2008

A hierarchical learning network for face detection with in-plane rotation

Fok Hing Chi Tivive
University of Wollongong, tivive@uow.edu.au

Abdesselam Bouzerdoum
University of Wollongong, bouzer@uow.edu.au

A hierarchical learning network for face detection with in-plane rotation

Abstract

This paper presents a scale and rotation invariant face detection system. The system employs a hierarchical neural network, called SICoNNet, whose processing elements are governed by the nonlinear mechanism of shunting inhibition. The neural network is used as a face/nonface classifier that can handle in-plane rotated patterns. To train the network as a rotation invariant face classifier, an enhanced bootstrap training technique is developed, which prevents bias towards the nonface class. Furthermore, a multi-resolution processing is employed for scale invariance: an image pyramid is formed through subsampling and face detection is performed at each scale of the pyramid using an adaptive threshold. Evaluated on the benchmark CMU rotated face database, the proposed face detection system outperforms some of the existing rotation invariant face detectors; it has fewer false positives and higher detection accuracy.

Disciplines

Physical Sciences and Mathematics

Publication Details

F. Tivive & A. Bouzerdoum, "A hierarchical learning network for face detection with in-plane rotation," Neurocomputing, vol. 71, (16-18) pp. 3253-3263, 2008.

A Hierarchical Learning Network for Face Detection with In-Plane Rotation

Fok Hing Chi Tivive and Abdesselam Bouzerdoum *

School of Electrical, Computer and Telecommunications Engineering, University of Wollongong, Northfields Avenue, Wollongong, NSW 2522, AUSTRALIA

Abstract

This paper presents a scale and rotation invariant face detection system. The system employs a hierarchical neural network, called SICoNNet, whose processing elements are governed by the nonlinear mechanism of shunting inhibition. The neural network is used as a face/nonface classifier that can handle in-plane rotated patterns. To train the network as a rotation invariant face classifier, an enhanced bootstrap training technique is developed, which prevents bias towards the nonface class. Furthermore, a multi-resolution processing is employed for scale invariance: an image pyramid is formed through subsampling and face detection is performed at each scale of the pyramid using an adaptive threshold. Evaluated on the benchmark CMU rotated face database, the proposed face detection system outperforms some of the existing rotation invariant face detectors; it has fewer false positives and higher detection accuracy.

Key words: Feedforward neural network, Convolutional neural network, Rotation invariant face detection, Scale invariant face detection, Shunting inhibitory neurons, Bootstrap training method.

1 Introduction

In recent years, the task of detecting human faces in an image has attracted much attention, as the human face is becoming an acceptable distinctive per-

* Corresponding author
Email addresses: f.tivive@ieee.org (Fok Hing Chi Tivive), a.bouzerdoum@uow.edu.au (Abdesselam Bouzerdoum).

[1] This work is supported in part by a grant from the Australian Research Council (ARC)

sonal trait that has the potential to be applied in many surveillance and biometric applications. However, developing systems capable of recognizing faces independent of appearance (i.e., size, pose and location) in an image is a challenging visual pattern recognition problem; perfect invariance is very difficult to achieve because of the computation inaccuracies and the continuous nature of some transformations [1]. Therefore, many researchers have proposed approaches for detecting frontal or quasi-frontal faces; for a comprehensive review on face detection, see, e.g., [2,3]. There are a few face detection systems which have been reported to be rotation invariant [4–8]. These systems often employ an extra module to estimate the rotation angle of the face pattern, which is used to transform the pattern into an upright position before it is passed to the face classifier. To overcome the problem of estimating the rotation angle, we propose a face detector based on a hierarchical neural network that can classify in-plane rotated faces in an image, regardless of their orientation.

The rest of this paper is organized as follows: Section 2 presents an overview of some popular techniques applied to invariant pattern recognition problems. Section 3 describes the proposed network architecture (SICoNNet) and its basic computing element, the shunting inhibitory neuron. Section 4 describes the proposed rotation invariant face detection system, which includes face classification and face localization. This is followed by the experimental results and performance analysis, presented in Section 5. Finally, Section 6 presents concluding remarks.

2 Background

Over the past three decades, a variety of techniques have been developed to deal with specific or general instances of invariance in pattern recognition [1,9]. Such techniques can broadly be grouped into three approaches: integral invariance, algebraic invariance and machine learning approaches. The first two approaches usually employ a pre-processing stage to eliminate certain variations in the input pattern (see, e.g., [10, 11]). This usually involves the transformation of the input space into another space in which the extracted features are invariant to some geometric transformations. A simple transform which is invariant to translations is the magnitude spectrum (i.e., magnitude of the Fourier transform on the input pattern). More advanced transforms, such as Fourier-Mellin integral [12] and moment functions [13, 14], have been used to define a set of descriptors that are invariant to rotation, translation, and scaling. However, all these invariant transforms have their own drawbacks. For example, the Fourier-Mellin integral is costly to compute and converges only under certain strong conditions [15]. The geometric moments, on the other hand, suffer from a high degree of information redundancy [16] and are sensitive to noise; such problems have been investigated by many researchers,

e.g., [17–19].

Various neural network algorithms have been developed to address the issue of invariant recognition [1, 9]. One of the first neural-based approaches for rotation invariance was proposed by Rumelhart and his colleagues [20], who employed the concept of weight sharing in a neural network to achieve translation and 90-degree rotation invariance. More powerful neural architectures, such as high-order networks, have been developed to achieve the desired invariant recognition [21, 22]. These networks can incorporate domain specific knowledge into the network structure and capture the higher-order correlations of the inputs that are necessary for the network to learn geometrically invariant properties [23]. However, higher-order networks have a major drawback in that the number of trainable weights increases combinatorially with the order of the network, and thus increases the computational cost. Other researchers, inspired by the mammalian visual system, developed biologically plausible network by representing the geometric variation internally with forward and lateral connections, whose weights are adapted by a hybrid competitive/Hebbian learning rule [24]. Satoh and his colleagues, on the other hand, extended the Neocognitron architecture, developed by Fukushima for visual pattern recognition [25,26], to cope with the recognition of rotated patterns [27]. They used a stack of cell planes in each layer, where each cell plane represents a different rotation angle of the extracted features; However, this strategy results in a network architecture with a huge number of connections. Convolutional neural networks (CoNNs) have also been used for invariant pattern recognition. Fasel and Gatica-Perez [28] developed a rotation invariant CoNN for facial expression recognition. Their network consists of three hidden layers, and each hidden layer has two sub-layers: a shared feature group sub-layer and a blurring feature group sub-layer. The shared feature group is used for feature extraction, whereas the blurring feature group is employed to reduce the number of feature planes in the previous sub-layer.

