A New Image Filtering Technique Combining a Wavelet Transform with a Linear Neural Network : Application to Face Recognition

Fan Yang (fanyang@u-bourgogne.fr)
Michel Paindavoine (paindav@u-bourgogne.fr) *SPIE member
Laboratoire d'Electronique, d'Informatique et d'Image
Aile de l'ingénieur - Mirande
Université de Bourgogne, BP 47870
21078 DIJON France

Hervé Abdi (http://www.utdallas.edu/ herve)
The University of Texas at Dallas
Program in Cognition
MS GR:4.1
Richardson TX 75083-0688 U.S.A.

Abstract

(2000, Optical Engineering, **38**, 2894–2899)

In order to improve the performance of a linear auto-associator (which is a neural network model), we explore the use of several pre-processing techniques. The gist of our approach is to represent each pattern by one or several pre-processed (i.e. filtered) versions of the original pattern (plus the original pattern). First, we compare the performance of several pre-processing techniques (a plain vanilla version of the auto-associator as a control, a Sobel operator, a Canny-Deriche operator, and a multiscale Canny-Deriche operator) and a Wiener filter on a pattern completion task using a noise degraded version of faces stored. We found that the multiscale Canny-Deriche operator gives the best performance of all models. Second, we compare the performance of the multiscale Canny-Deriche operator with the control condition on a pattern completion task of noise degraded versions (with several levels of noise) of learned faces and new faces of the same or another race than the learned faces. In all cases, the multiscale Canny-Deriche operator performs significantly better than the control.

keywords : auto-associator, multiscale Canny-Deriche operator, pattern recognition, wavelet transform, Wiener filter, Principal Component Analysis.

1 Introduction

Linear auto-associative memories are one of the most simple and well studied neural-network model[1] [2]. They are widely used as models for cognitive tasks as well as pattern recognition, or digital signal processing, in part because they are formally equivalent to well-known techniques such as the Karhunen-Loève transform or principal component analysis [3]. Even though linear auto-associators are known to be quite robust when noise is added to the patterns to be recognized, their performance is rather bad when a *lot* of noise is added to the stimulus. One of the ways of improving performance could be to use some pre- and post- processing of the patterns to be recognized. In this paper, we evaluate the performance of a pre-processing techniques using wavelet transform applied to face images. This paper is organized as follows. First, we describe the linear auto-associator model applied to face images. In the second part, we compare the wavelet approach with other pre-processing or filtering techniques. In the third part, we look at the performance of the wavelet approach under various conditions of noise degradation.

2 Description of linear associators

The class of models presented in this section are known as linear associators[3] [4] [5]. They come in two forms: hetero- and auto-associators. The hetero-associator can be used to learn arbitrary associations between input and output patterns. The auto-associator is a special case of the hetero-associator in which the association between an input pattern and itself is learned. In this paper, we will consider only the linear auto-associator.

The advantage of linear associators in comparison with non-linear models is that they provide for the integration of a very large number of cells in the network. Their implementation is quite easy, because they can be analyzed in terms of the singular value decomposition of a matrix[3][5]. Besides, linear models constitute a first processing stage for numerous applications using more sophisticated approaches (see [3],[6] for reviews).

When these models are applied to images, the first step is to transform each digitized image into a (image) vector by concatening the columns of the matrix of the pixel values of the image. Images are “stored” into a connection weight matrix, which models neural synaptic connections between neural cells associated to the image pixels (see Fig. 1).

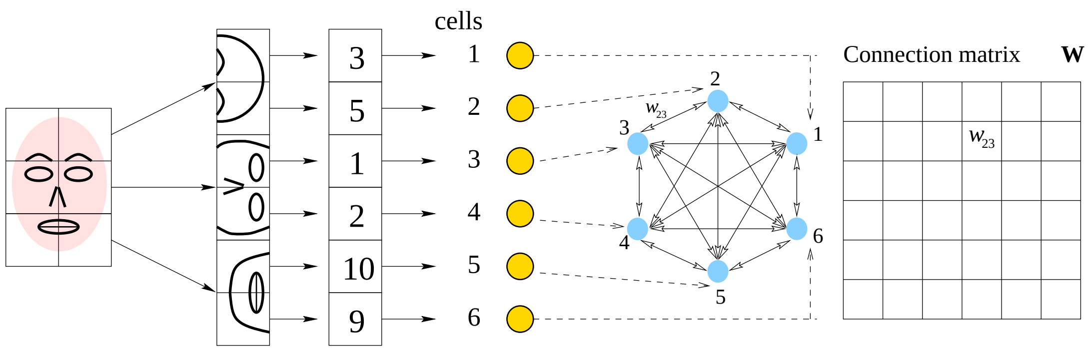

Figure 1: The linear auto-associator is applied to images : transform each digitized image into a vector by concatening the columns of the matrix of the pixel values of the image.

