{(j)}, \cdots, x_n^{(j)}); j = 1, 2, \cdots, p\}$ are the corresponding components retained from the training images, where n is the number of DCT components retained and p is the number of training images. Then the upper bounds (b_i) and lower bounds (a_i) can be estimated by

$$b_i = \beta \cdot \max\{1, x_i^{(1)}, \cdots, x_i^{(p)}\},\ i = 1, 2, \cdots, n; \tag{10}$$

and

$$a_i = \beta \cdot \min\{-1, x_i^{(1)}, \cdots, x_i^{(p)}\}, \; i = 1, 2, \cdots, n. \tag{11}$$

where $\beta > 1$ is a factor to extend the bounds. Then the input vectors $\{(z_1^{(j)}, z_2^{(j)}, \cdots, z_n^{(j)}); j = 1, 2, \cdots, p\}$ to the neural network can be determined by

$$z_i^{(j)} = 2 \cdot \frac{x_i^{(j)} - a_i}{b_i - a_i} - 1, \quad i = 1, 2, \cdots, n. \tag{12}$$

For an unknown image, the scaling factors obtained from the training set are applied to the selected DCT components to obtain the input vector to a classifier.

3 Simulations

3.1 System Description

The main idea of our approach is to apply DCTs to reduce information redundancy in images and to use the packed information for classification. For a face image, the system first computes its DCT components (DCT coefficients of the difference image), then selects a fixed number of components before presenting them as inputs to a classifier. The classifier used in our system is a multi-layer perceptron (MLP) with only one hidden layer, where the quick backpropagation algorithm (Quickprop) is used as the training algorithm. A block diagram of our DCT-based system for face recognition is shown in Figure 1.

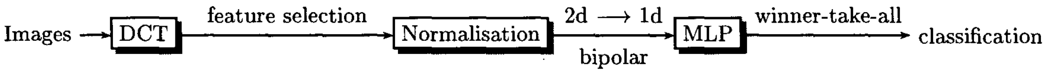

Figure 1: A functional block diagram of our proposed DCT-based face recognition system.

3.2 Experimental Setup

In our experiments, the ORL database (available from `www.cam-orl.co.uk/facedatabase.html`) is used as a benchmark database. The database consists of 400 different face images, 10 for each of 40 distinct subjects. The size of each image is 92 × 112 pixels, with 256 grey levels per pixel.

The weights and biases of the MLP used as the classifier are initialised to random values in the range $[-0.5, 0.5]$. Three learning parameters, $\alpha_{\max}, \epsilon_0$, and *decay*, used in Quickprop are set to 0.02, 0.008, 0.0001, respectively. The maximum number of training epochs is 1000. The multiplication factor β in (10) and (11) is set to 1.1. No attempt was made to optimise these parameters. To reduce the influence of the presentation order of training samples, the training samples were randomly shuffled after every training loop. For the ORL database, the number of outputs of the MLP was always 40 and a winner-take-all strategy was used for classification.

To allow comparisons, the same training and test set sizes are used as in [9, 12], i.e., the first 5 images for each subject are the training images and the remaining 5 images are used for testing. Hence there are 200 training images and 200 test images in total and no overlap exists between the training and test images. Due to the small size of the available data, a validation set was not used and the *best-so-far* recognition rate on test images is reported as the testing recognition rate.

In each of the following statistical results, 30 separate runs were carried out with randomly initialised weights and biases for each MLP. The T-tests are based on the 0.05 (95%) level of significance, which means that the T-test statistic has to exceed 1.645 for experimental results to be classified as statistically different.

Figure 2 shows the proportion of total sample variance associated with sorted DCT components on the training images of ORL database. It is demonstrated that most DCT components are insignificant and with only 20 (0.194% of 10304=92 × 112 available) DCT components we can keep 55% of the total variance.