For rotation invariant face detection, Rowley and co-workers [29] developed a system that uses two neural networks: one for estimating the pose of the face and the other for classifying the rectified image window into a face or nonface pattern. Other researchers proposed invariant face detection systems by combing a skin color model to detect regions of interest (ROIs) and a face detector to locate faces in the ROIs. Zhang *et al.* [30] employed a trained neural network to estimate the rotation angle, and a template matching technique to classify the de-rotated input pattern. Phung *et al.* [31], on the other hand, employed an eye detector to locate the potential face candidates in the ROI, followed by a Bayesian classifier to determine whether the input window is a face or nonface. Wu *et al.* [32] employed the real Adaboost algorithm, similar approach as Viola *et al.* [33], to develop a multi-view face detection system consisting of three view-based detectors. Each detector uses three layers of cascade look-up table type weak classifiers for estimating the pose of the face

and six layers of weak classifiers for face detection.

3 Shunting Inhibitory Convolutional Neural Networks

Recently, we have developed a new class of convolutional neural networks called *SICoNNets*, which use *shunting inhibitory neurons* as feature detectors. The motivation for using such neurons for feature detection is that the biophysical mechanism of shunting inhibition has been employed successfully for modeling a number of visual and cognitive functions, see, e.g., [34]. Furthermore, it has been shown that shunting inhibitory artificial neuron networks can solve classification and regression problems more efficiently than their tradition counterparts, *multilayer perceptrons* (MLPs) [35–37]; for example, a single shunting inhibitory neuron is capable of forming complex nonlinear decision boundaries, and hence it can solve linearly non-separable classification problems. SICoNNets have been applied to several visual pattern recognition problems, namely face detection [38, 39], texture segmentation [40, 41], handwritten digit recognition [42], and gender recognition [43, 44]; here, SICoNNets are applied to in-plane rotation invariant face detection.

3.1 SICoNNet Architecture

The SICoNNet architecture is based on the three concepts of local receptive field, weight sharing and sub-sampling, borrowed from its predecessors LeNet [45] and the neocognitron [25]. The input layer, also known as the network retina, is a two-dimensional (2-D) array used to receive inputs from the environment. After the input layer there are several hidden layers, each comprising a number of planes of shunting neurons, called *feature maps*. Each neuron in a feature map receives inputs from a small local region in the preceding hidden layer, known as the *local receptive field*, thus reducing the number of interconnections between hidden layers. Furthermore, all neurons in a feature map share the same connection strengths or weights; thereby, the number of free parameters that needs to be trained is reduced since the number of weights is related to the size of the receptive field, not the size of the input plane. The idea of using a local receptive field and a weight sharing mechanism is to constrain every neuron in the feature map to perform the same operation on different parts of the input plane; subsequently, the same visual feature is extracted from different positions in the input plane. However, different feature maps in the same convolutional layer are used to extract different types of local features, i.e., they have different sets of weights. The feature maps in successive layers are connected using one of the three interconnection schemes: one full-connection scheme and two partial-connection schemes.

In the full-connection scheme, each feature map is connected to all feature maps of the succeeding layer, and each hidden layer can have an arbitrary number of feature maps. In the first partial-connection scheme, Fig. 3a, each feature map may have one-to-one or one-to-many interconnections with feature maps of the preceding layer, forming a toeplitz connection matrix; hence, it is referred to as a toeplitz-connection scheme. Figure 3b depicts the second partial-connection scheme where each feature map is connected to two feature maps in the following layer, forming a binary tree; therefore, it is referred to as a binary-connection scheme [46].

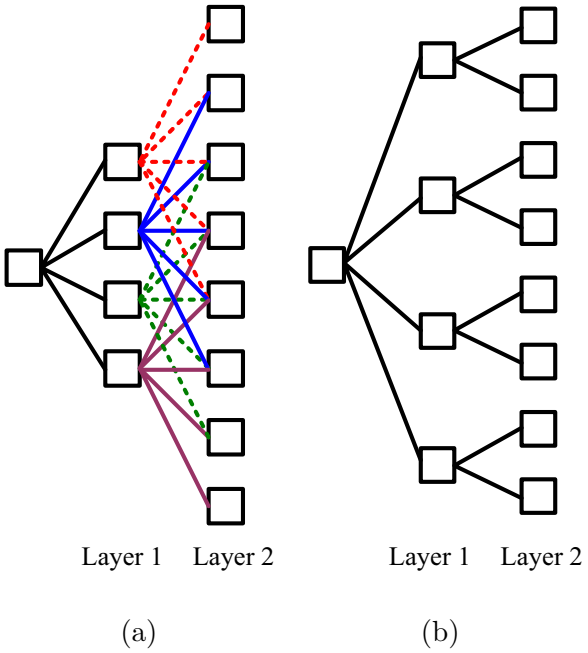


Fig. 1. Partial connection schemes: (a) toeplitz-connection scheme, (b) binary-connection scheme.

The feature maps of the last hidden layer are smoothed out with a Gaussian filter and fed to the output layer. The processing units at the output layer can be sigmoid type neurons, radial basis functions, or shunting inhibitory neurons. Although more complex classifier can be employed for the classification stage, the aim here is to investigate the capability of the proposed network to extract features from the image which can be used to detect rotated faces. To this end, a linear neuron is used to compute the network response

$$y = \sum_{i=1}^{S_N} w_i z_i + b, \tag{1}$$

where y is the response of the output neuron, w_i's are the connection weights, z_i's are the smoothed features from the last hidden layer, S_N is the number of input signals, and b is the bias term.

The main difference between SICoNNets and other convolutional neural networks is that the feature maps of SICoNNets are made up of shunting inhibitory neurons. Mathematically, the response of a static feedforward shunting inhibitory neuron located at position (i, j) in the feature map of layer L is expressed as

$$Z_L(i,j) = \frac{f\Big(\sum[C * Z_{L-1}]_{(2i)(2j)} + b(i,j)\Big)}{a(i,j) + g\Big(\sum[D * Z_{L-1}]_{(2i)(2j)} + d(i,j)\Big)}, \tag{2}$$

where "$*$" denotes 2-D convolution, $[C]$ and $[D]$ are convolution masks comprising the set of trainable weights, $b(i,j)$ and $d(i,j)$ are bias terms, $a(i,j)$ is the passive decay term, and f and g are the activation functions. To avoid dividing by zero in (2), the following condition is imposed on the decay rate parameter:

$$\Big[a(i,j) + g\Big(\sum[D * Z_{L-1}]_{(2i)(2j)} + d(i,j)\Big)\Big] \geq \varepsilon > 0. \tag{3}$$

This constraint is enforced during both the initialization and training phases.

One interesting property of the shunting inhibitory neuron is that its input-output transfer characteristic is adaptive, even with fixed activation functions, f and g; in other words, the shape of the input-output transfer characteristic can be altered by varying the neuron weights and biases only. Figure 2 illustrates some examples of the input-output transfer characteristics of the shunting neuron.

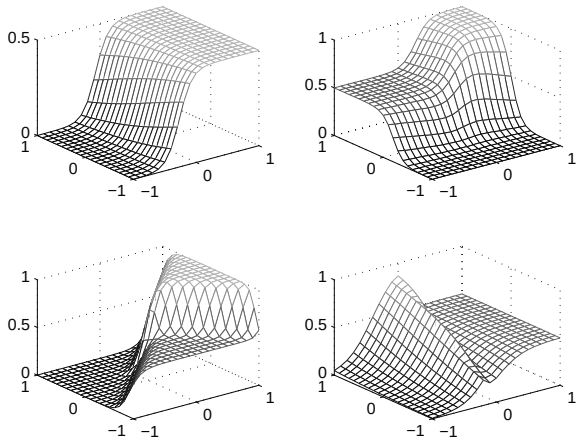


Fig. 2. Input-output transfer characteristic of the shunting inhibitory neuron using a logarithmic sigmoid for the two activation functions, f and g.