In our description, we follow closely the formulation detailed in[5]. The patterns to be learned are represented by $L \times 1$ vectors $\mathbf{a}_k$ where k is the stimulus number. The components of $\mathbf{a}_k$ specify the values of the pattern to be applied to the L cells of the input layer for the k-th stimulus. The complete set of K stimuli is represented with a $L \times K$ matrix noted $\mathbf{A}$ (i.e., $\mathbf{a}_k$ is the k-th column of $\mathbf{A}$). The $L \times L$ synaptic weight connection matrix between the L input cells is denoted $\mathbf{W}$. Learning occurs by modifying the values of the connection weight between cells as explained later in this section. The response (or estimation) of the model to a pattern $\mathbf{x}$ (which may or may not have been learned) is obtained as

$$\widehat{\mathbf{x}} = \mathbf{W}\mathbf{x} \tag{1}$$

Because auto-associators are generally interpreted as content addressable memories, their performance is evaluated by comparing the output of the system with a test pattern which can be a copy or a degraded version of one of the patterns previously learned by the system. This is achieved by computing similarity measures (most of a time a cosine) between input and output. The *coefficient of correlation* between $\widehat{\mathbf{x}}$ and $\mathbf{x}$ is an index of the quality of the estimation and so on the larger the similarity between input and output, the better the performance.

In order to achieve a high level of performance, several iterative learning rules have been proposed. The most popular one is clearly the Widrow-Hoff learning rule. This is an iterative procedure which corrects the connection matrix $\mathbf{W}$ using the difference between the target response and the actual response of the network. In matrix notation, the Widrow-Hoff rule is written as:

$$\mathbf{W}_{[t+1]} \quad = \quad \mathbf{W}_{[t]} + \eta(\mathbf{a}_k - \mathbf{W}_{[t]}\mathbf{a}_k)\mathbf{a}_k^T \tag{2}$$

with $\mathbf{W}_{[t]}$ being the weight matrix at step *t*, η being a small positive constant, and the index k being chosen randomly.

The linear associator has been often analyzed in terms of the eigenvalue decomposition or singular value decomposition of a matrix. In terms of the singular value decomposition, the rectangular stimulus matrix $\mathbf{A}$ is decomposed as:

$$\mathbf{A} = \mathbf{P}\boldsymbol{\Delta}\mathbf{Q}^T \tag{3}$$

with $\boldsymbol{\Delta}$, diagonal matrix of singular values ($\boldsymbol{\Delta} = \boldsymbol{\Lambda}^{1/2}$); $\boldsymbol{\Lambda}$, diagonal matrix of non zero eigenvalues of $\mathbf{A}\mathbf{A}^T$ and $\mathbf{A}^T\mathbf{A}$; $\mathbf{P}$, matrix of eigenvectors $\mathbf{A}\mathbf{A}^T$; and $\mathbf{Q}$, matrix of eigenvectors $\mathbf{A}^T\mathbf{A}$. The Widrow-Hoff learning rule can be analyzed in terms of the singular value decomposition of matrix $\mathbf{A}$ [3]. Specifically, $\mathbf{W}_{[t]}$ is expressed as[5]:

$$\mathbf{W}_{[t]} = \mathbf{P}\{\mathbf{I} - (\mathbf{I} - \eta\boldsymbol{\Lambda})^t\}\mathbf{P}^T \tag{4}$$