3.3 Experimental Results

Table 1 shows the recognition performances for different numbers of DCT components in conjunction with different numbers of hidden neurons in the MLP. Table 1 also records the T-test result on the 0.05 level of significance compared to the best case (with 25 DCT components retained and 60 hidden neurons used in the MLP). It is demonstrated that the recognition rate decreases as more DCT components are retained. This is because by using more DCT components more person specific information is introduced which reduces

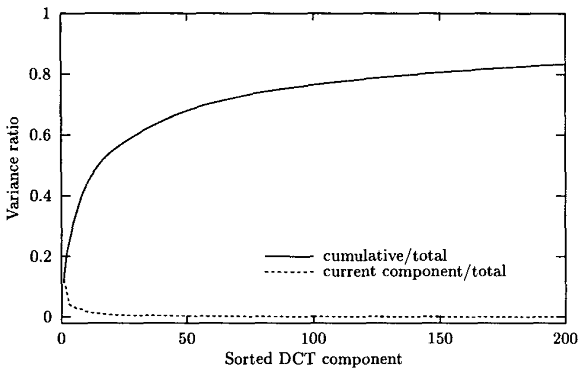

Figure 2: The proportion of total sample variance exploited by DCT components.

#	k	mean (%)	σ	max (%)	min (%)	T-test statistic	T
20	60	92.47	0.0121	94.5	89.5	5.343	✓
25	60	94.15	0.0123	97.0	92.5	—	—
25	80	93.67	0.0107	96.0	91.5	1.624	×
30	60	93.85	0.0097	96.0	92.5	1.051	×
30	80	94.00	0.0107	96.0	91.0	0.505	×
35	80	92.23	0.0123	94.5	90.0	6.035	✓
40	80	91.48	0.0138	94.0	89.0	7.901	✓
50	80	90.63	0.0119	92.5	88.5	11.261	✓
70	80	88.38	0.0166	91.5	84.5	15.261	✓

Table 1: Recognition performance on test images versus number of DCT coefficients retained (#) and number of hidden neurons in the MLP (k). The last column (T) shows the T-test result on the 0.05 level of significance. The best mean recorded performance is indicated in the shaded row.

the generalisation capabilities of the classifier. The most important facial features for remembering faces are hair, face outline, eyes and mouth which can be perceived by components at lower spatial frequencies[11].

The best average recognition rate is 94.15% obtained by retaining 25 DCT components and using an MLP with 60 hidden neurons. By comparison, for our earlier zonal selection scheme [11], the best mean recognition rate was 92.87% (where 35 DCT coefficients are retained and 75 hidden nodes are used in the MLP). The reason is that the zonal selection strategy can miss a few features with large variance which unfortunately are not in the zonal area described in [11]. A complete training run typically takes about 1 minute on a Pentium II PC with a 450MHz CPU. The T-tests test the hypothesis that the recognition rate is statistically different from the configuration with the best mean recognition rate.

3.4 Comparison of Different Recognition Approaches

The ORL database has been used to test several face recognition approaches[9, 11, 12]. The recognition rates of the best models and the training/classification times (if available) are shown in Table 2.

As shown in Table 2, the recognition rates of our DCT-based system are comparable to the best reported results (the Convolutional NN and the P2D-HMM) and better than that achieved by PCA. The relative recognition speeds given in Table 2 are extrapolated from benchmark evaluations using the MATLAB benchmark utility and the published SPEC CPUfp92 data (available from www.spec.org). The relative speed calculation was discussed in [11]. Note that the classification time for our DCT-based method is around 600 times faster than the convolutional neural network approach. The classification speed of the convolutional neural network (CNN) approach is itself about 200 times faster than the P2D-HMM approach (Lawrence et al[9] report CNN to be 500 times faster than P2D-HMM, but their comparison ignores processor speed differences). Furthermore, for the above comparison, input images to the CNN and the MLP were a quarter of full resolution. For $N \times N$ images, the computational cost of these approaches is proportional to $O(N^2)$ while the computational complexity of fast DCT is only $O(N \log N)$ when N is a power of 2.

approach	recognition rate			training time	recognition time	relative speed*
	best	mean	σ			
HMM[12]	87%	—	—	—	—	
eigenfaces(PCA)[12]	90%	—	—	—	—	
P2D-HMM[12]	95%	—	—	—	≈4 minutes†	1/192
convolutional NN[9]	98.5%‡	96.2%§	0.004§	≈4 hours¶	< 0.5 seconds¶	1
MLP[11]	84.0%	77.2%	0.0353	10 minutes‖	0.0014 seconds‖	89
DCT+MLP	97.0%	94.2%	0.0123	≈ 1 minute‖	0.0002 seconds‖	625

*relative recognition speed, see text for details.