4 Rotation Invariant Face Detection

This section describes the design of a rotation invariant face detection system based on SICoNNets. The system comprises two stages: face classification and face localization. Face classification is achieved with a SICoNNet trained to detect faces with arbitrary in-plane rotation. The network architecture of the face classifier is described in the next subsection. Subsection 2 introduces a bootstrap training procedure employed with the SICoNNet face classifier. The last subsection describes a multiresolution procedure for locating faces of arbitrary size in a given image.

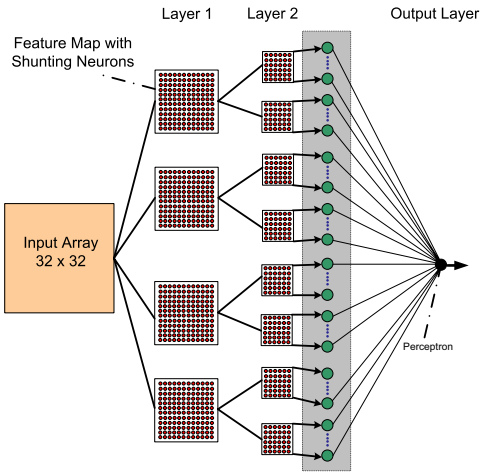


Fig. 3. A schematic diagram of the face classifier.

4.1 Face Classification

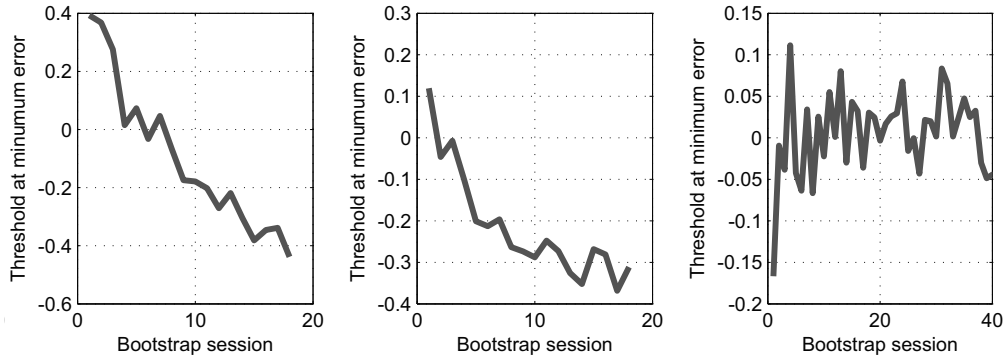
To achieve in-plane rotation invariant face classification a SICoNNet is employed. The network architecture of the face classifier consists of a 32×32 input retina, one output unit, and two hidden layers: the first hidden layer comprises four feature maps, and the second one has eight feature maps. Figure 3 presents a schematic diagram of the network architecture of the face classifier. The output neuron has a linear activation function, whereas the shunting inhibitory neurons, in the hidden layers, use the hyperbolic tangent and exponential as activation functions, f and g, respectively. A 5×5 receptive field is used throughout the network. This receptive field size is chosen by performing a series of preliminary experiments, in which a number of networks are trained with receptive field sizes ranging from 5×5 to 9×9. The results of the preliminary experiments did not show significant differences in performance between various networks; therefore, a 5×5 receptive field is chosen since it has the lowest computational cost. For the smoothing filter, a Gaussian kernel of size 5×5 is used for filtering the feature maps of the last

hidden layer. After training, the SICoNNet is used as a rotation invariant face classifier in the face detection system; it receives a 32×32 input pattern and produces an output indicating whether or not the input is a face, regardless of its orientation.

4.2 Bootstrap Training

After the preliminary experiments which identified the network with the best generalization performance, the chosen network is subjected to further training using a bootstrap procedure. The bootstrap training technique was first used by Sung and Poggio [47] to augment the training set with difficult nonface patterns, in order to improve the performance of the classifier; such technique is now widely used for training pattern classifiers when the negative class is broad and not very well defined [4, 48–50]. In Sung and Poggio's method, an initial training set, which contains more face than nonface patterns, is used to train a classifier. The trained classifier is then applied to scenery images to extract background windows that are falsely classified as faces (false detections). These false detections are added to the existing training set and the training is resumed based on the augmented training set; thereby, the number of nonface patterns is increased for each bootstrap session. Garcia and Delakis, on the other hand, improved the bootstrap training scheme using an adaptive threshold to collect nonface patterns [51]. Initially, a training set is created with equal number of face and nonface patterns. In the first bootstrap session, a high threshold of 0.8 is used so as to avoid collecting a large number of nonface patterns with low positive responses. In the subsequent bootstrap sessions, the threshold is gradually reduced until it reaches zero. Furthermore, random face patterns are added to the training set to prevent the trained network from biasing towards the nonface class.

Here the two aforementioned bootstrap schemes are evaluated by testing the trained network at each bootstrap session and recording the threshold at which the minimum classification error occurs. Figures 4a and 4b show that the minimum classification error for each trained network occurs at a negative threshold, indicating that the trained network is favoring the negative or nonface class. To correct this problem, we employ an enhanced bootstrap training procedure. Firstly a small training set with equal number of face and nonface patterns is used to start the training process (e.g., 200 face and nonface patterns). At each bootstrap session, the threshold is taken as the point at which the total number of false detections and false dismissals is minimized over the validation set. The network is then applied as a face filter to scan scenery images. The scanned windows that are falsely classified as faces are collected, and a small number of them (e.g., 200) is added to the training set. The numbers of face and nonface patterns are balanced by adding new face patterns



(a) (b) (c)

Fig. 4. The threshold computed at the minimum classification error on the validation set at each bootstrap session, based on (a) Sung's method, (b) Garcia's method, and (c) our approach.

to the training set. The new face patterns are not selected randomly, but they are chosen as the face patterns that cause false dismissals. Figure 4c shows the optimum threshold (i.e., the threshold yielding minimum classification error) of the enhanced bootstrap technique: this threshold oscillates around zero, and it is smaller in magnitude than the thresholds in Figs. 4a and 4b. This shows that determining the thresholds using a validation set prevents biasing in bootstrap training.