When η is smaller than $2\lambda_{\text{max}}^{-1}$ (λ_{max} being the largest eigenvalue of $\boldsymbol{\Lambda}$), the Widrow-Hoff learning rule will converge to:

$$\mathbf{W}_{[\infty]} = \mathbf{P}\mathbf{P}^T \tag{5}$$

which is the value we used in this paper. The matrix $\mathbf{P}$ is a $L \times N$ matrix with N being the number of non zero eigenvalues. Typically, L is significantly smaller than N (*i.e.*, $N \ll L$). As a consequence, using $\mathbf{P}$ directly instead of $\mathbf{W}$ will lead to an important gain in processing speed as well as storage. For example, when dealing with a face recognition application, the matrix $\mathbf{W}$ was a 33975×33975 matrix whereas the eigenvector matrix $\mathbf{P}$ was only a 33975×400 matrix. In terms of the eigenvector matrix $\mathbf{P}$, the response of the model $\widehat{\mathbf{x}}$ to a pattern $\mathbf{x}$ (equation 1) is obtained as :

$$\widehat{\mathbf{x}} = \mathbf{P}\mathbf{P}^T\mathbf{x} \tag{6}$$

3 A pre-processing using multiscale edges

The goal of learning is to find values for the connections between cells such that the response of the model approximates the input as well as possible. To assess the performance of the model, degraded versions of the previously learned stimuli are presented to the model as a test. If learning has been successful, then the response pattern will be more similar to the original pattern than the degraded stimulus was (see [1] for an illustration). In other words, auto-associators can act as pattern completion devices.

In this paper we explore different approaches for improving the performance of a linear auto-associator storing face images. The general strategy is to store, in addition to the original images, several filtered versions of the images (see figure 3). We refer to this technique as *pre-processing*. Then, the model is evaluated by its reconstruction performance when presented with *probes* which are versions of the original faces degraded by the addition of Gaussian random noise.

As we are interested in image patterns, we choose filtering techniques meaningful in the this context. Because it is generally agreed that edges are essential for recognition[7], we decided to increase their importance in the image. Quite a number of algorithms have been proposed in the literature to perform edge extraction. We decided to implement three algorithms, namely:

1. the *Sobel* operator (a differential operator) as it is considered as a standard procedure well suited for noiseless images.

2. the *Optimized Canny-Deriche* operator because it is known to be optimal for edge extraction in noisy images [8][9].

3. the *Multiscale optimized Canny-Deriche* edge detector[9] which is equivalent to finding local maxima of a wavelet transform as suggested in [10]. The optimized Canny-Deriche filter is a separable filter when applied to 2D images. Its impulse response for a 1D signal (because of its separability, the filter can be seen as two 1D filters) is given by:

$$f(x) = ksxe^{msx} + e^{msx} - e^{sx} \tag{7}$$

with $k = 0.564$, $m = 0.215$, and where $s = 2^j$ is the scale factor (with, in our case, $j \in \{0, 1, 2, 3\}$), and x being the pixel position. The figure 2 displays the impulse response of optimized Canny-Deriche filter for different scales.

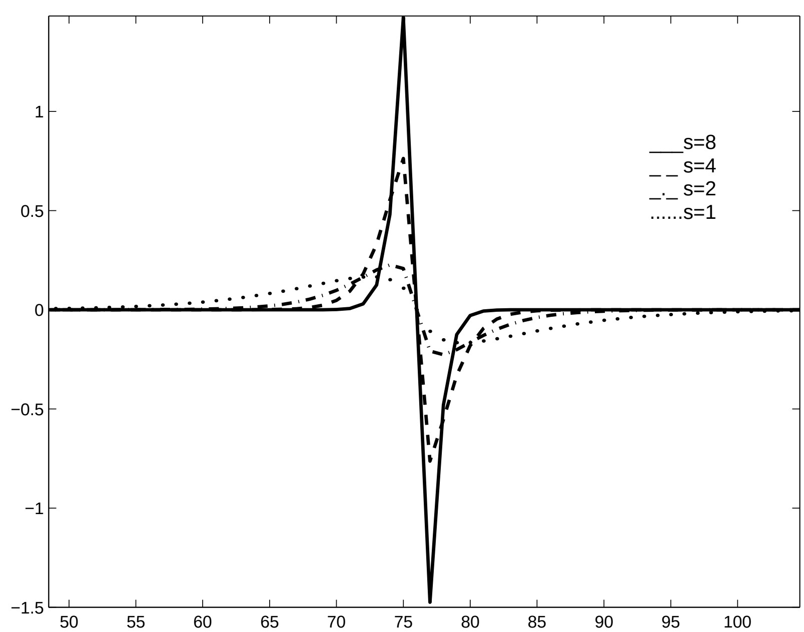

Figure 2: Impulse response of optimized Canny-Deriche filter for different scales.

This method is implemented as a wavelet transform using a convolution between the image and the edge detection filter for different scales ($s = 2^j$). As a result, this filter detects edges occurring at different scale resolutions in the image [10].