†on a Sun Sparc II workstation, p.92 of [12],'—' means data not available.

‡not included in the calculation of the mean and standard deviation (personal communication).

§average of 3 simulations, the value of the standard deviation is from a personal communication.

¶on an SGI Indy MIPS R4400 100MHz system.

‖on a 450MHz IBM compatible PC with 128M RAM.

Table 2: Performance comparison of different approaches to recognition applied to the ORL database.

4 Conclusions

Reducing the dimensionality of face images simplifies neural network based recognition systems. In this paper, dimensionality reduction is achieved by applying a DCT to the face images and truncating unimportant features. The importance of DCT features can be determined by their variance over the training images. Selected transform features, rather than the raw pixel data, are used for neural network classification. Experiments demonstrated that for the ORL database, using less than 0.2% ***correct?*** of all the available DCT features, our DCT-based approach produces a recognition rate comparable to the best results reported to date while recognition speed is more than 2 orders of magnitude faster. Since DCTs (unlike the popular dimensionality reduction technique Principle Component Analysis) are data independent and obtaining DCT components is much less computationally expensive than PCA, our DCT-based approach to face recognition is much faster than other comparable methods.

References

1. C. M. Bishop, *Neural Networks for Pattern Recognition.* Oxford: Oxford University Press, 1995.
2. B. Chalmond and S. Girard, "Nonlinear modeling of scattered multivariate data and its application to shape change," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 5, pp. 422–432, 1999.
3. R. Chellappa, C. L. Wilson, and S. Sirohey, "Human and machine recognition of faces: A survey," *Proceedings of the IEEE*, vol. 83, no. 5, pp. 705–740, 1995.
4. C. Christopoulos, J. Bormans, A. Skodras, and J. Cornelis, "Efficient computation of the two-dimensional fast cosine transform," in *SPIE Hybrid Image and Signal Processing IV*, pp. 229–237, 1994.
5. R. Gonzalez and R. Woods, *Digital Image Processing.* Reading, MA: Addison-Wesley, 1992.
6. A. Hyvärinen, "Survey on independent component analysis," *Neural Computing Surveys*, 2, pp. 94–128, 1999.
7. J. Karhunen and J. Joutsensalo, "Generalization of principal component analysis, optimization problems and neural networks," *Neural Networks*, vol. 8, no. 4, pp. 549–562, 1995.
8. M. Kirby and L. Sirovich, "Application of the Karhunen-Loeve procedure for the characterization of human faces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 1, pp. 103–108, 1990.
9. S. Lawrence, C. Lee Giles, A. Tsoi, and A. Back, "Face recognition: A convolutional neural network approach," *IEEE Transactions on Neural Networks*, vol. 8, no. 1, pp. 98–113, 1997.
10. C. Nebauer, "Evaluation of convolutional neural networks for visual recognition," *IEEE Transactions on Neural Networks*, vol. 9, no. 4, pp. 685–696, 1998.
11. Z. Pan, R. Adams, and H. Bolouri, "Dimensionality reduction of face images using discrete cosine transforms for recognition." submitted to *IEEE Conference on Computer Vision and Pattern Recognition*, 2000.
12. F. Samaria, *Face Recognition using Hidden Markov Models.* PhD thesis, Cambridge University, 1994.
13. E. Saund, "Dimensionality-reduction using connectionist networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, no. 3, pp. 304–314, 1989.
14. D. Valentin, H. Abdi, A. O'Toole, and G. Cottrell, "Connectionist models of face processing: A survey," *Pattern Recognition*, vol. 27, pp. 1209–1230, 1994.