4.3 Face Localization Procedure

Once the network has been successfully trained as a rotation invariant face classifier, the next step is to design a scheme for detecting and locating faces of arbitrary scale in digital images. Since the face classifier can only detect faces of size 32×32 pixels, the input image is subsampled at different scales to form an image pyramid. At each level of the image pyramid, the subsampled image is processed by the network to generate a map of network responses: the entire subsampled image is used as input to the network, where it is convolved with the receptive field masks and processed according to (2). Furthermore, the output of the feature maps are subsampled before they are sent to the next layer. Normally, the down-sampling operation is performed by discarding the feature map response located at odd/even rows and odd/even columns; as a result, the subsampled feature map is 1/4 of its original size. Here, the feature map responses of the first hidden layer are arranged into four different planes, as shown Figs. 5b-5e. These planes are down-sampled by a factor of two (vertically and horizontally) after the output of the second hidden layer. The subsampled feature maps are then merged together to form the final network

response map; this reduces the response map to 1/2 the size of the original input image. The map of network responses generated from the network is compared to a decision threshold, and those responses that are above the threshold are considered face candidates, and retained for further processing. This process is repeated for every subsampled image in the pyramid.

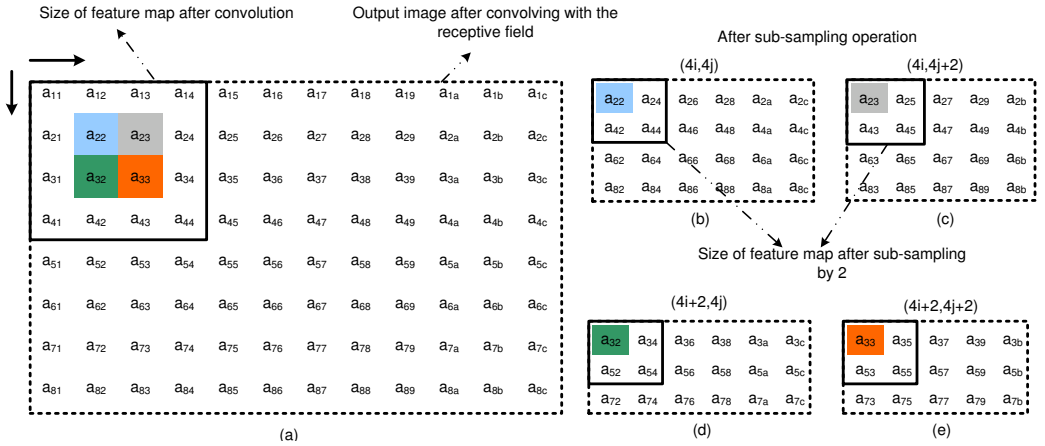


Fig. 5. The arrangement of the output images resulting from down-sampling the first hidden layer feature maps: (a) feature map before down-sampling, (b) subsampled feature map obtained by discarding odd rows and columns, (c) subsampled feature map taken at even rows and odd columns, (d) subsampled feature map taken at odd rows and even columns,(e) subsampled feature map taken at odd rows and odd columns.

In general, the decision threshold is often computed from a given test set, where the minimum error occurs. However, in practice such threshold does not work well for every test image. Therefore, we have developed a scheme to compute an adaptive threshold. At each level of the image pyramid, the threshold is taken as the average of positive network responses at the current and higher levels. At Level L_0, the top level of the image pyramid, i.e. the smallest image, the threshold is

$$T_0 = \frac{1}{P_0} \sum_{i=1}^{P_0} y_{0,i}^{+}, \tag{4}$$

where $y_{0,i}^{+}$'s are the positive network responses, and P_0 is the number of positive responses at level L_0. At the next level, L_1, the threshold is given by

$$T_1 = \frac{1}{P_0 + P_1} (\sum_{i=1}^{P_0} y_{0,i}^{+} + \sum_{i=1}^{P_1} y_{1,i}^{+}) = \frac{K_0 T_0 + \sum_{i=1}^{P_1} y_{1,i}^{+}}{K_1}, \tag{5}$$

where $K_0 = P_0$, $K_1 = K_0 + P_1$, and $y_{1,i}^{+}$'s are the positive network responses

at Level L_1. Hence, for the L_Nth level, the threshold is computed as follows:

$$T_N = \frac{K_{N-1}T_{N-1} + \sum_{i=1}^{P_N} y_{N,i}^{+}}{K_{N-1} + P_N}. \tag{6}$$

From (6), it is clear that only the threshold and the total number of positive network responses of the previous level are required to store for computing the decision threshold of the current level of the image pyramid.

However, it is unavoidable that during the face detection process a certain number of background windows will generate high network responses, and hence be misclassified as face candidates. Therefore, a number of post-processing steps are performed to reduce the number of false detections. First each detected face candidate is folded along the Y-axis (to obtain the mirror-image pattern) and passed to the face classifier; the average response of the input window and its mirror-image is taken as the final network response. If the network response exceeds the threshold T_{net}, where T_{net} is set to the computed threshold of the last layer, N, of the pyramid, the input window is deemed a face candidate. Usually, a number of overlapping detections occur around the true face forming a cluster of face candidates. Furthermore, the true face is not always the face candidate with the highest network response. To improve the confidence score of the true face, the number of overlapping detections around the face is taken as the face score. The overlapping detections are then merged into a single representative face candidate using the following clustering technique. First, all the face candidates from the series of output images are sorted in descending order according to their face scores. Suppose that S_{max} is the size of the face candidate with the highest score. All the face candidates whose centers are within a neighborhood of $0.25S_{max}$ from the center of the top face candidate are grouped into a single face cluster, and the cluster is removed from the list of face candidates. The process is repeated until all face candidates in the list are clustered. For each cluster, the center of the representative face is taken as the centroid of the cluster and its confidence score is computed as the sum of all face candidate scores in the cluster. The confidence score of the cluster is then used to verify the corresponding representative face candidate by comparing it to a threshold. This verification strategy has been used in many face detection systems to reject false detections [47, 51].

To estimate the true size and position of the detected face, two fine searches are performed. The representative face is first tested at nine scales, ranging from 0.4 to 1.6 of the detected size. The size of the representative face is computed as the average size of the positive detections and the face score is the volume of positive network responses. Then, the position of the face is searched within a grid of eight pixels around its center. The location of all detected face candidates are weighted by their network responses and averaged to give the final location. The confidence score of the face is calculated by

summing all the positive network responses together with its original score, i.e., the volume of positive network responses computed in the search of the true face size. Now since the remaining face candidates have high confidence scores, they are stored in a list and sorted in descending order. A search for overlapping face candidates is then performed, starting around the center of the face with the highest confidence score. Those overlapping face candidates whose centers are within a search region of size $0.5S_{max}$ are rejected, where S_{max} is the size of the face with the highest confidence score. Furthermore, if the intersecting area of the overlapping face candidate is greater than $0.2S_{max}^2$, the face candidate is also removed.

5 Experimental Results and Performance Analysis

In this section, the performances of the face classifier and face detector are evaluated and compared to those of other classifiers and face detectors. The next subsection presents the experimental results of the face classifier, and the last subsection analyzes the performance of the face detection system.

5.1 Performance Analysis of the Face Classifier

In this subsection we analyze the performance of the SICoNNet face classifiers, assess their robustness to in-plane rotations, and compare their performances with those of some popular neural networks. The training process of the face classifiers, including the database used for training and testing, is presented next.