In order to compare these different techniques, we implemented these 3 models along with 2 control conditions corresponding to a standard associator, and to denoising using a Wiener filter (because it is a standard algorithm applied one image at a time). All the models were tested using the same procedure. The patterns stored were a set of 80 face images of size 225×151 (with 16 gray levels per pixel). The performance of each model was assessed as follows. Each face was then degraded by adding to each pixel a random number (chosen such that the noise values belong to the interval $[0\ 45]$). The degraded face was then recalled by the model (or filtered in the case of the Wiener filter model). The correlation between the original face and the model reconstructed face reflects the performance of the model for this face: The higher the correlation, the better the performance. Specifically, the models were a Wiener filter applied directly to the noise degraded stimulus and 4 auto-associators (see figure 3) :

1. a standard auto-associator storing the original 80 faces images;
2. an auto-associator storing the original 80 face images plus, for each face, a *Sobel* filtered image of the face (hence a total of 160 face images);
3. an auto-associator storing the original 80 face images plus, for each face, a *Canny-Deriche* filtered image of the face (hence a total of 160 face images);
4. an auto-associator storing the original 80 face images plus, for each face, four *wavelet transformed* (by a multiscale Canny-Deriche filter) face images (one face image per scale resolution, hence a total of 400 images).

For the last three models, the complete set of patterns to be learned (matrix $\mathbf{A}$) is composed by original images and filtered images. The eigenvector matrix $\mathbf{P}$ and the synaptic connection matrix $\mathbf{W}$ have be obtained using equations (3) and (5).

Fig. 4 displays an example of the responses of the models to the test face. The top panels present: A face previously learned by the system and a stimulus with additive random noise added used as a probe to evaluate the performance of the models. The bottom panels show the estimation of the original face by: a standard auto-associator, a Wiener filter, auto-associator plus Sobel preprocessing, auto-associator plus Canny-Deriche filter, and auto-associator plus wavelet transform.