5.1.1 Training of the Face Classifier

The training of SICoNNet as a rotation invariant face classifier is based on a large face database, which includes rotated and quasi-frontal face patterns. The quasi-frontal face patterns are taken from the image database created by Phung *et al.* [52], which contains people of different ages, ethnic backgrounds, and different lighting conditions. The rotated face patterns are generated by rotating face images at different angles, in the range $\pm 90°$, at $15°$ steps. Figure 6 illustrates some examples of the rotated and quasi-frontal face patterns used in the training and testing of the SICoNNets face classifier. The training set consists of 2,000 quasi-frontal face patterns, 4,000 rotated face patterns and 6,000 nonface patterns. A set of 20,000 rotated face patterns, including their mirror-image counterparts, and 20,000 nonface patterns is used for testing. The nonface patterns were collected using the bootstrap technique. The

input patterns are normalized by scaling linearly every image pixel to the range $[-1, 1]$, and the desired output labels are set to 1 for a face and -1 for nonface. To train the network, a variant of Levenberg-Marquardt algorithm (LM), proposed by Ampazis and Perantonis [53], is used for training the face classifier; its derivation for SICoNNet is reported in [46].

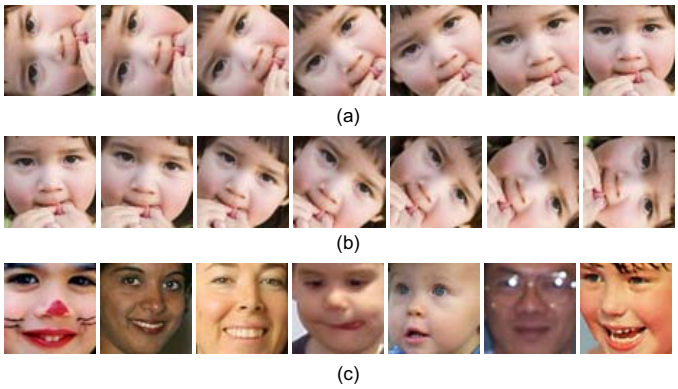


Fig. 6. Face patterns: (a) face rotated in the range $[-90°, 0°]$, (b) face rotated in the range $[0°, 90°]$, and (c) quasi-frontal faces.

To investigate the training performance of the proposed face classifier, three network topologies, the binary-, toeplitz- and fully-connected networks, are trained for 100 epochs using batch training, i.e., all training patterns are used to update the network parameters at each epoch. The training mean square error (MSE) of the three networks are recorded, and then plotted versus the number of epochs, as shown in Fig. 7. The results in this figure show that all three networks are successfully trained after 40 epochs, with the binary-connected network having the best convergence speed among the three. This may be due to the fact that the binary-connected network has fewer connections. For example, for a network architecture with 12 feature maps (4-8 configuration), the binary-connected network has to perform 24 convolution operations (i.e., 12×2: 12 interconnections by 2 receptive fields per feature map); the toeplitz- and fully-connected networks, on the other hand, perform 48 and 72 convolution operations, respectively. In terms of classification accuracy, the correct face classification rates are presented Fig. 8, as a function of training epochs. Based on both the training and test sets, all three networks achieve over 95% classification accuracy. Among the three networks, both partially-connected networks have faster training speed and better generalization performance than the fully-connected network.

5.1.2 Rotation Invariance of the Face Classifier

Another experiment is conducted to determine the built-in rotation invariance of the proposed face classifier. For this experiment, a set of 2000 quasi-front

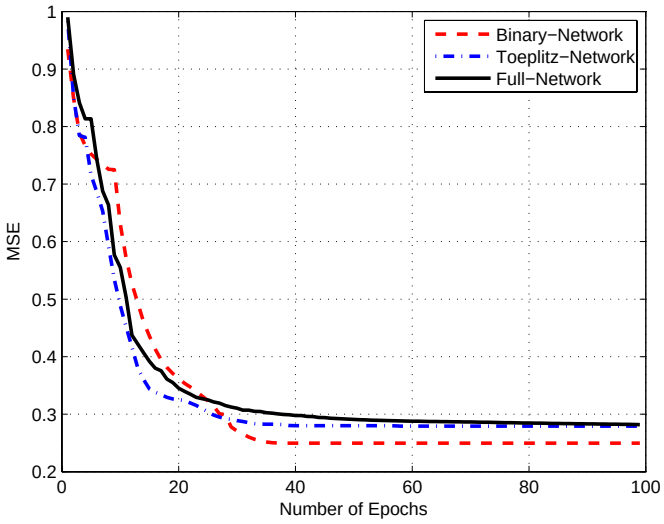


Fig. 7. The training speed of the proposed three networks.

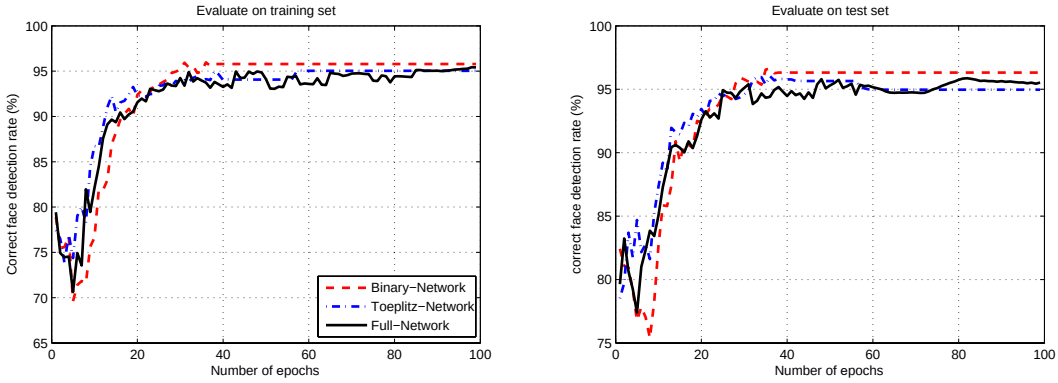
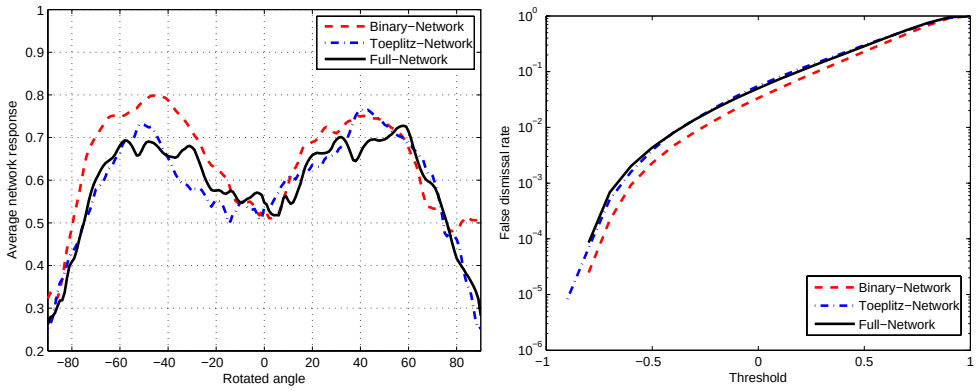


Fig. 8. The correct face classification rate of the proposed classifier based on the training and test sets.

face images are selected from Phung *et al.* face database [52]. The face images are rotated at steps of one degree between $-90°$ and $90°$; consequently, 181 rotated face patterns are generated from each face image. To ascertain whether the trained networks always generate positive responses for rotated faces, the three proposed network architectures are tested on the 362,000 rotated face patterns. Figure 9a presents the average network responses at each rotation angle. All average responses remain positive, but the average response of the binary network is generally higher than the other two, which suggests that the binary network is more tolerant to rotations. Figure 9b illustrates the false dismissal rates as a function of the decision threshold. The false dismissal rates of all three networks remain small at zero threshold, with the binary network having the lowest false dismissal rate. Another way to assess the rotation invariance of the classifier is to measure the difference between the response of the rotated pattern and that of its reference (non-rotated) face; ideally, there should be no difference between the two responses, but full rotation invariance is very hard to achieve. Figure 10 shows the histograms of response deviations

from the reference responses. The width of the histogram indicates the degree of rotation invariance achieved by the network: the wider is the histogram, the less robust the network response to rotations. All three histograms are unimodal with standard deviations of 0.322 for binary network, 0.323 for toeplitz network, and 0.326 for fully-connected network. This again suggests that the binary-connected network is slightly more tolerant to rotations. Therefore, based on these experimental results, the binary-connected network is chosen for the implementation of the face detection system.



(a) (b)

Fig. 9. Comparison among three SICoNNets: (a) the average response versus the rotated angle of the face pattern and (b) the false dismissal rate versus the threshold.

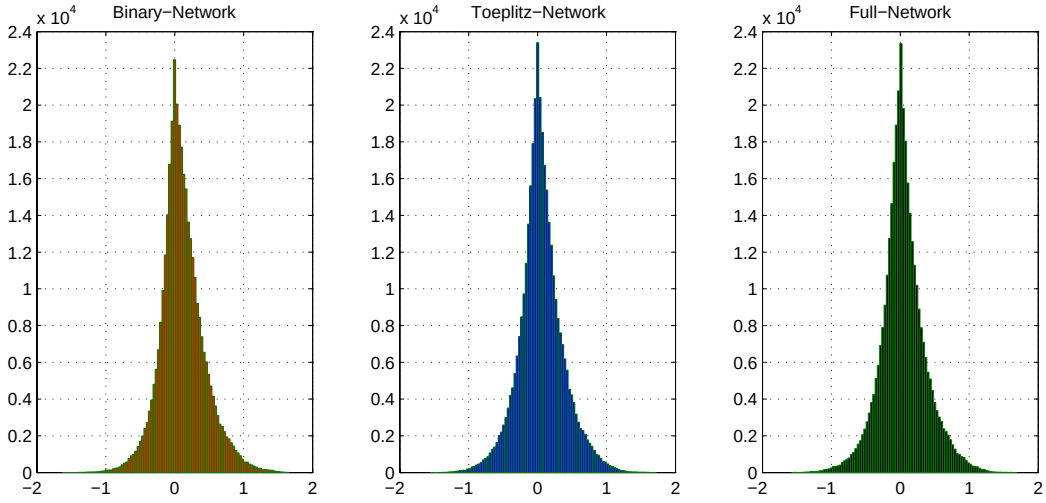


Fig. 10. The histograms of the network responses deviated from their respective frontal face patterns.

5.1.3 Performance Comparison of Face Classifiers

In this subsection, the performance of the SICoNNet, as a rotation invariant face classifier, is compared with those of two other popular neural networks:

the multilayer perceptron (MLP) and LeNet, the convolutional neural network proposed by LeCun [54]. Here a fully-connected SICoNNet is used for comparison because it has similar network topology as LeNet. The latest in a series of convolutional neural networks proposed by LeCun and his colleagues is called LeNet-5 [45]. In this network, a layer is added after the convolution layer to downsample the feature maps. This is done by averaging every four outputs of the feature map, multiplying it with a trainable weight and then passing it to an activation function. As a result, the classification performance of LeNet is improved [45]. However, to make a fair comparison, we have implemented the earlier version of LeNet [54], which has similar network topology as the fully-connected SICoNNet, except for the processing elements—the SICoNNet uses the shunting neuron, whereas LeNet uses the sigmoid neuron. Both the MLP and LeNet are trained and tested on the same data set as the SICoNNet. For the MLP, a number of networks were trained and tested. The networks contain one and two hidden layers, and the number of hidden units in each layer ranges from five to fifty neurons. Here we report only results from the best performing MLP networks. Figure 11 presents the *Receiver Operating Characteristic* (ROC) curves of the different face classifiers, and Table 1 lists their detection rates for a 10% false positive rate. Clearly the SICoNNet outperforms both LeNet and the MLPs: it has a correct detection rate of 97%, followed by LeNet with 95% and MLPs with 88%.

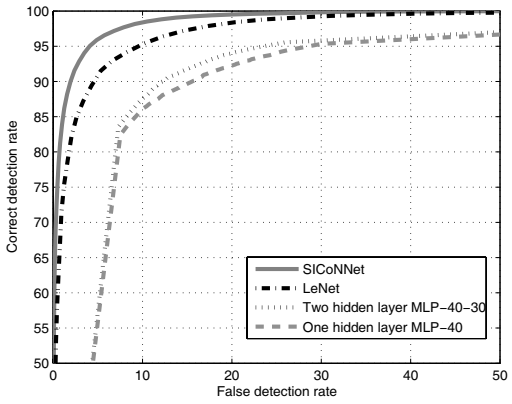


Fig. 11. ROC curves of three different networks - SICoNNet, LeNet and MLPs, tested on rotated face patterns.

Table 1
The detection rates of SICoNNet, LeNet and MLP at 10% false positive rate.

Face classifier	Detection rate
SICONNet	97%
LeNet	95%
MLP	88%

In the previous subsection, it was shown that a properly trained SICoNNet is capable of classifying face/nonface patterns with in-plane rotation; however, the rotation angle considered so far was limited to $\pm 90°$. To design a face detection system that can handle 360° in-plane rotation, the binary-connected network is trained on a larger training set that includes rotated face patterns in the range $[-180°, 180°]$. In addition, a two-layer MLP (one hidden layer and one output) is trained on local features, extracted from the feature maps of the last hidden layer, and used for face verification at the post-processing stage. However, in the multi-resolution scanning phase, only the output of the original network is used to detect the potential face candidates; this is to speed up the search for face candidates in the image pyramid.

Table 2
Detection performances of several face detectors including our face detection system.