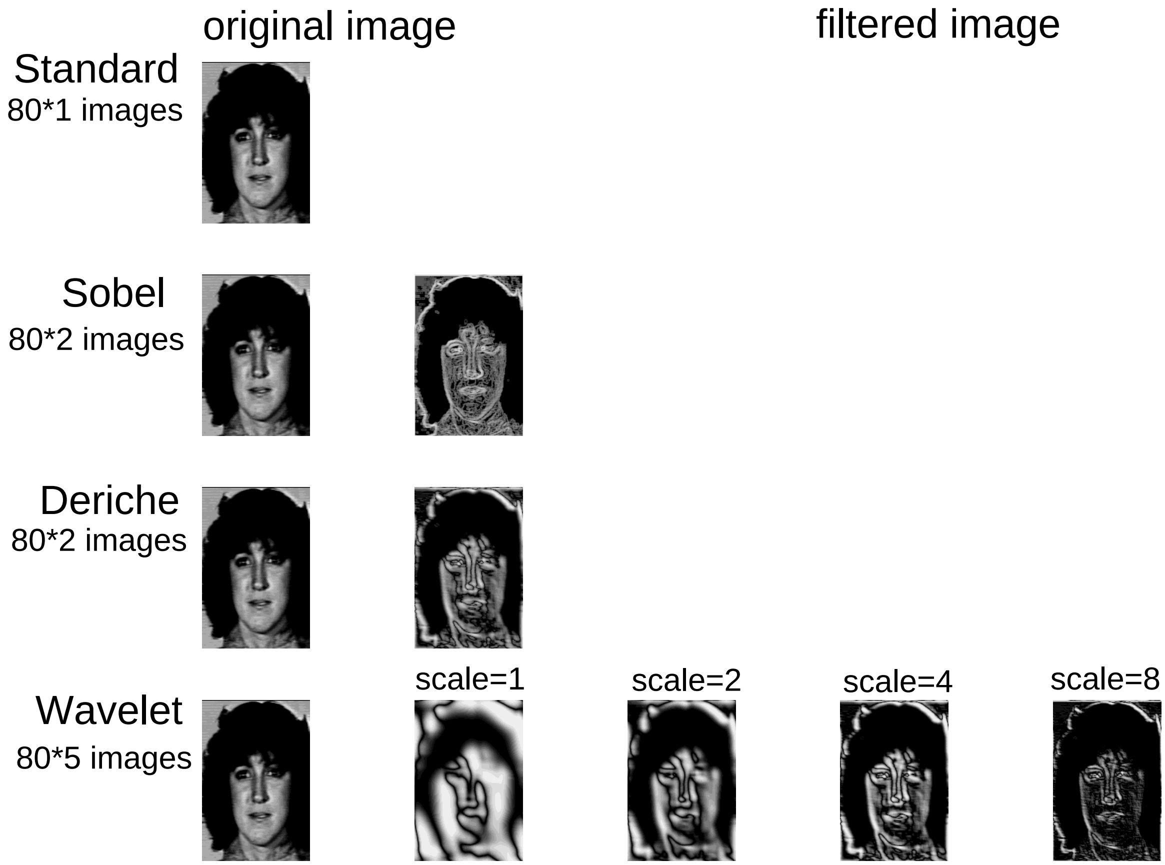

Figure 3: Patterns to be learned by 4 auto-associators : Filtered images have be obtained respectively with the Sorel operator, optimized Canny-Deriche operator and the Wavelet Transform.

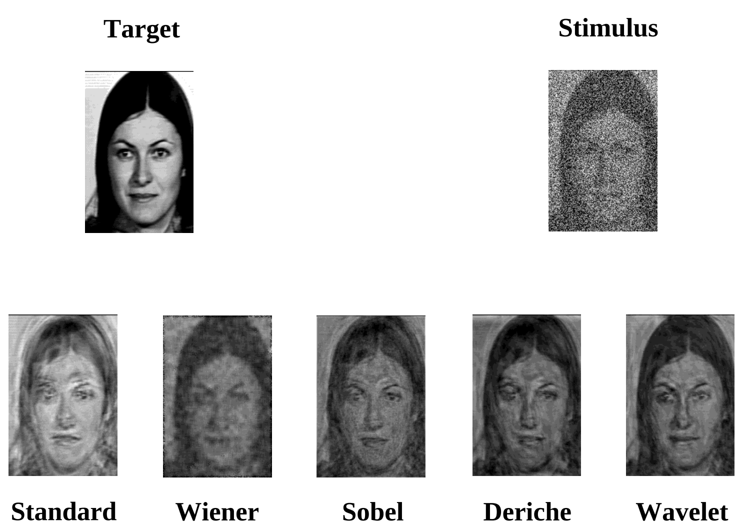

Figure 4: Response of the models. The top panels show a target face (previously learned by the auto-associator) and the stimulus used to test the model (the stimulus is a noisy version of the target). The bottom panels show the responses of the different models. The wavelet model performs best.

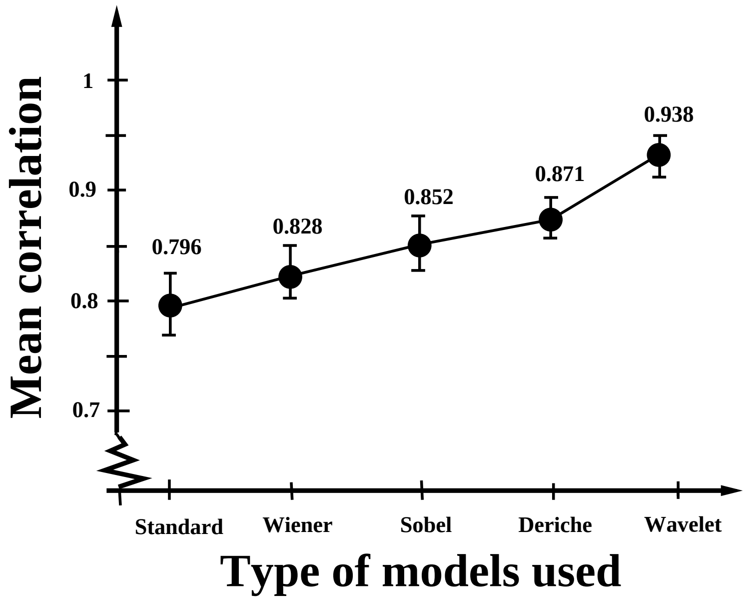

Figure 5: Mean correlation between the model responses and their targets (the higher the correlation, the better the performance). The wavelet model performs best.