Face Detection System	CMU rotated face database	
	Correct detection rate	False detections
Proposed face detection system	95.96%	107
Neural-based face detector [4]	90.1%	303
Cost-sensitive Adaboost algorithm [7]	93.7%	303
Original Adaboost algorithm [6]	89.7%	221
Spectral histogram/SVM [8]	93.0%	137
Model-based clustering algorithm [5]	91.0%	196

The proposed face detection system is compared with five other well-known face detectors, which have been proposed recently. These detectors include the face detector developed by Rowley *et al.* [4], which uses two neural networks, the Jones and Viola Adaboost cascade classifier [6], the cost-sensitive Adaboost algorithm of Ma and Ding [7], the spectral histogram/SVM (support vector machine) face detector proposed by Waring and Xiuwen [8], and the moded-based clustering algorithm of Jeon *et al.* [5]. All these systems were tested on the benchmark CMU face database, which has a rotated test set of 50 images with 223 faces. Some of these test images were scanned from newspapers or magazines, and others were taken from the Web [4]. Few images are quite noisy, and hence they are filtered with a Gaussian filter before they are passed to the face detector. The detection performance of the proposed network, along with those of the aforementioned face detectors, are listed in Table 2, in terms of correct detection rate and the number of false detections.

Among the existing face detection systems listed in the table, the cost-sensitive Adaboost method has the highest detection rate (93.7% with 303 false detections), followed by the spectral histogram/SVM (93.0% correct detection rate and 137 false detections). In comparison to these approaches, the proposed method has the highest detection rate of 95.96% and the lowest number of false detections, 107. Figure 12 illustrates some face detection examples using the proposed system.

Fig. 12. Examples of detected face images from the CMU rotated test set.

6 Conclusion

This paper presented a rotation invariant face detection system based on a hierarchical neural network, known as SICoNNet. The proposed neural network has a simple architecture and employs shunting inhibitory neurons for information processing; it was trained to discriminate between face and nonface patterns with in-plane rotation. The training was conducted using an enhanced bootstrap training method, which avoids biasing the trained network towards the nonface class. Compared to other neural networks, namely LeNet and the MLP, the proposed network achieves better classification rates at the same false alarm rate. Furthermore, a multi-resolution face localization procedure has been developed, which employs an adaptive decision threshold for face detection. The proposed face detection system was tested on the CMU face database that contains rotated face images, and compared to five existing face detectors. The results of the comparison show that the proposed system achieves the highest detection rate with the lowest number of false detections.

References

[1] J. Wood, Invariant pattern recognition: a review, Pattern Recognition 29 (1) (1996) 1–17.

[2] M.-H. Yang, D. J. Kriegman, N. Ahuja, Detecting faces in images: a survey, IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (1) (2002) 34–58.

[3] E. Hjelmås, B. K. Low, Face detection: a survey, Computer Vision and Image Understanding 83 (3) (2001) 236–274.

[4] H. A. Rowley, S. Baluja, T. Kanade, Neural network-based face detection, IEEE Transactions on Pattern Analysis and Machine Intelligence 20 (1) (1998) 23–38.

[5] B. H. Jeon, S. U. Lee, K. M. Lee, Rotation invariant face detection using a model-based clustering algorithm, in: IEEE International Conference on Multimedia and Expo, Vol. 2, 2000, pp. 1149–1152.

[6] M. Jones, P. Viola, Fast multi-view face detection, Tech. report TR2003-96, Mitsubishi Electric Research Laboratories (2003).

[7] Y. Ma, X. Ding, Real-time rotation invariant face detection based on cost-sensitive adaboost, in: Proc. of the International Conference on Image Processing, Vol. 3, 2003, pp. III – 921–4.

[8] C. A. Waring, L. Xiuwen, Rotation invariant face detection using spectral histograms and support vector machines, in: IEEE International Conference on Image Processing, 2006, pp. 677–680.

[9] E. Barnard, D. Casasent, Invariance and neural nets, IEEE Transactions on Neural Networks 2 (1991) 498–508.

[10] R. Lenz, Group invariant pattern recognition, Pattern Recognition 23 (1990) 199–217.

[11] M. Fukumi, S. Omatu, F. Takeda, T. Kosada, Rotation-invariant neural pattern recognition system with application to coin recognition, IEEE Transactions on Neural Networks 3 (2) (1992) 272–279.

[12] D. Casasent, D. Psaltis, Position, rotation and scale-invariant optical correlation, Applied Optics 15 (7) (1976) 1795–1799.

[13] M. K. Hu, Visual pattern recognition by moment invariants, IRE Transactions Information Theory IT-8 (1962) 179–187.

[14] M. Teague, Image analysis via the general theory of moments, Journal of the Optical Society of America 70 (8) (1980) 920–930.

[15] M. A. Rodrigues, Invariants for pattern recognition and classification, Vol. 42 of Series in Machine Perception and Artificial Intelligence, World Scientific, Singapore, 2000.

[16] C.-W. Chong, P. Raveendran, R. Mukundan, Translation invariants of zernike moments, Pattern Recognition 36 (8) (2003) 1765–1773.

[17] C. H. Teh, R. T. Chin, On image analysis by the method of moments, IEEE Transactions on Pattern Analysis and Machine Intelligence 10 (4) (1988) 496–513.

[18] A. Khotanzad, J. H. Lu, Classification on invariant image representations using a neural network, IEEE Transactions on Acoustics, Speech, and Signal Processing 38 (1990) 1028–1038.

[19] M. Gruber, K. Y. Hsu, Moment-based image normalization with high noise tolerance, IEEE Transactions on Pattern Analysis and Machine Intelligence 19 (2) (1997) 136–139.

[20] D. E. Rumelhart, G. E. Hinton, R. J. Williams, Learning internal representations by error propagation, in: D. E. Rumelhart, J. L. McClelland (Eds.), Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1: Foundations, MIT Press, Cambridge Massachusetts, 1986, pp. 318–362.

[21] M. B. Reid, L. Spirkovska, E. Ochoa, Rapid training of higher-order neural networks for invariant pattern recognition, in: Proc. of the International Joint Conference on Neural Networks, Vol. 1, 1989, pp. 689–692.

[22] S. J. Perantonis, P. J. Lisboa, Translation, rotation, and scale invariant pattern recognition by higher-order neural networks and moment classifiers, IEEE Transactions on Neural Networks 3 (2) (1992) 241–251.

[23] C. L. Giles, T. Maxwell, Learning, invariance, and generalization in a high-order neural network, Applied Optics 26 (23) (1987) 4972–4978.

[24] R. Wang, A hybrid learning network for shift-invariant recognition, Neural Networks 14 (8) (2001) 1061–1073.

[25] K. Fukushima, Neocognitron: a hierarchical neural network capable of visual pattern recognition, Neural Networks 1 (2) (1988) 119–130.

[26] K. Fukushima, Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position, Biological Cybernetics 36 (4) (1980) 193–202.

[27] S. Satoh, J. Kuroiwa, H. Aso, S. Miyake, Recognition of rotated patterns using neocognitron, in: Proc. of the. International Conference on Neural Information Processing, Vol. 1, 1997, pp. 112–116.

[28] B. Fasel, D. Gatica-Perez, Rotation-invariant neoperceptron, in: Proc. of the 18th International Conference on Pattern Recognition, Hong Kong, 2006, pp. 336–339.

[29] H. Rowley, S. Baluja, T. Kanade, Rotation invariant neural network-based face detection, in: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, 1998, pp. 38–44.

[30] H. Zhang, D. Zhao, W. Gao, X. Chen, Combining skin color model and neural network for rotation invariant face detection, in: Proc. of the Third International Conference on Advances in Multimodal Interfaces, 2000, pp. 237–244.

[31] S. L. Phung, A. Bouzerdoum, D. Chai, A. Watson, Naive bayes face-nonface classifier: a study of preprocessing and feature extraction techniques, in: Proc. of the IEEE International Conference on Image Processing, Vol. 2, Singapore, 2004, pp. 1385–1388.

[32] B. Wu, H. Ai, C. Huang, S. Lao, Fast rotation invariant multi-view face detection based on real adaboost, in: Proc of the Sixth IEEE International Conference on Automatic Face and Gesture Recognition, 2004, pp. 79–84.

[33] P. Viola, M. Jones, Rapid object detection using a boosted cascade of simple features, in: Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol. 1, Kauai, Hawaii, 2001, pp. 511–518.

[34] S. Grossberg (Ed.), Neural Networks and Natural Intelligence, MIT Press, Cambridge, Massachusetts, 1988.

[35] G. Arulampalam, A. Bouzerdoum, Application of shunting inhibitory artificial neural networks to medical diagnosis, in: Proc. of the Seventh Australian and New Zealand Intelligent Information Systems Conference, Perth, 2001, pp. 89–94.

[36] A. Bouzerdoum, A new class of high-order neural networks with nonlinear decision boundaries, in: Proc. of the Sixth International Conference on Neural Information Processing, Vol. 3, Perth, 1999, pp. 1004–1009.

[37] A. Bouzerdoum, Classification and function approximation using feed-forward shunting inhibitory artificial neural networks, in: Proc. of the IEEE-INNS-ENNS International Joint Conference on Neural Networks, Vol. 6, 2000, pp. 613–618.

[38] F. H. C. Tivive, A. Bouzerdoum, A face detection system using shunting inhibitory convolutional neural networks, Proc. of the International Joint Conference on Neural Networks 4 (2004) 2571 – 2575.

[39] F. H. C. Tivive, A. Bouzerdoum, Rotation invariant face detection using convolutional neural networks, in: Neural Information Processing, Vol. 4233, Springer Berlin / Heidelberg, 2006, pp. 260–269.

[40] F. H. C. Tivive, A. Bouzerdoum, Texture classification using convolutional neural networks, 2006, pp. 1–4.

[41] F. H. C. Tivive, A. Bouzerdoum, A nonlinear feature extractor for texture segmentation, in: IEEE International Conference on Image Processing, Vol. 2, 2007, pp. II–37–II–40.

[42] F. H. C. Tivive, A. Bouzerdoum, Application of siconnets to handwritten digit recognition, International Journal of Computational Intelligence and Applications 6 (1) (2006) 45–59.

[43] F. H. C. Tivive, A. Bouzerdoum, A shunting inhibitory convolutional neural network for gender classification, in: Proc. of the 18th International Conference on Pattern Recognition (ICPR'06), 2006, pp. 421–424.

[44] F. H. C. Tivive, A. Bouzerdoum, A gender recognition system using shunting inhibitory convolutional neural networks, in: Proc. International Joint Conference on Neural Networks, 2006.

[45] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proc. of the IEEE 86 (11) (1998) 2278–2324.

[46] F. H. C. Tivive, A. Bouzerdoum, Efficient training algorithms for a class of shunting inhibitory convolutional neural networks, IEEE Transactions on Neural Networks 16 (3) (2005) 541– 556.

[47] K. Sung, T. Poggio, Example-based learning for view-based human face detection, IEEE Transactions on Pattern Recognition and Machine Intelligence 20 (1) (1998) 31–59.

[48] E. Osuna, R. Freund, F. Girosi, Training support vector machines: an application to face detection, in: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, 1997, pp. 130–136.

[49] S. Ben-Yacoub, B. Fasel, J. Luettin, Fast face detection using MLP and FFT, in: Proc. of the Second International Conference on Audio and Video-based Biometric Person Authentication, Washington D. C., USA, 1999, pp. 31–36.

[50] M.-H. Yang, D. Roth, N. Ahuja, A SNoW-based face detector, in: S. Solla, T. K. Leen, K.-R. Müller (Eds.), Advances in Neural Information Processing Systems, Vol. 12, MIT Press, 2000, pp. 855–861.

[51] C. Garcia, M. Delakis, Convolutional face finder: a neural architecture for fast and robust face detection, IEEE Transactions on Pattern Analysis and Machine Intelligence 26 (11) (2004) 1408–1423.

[52] S. L. Phung, A. Bouzerdoum, D. Chai, Skin segmentation using color pixel classification: analysis and comparison, IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (1) (2005) 148 – 154.

[53] N. Ampazis, S. J. Perantonis, Two highly efficient second-order algorithms for training feedforward networks, IEEE Transactions on Neural Networks 13 (5) (2002) 1064–1074.

[54] Y. LeCun, Generalization and network design strategies, in: R. Pfeifer, Z.Schreter, F. Fogelman, L. Steels (Eds.), Connectionism in Perspective, Elsevier, Zurich, Switzerland, 1989.

Fok Hing Chi Tivive received the B. Eng. degree with first-class honours from Edith Cowan University, Australia, in 2001 and the Ph.D from University of Wollongong, Australia, in 2006. He is a Research Fellow at the University of Wollongong. His general research interests are in the areas of image and video processing, neural networks and machine learning, and pattern recognition.

Abdesselam Bouzerdoum is Professor of Computer Engineering and Associate Dean Research (Faculty of Informatics), University of Wollongong, Australia. He received the M.Sc. and Ph.D. degrees, both in electrical engineering, from the University of Washington, Seattle. In 1991, He joined Adelaide University as a Research Associate, and became a faculty member in June 1992. In 1998 he joined Edith Cowan University, Western Australia, as an Associate Professor. In 2004, he was appointed Professor of Computer Engineering and Head of School of Electrical, Computer and Telecommunications Engineering at the University of Wollongong. He held several Visiting Professor Appointments at Institut Galile, University of Paris-13 in (2004, 2005, and 2007), and the Hong Kong University of Science and Technology (2007). He has published over 200 technical articles and graduated fifteen Ph.D. and six Research Masters students.

Prof. Bouzerdoum has received several fellowships and distinguished awards; amongst them are the Vice Chancellor's Distinguished Researcher Award in 1998 and 1999, Awards for Excellence in Research Leadership and Excellence in Postgraduate Supervision, and the Chester Sall Award for best paper in IEEE Trans. on Consumer Electronics in 2004. In 2001 he was awarded a Distinguished Researcher (Chercheur de Haut Niveau) Fellowship from the French Ministry of Research. He served as Chair of the IEEE WA Section Signal Processing Chapter in 2004, and was Chair of the IEEE SA Section NN RIG from 1995 to 1997. From 1999 to 2006, he served as Associate Editor of IEEE Transactions on Systems, Man and Cybernetics. Currently, he is serving as Associate Editor of International Journal of Computational Intelligence and Applications and member of the governing board of the Asia Pacific Neural Network Assembly.