The quality of recognition (estimation) can be measured by computing the cosine between the vector $\widehat{\mathbf{x}}$ (i.e., the response of model) and $\mathbf{x}_k$ (i.e., the original stimulus which is also the desired response or target). Fig. 5 gives the average correlation between response and target for the 5 models used. Clearly, the standard method is the worst of all. Pre-processing the images improves the performance of the auto-associator, and the wavelet transform gives the best result. In conclusion, the multiscale resolution (i.e., wavelet pre-processing) approach leads to the best performance for the auto-associator. Therefore, we decided, in what follows, to consider only this approach.

4 Pattern completion of noisy patterns

We have applied the multiscale edge pre-processing to store a set of 80 Caucasian faces (40 males and 40 females). In order to evaluate the effect due to pre-processing, we tested the model with different levels of Gaussian random noise added to the test stimulus. Learning was implemented as described in the previous section. For simplicity, we decided to keep only two models: the standard auto-associator and the wavelet enhanced auto-associator. Testing was implemented as described in the previous section except that faces were tested under 4 different levels of noise. The noise intensity was chosen such that its range was respectively: $[0..15]$, $[0..30]$, $[0..45]$, and $[0..60]$. Fig. 6 displays an example of the noisy test stimuli used along with the response of each model (standard and wavelet).

Figure 6: The top panels show 4 stimuli obtained by adding to a target stimulus a noise component of magnitude equal to 1, 2, 3, and 4 times the magnitude of the signal of the original stimulus. The middle panels show the responses produced by the standard auto-associator, and the bottom panels the response of the auto-associator when learning is “enhanced” with wavelet filtered stimuli.

Figure 7: Stimuli and responses. The top panels illustrates performance for a new Caucasian face, the bottom panels for a new Japanese face. From left to right the panels show: 1) a noise degraded stimulus (the magnitude of the noise is equal to twice the magnitude of the original signal); 2) the response of the standard auto-associator; 3) the response of the wavelet enhanced auto-associator.

We also decided to explore the performance of the model with 3 different types of face stimuli: 1) previously learned faces, 2) new faces similar to the learned faces, and 3) new faces coming from an other race than the learned faces. This was done in order to evaluate the robustness of the models in terms of response generalization to new stimuli. Fig. 7 displays, as an example, the responses of both models for 2 new faces (from top to bottom): 1) a new face similar to the set of learned faces (Caucasian face), and 2) a new face different from the set of learned faces (Japanese face). The auto-associator trained with the standard learning is not able to produce distinguishable responses. As can be seen in Fig. 7, better results are obtained with wavelet preprocessing.

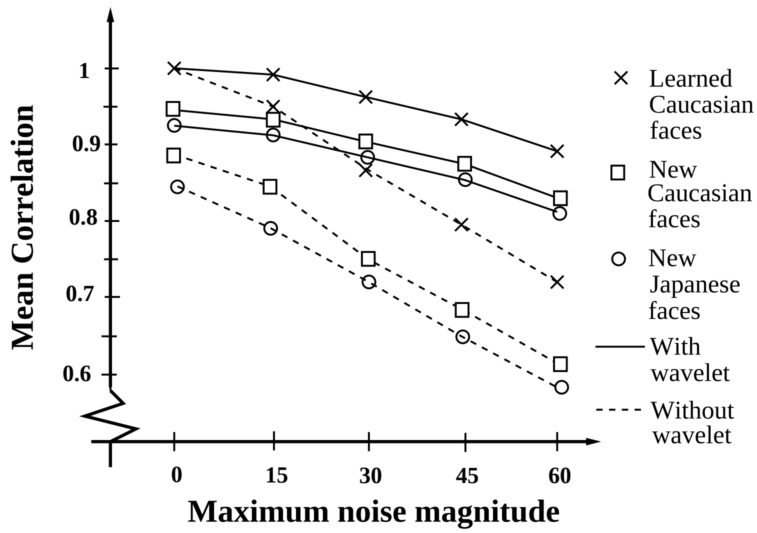

Figure 8: Mean correlation between model response as a function of the magnitude of the noise for Caucasian faces previously learned, new Caucasian faces, and new Japanese faces. The filled lines correspond to the wavelet enhanced model, the dotted line to the standard auto-associator. The wavelet enhanced auto-associator always performs best.

Fig. 8 displays the mean correlation between noiseless face images and the output for each model: 1) for 80 previously learned Caucasian faces, 2) for 80 new Caucasian faces, 3) for 80 new Japanese faces. In all cases, pre-processing the image improves the performance of auto-associator with the improvement being more important when the noise added is larger.

5 Conclusion

In this paper, we have explored the effects of storing, in a linear auto-associator, filtered versions of face images in addition to the original images. Compared to the Sobel operator and the simple Canny-Deriche operator the multiscale Canny-Deriche operator (i.e., a wavelet filter) gives the best performance for a pattern completion task involving degraded face images. The multiscale Canny-Deriche operator produces better generalization performance than the control with or without noise added to the image. The larger the amount of noise added, the larger the improvement in performance. We are now exploring the effects of using the multiscale Canny-Deriche operator for other traditional face processing tasks (see, e.g., [11]).

References

[1] T. Kohonen, "Associative memory: A system theoretic approach", *Springer-Verlag*, Berlin, 1977.

[2] J.A. Anderson, J.W. Silverstein, S.A. Ritz and R.S. Jones, "Distinctive features, categorical perception, and probability learning: Some applications of a neural model," *Psychological Review*, 84, 413–451, 1977.

[3] D. Valentin, H. Abdi and A.J. O'Toole, "Categorization and identification of human face images by neural networks: A review of the linear autoassociative and principal component approaches," *Journal of Biological Systems*, 2, 413–429, 1994.

[4] H. Abdi, D. Valentin and B. Edelman, "Neural networks," *Thousand Oaks (CA)*, Sage, 1998.

[5] H. Abdi, "*Les Rseaux de neurones,*" Presses Universitaires de Grenoble, Grenoble, 1994.

[6] D. Valentin, H. Abdi, A.J. O'Toole and G.W. Cottrell, "Connectionist models of face processing: A survey," *Pattern Recognition*, 27, 1209–1230, 1994.

[7] X. Jia and S. Nixon, "Extending the feature vector for automatic face recognition," *IEEE-Transactions on Patterns Analysis and Machine intelligence*, 17, 1995.

[8] R. Deriche, "Using canny's Criteria to Derive a Recursively Implemented optimal Edge Detector," *International Journal of Computer Vision*, 1, 167-187, 1987.

[9] E. Bourennane, M. Paindavoine and F. Truchetet, "Amlioration du filtre de Canny-Deriche pour la dtection des contours sous forme de rampe," *Traitement du signal, Recherche*, 10, 297-310, 1993.

[10] S. Mallat and Z. Zhong, "characterization of signal from multiscale edges," *IEEE-PAMI*, 14, 710-732, 1992.

[11] D. Valentin and H. Abdi, "Can a linear auto-associator recognize faces from new orientations? " *Journal of the Optical Society of America*, series A, 13, 717-724, 1996.

Fan Yang received a B.S degree in eletrical engineering from the University of Lanzhou (China) in 1982. She was a Scientific Assistant at the Departement of electronics in the University of Lanzhou. She received an M.S. (D.E.A.) degree in computer science and a Ph.D. degree in image processing from the University of Burgundy (France), respectively in 1994 and in 1998. Her research interests are in the areas of patterns recognition, neural network, motion estimation based on spatio-temporal Gabor filters, parallelism and real-time implementation, and more specifically, automatic face image processing : algorithms and architectures.

Michel Paindavoine received the Ph.D. in Electronics and Signal Processing from the Montpellier University, France 1982. He was with Fairchild CCD company for two years as engineer specialized on CCD sensors. He joined the Burgundy University in 1985 as “Maitre de Conferences” and is currently full professor and member of LE2I, Laboratory of Electronic, Computing and Imaging Sicences, Burgundy University, France. His main research topics is image acquisition and real time image processing. He is also a member of ISIS (a resarch group in Signal and Image Processing of the French National Scientific Research Committee).

Hervé Abdi received an M.S. in Psychology from the University of Franche-Comté (France) in 1975, an M.S. (D.E.A.) in Economics from the University of Clermond-Ferrand (France) in 1976, an M.S. (D.E.A.) in Neurology from the University Louis Pasteur in Strasbourg (France) in 1977, and a Ph.D. in Mathematical Psychology from the University of Aix-en-Provence (France) in 1980. He was an assistant professor in the University of Franche-Comté (France) in 1979, an associate professor in the University of Bourgogne at Dijon (France) in 1983, a full professor in the University of Bourgogne at Dijon (France) in 1988. He is currently a full professor in the School of Human Development at the University of Texas at Dallas. He was a visiting associate professor of Cognitive and Linguistic Sciences at Brown University in 1986 and 1987 and a Fulbright scholar. His interest includes neural network and cognitive modeling, experimental design and statistics.