Improving the Neural Network Training for Face Recognition using Adaptive Learning Rate, Resilient Back Propagation and Conjugate Gradient Algorithm

Hamed Azami
M.Sc. Student
Department of Electrical Engineering, Iran University of Science and Technology, Tehran, Iran

Saeid Sanei
Associate Professor
Department of Computing, Faculty of Engineering and Physical Sciences, University of Surrey, UK

Karim Mohammadi
Professor
Department of Electrical Engineering, Iran University of Science and Technology, Tehran, Iran

ABSTRACT
Face recognition is a method for verifying or identifying a person from a digital image. In this paper an approach for classifying images based on discrete wavelet transform (DWT) and neural network (NN) has been suggested. In the proposed approach, DWT decomposes an image into images with different frequency bands. An NN is a trainable and dynamic system which can acceptably estimate input-output functions. Although the basic BP has been the most popular learning algorithm throughout all NNs applications and can be used as estimator, detector or classifier. It usually requires a very long training time. To overcome the problem, we propose several high performance algorithms that can converge few times faster than the algorithm used previously (basic BP). In this paper, the BP with adaptive learning rate, resilient back propagation (RPROP), and conjugate gradient algorithm are used to train an MLP. The simulation results show the clear superiority of the proposed method by ORL face databases.

General Terms
Face recognition, discrete wavelet transform (DWT), and neural network (NN).

Keywords
Face recognition, discrete wavelet transform (DWT), back propagation (BP), adaptive learning rate, resilient BP (RPROP) and conjugate gradient algorithm.

1. INTRODUCTION
Face recognition is a complex object recognition problem as a result of large quantity of various face expressions, lighting changes and face position [1]. Face recognition has been an active research subject due to its extensive range of applications such as security access control systems, content-based indexing, and bank teller machines [2-3]. Because of these applications, many researchers try to enhance the performance in terms of accuracy and efficiency, learning time and robustness of the computational face recognition algorithm. However, it is an important challenge due to the inherent variability of face caused by background illumination, direction, possible occlusion, emotional expression, age, gender and race. Also, the same person's face has different facial expressions and orientation. In addition, images containing faces have a high degree of variability in size, texture, background, illumination and disguise. There are several techniques to identify face or extracted features in faces [4-8].

Discrete wavelet transform (DWT) has been played a significant role in the dimension reduction and feature extraction approach by decomposing an image into frequency sub-bands at different scales. DWT coefficients are earned by passing the image through a series of filter bank stages. The procedure of appropriate design of DWT and then selecting the low frequency approximation sub-band leads to enhance the robustness of features space with respect to variation in illumination [9]. After finding the robust features as face descriptors, our purpose is to find the relations through the different faces to make decision about the face class assignment by artificial neural network (NN).

NN simulates the neuro-structure of the human brain. The brain learns from a set of experiences and experiments [10]. One of the most well known and most popular NN among all the existing NN paradigms is multi-layer perceptron (MLP) [11]. The previous methods to classify faces used basic BP with momentum. In this paper, we suggest to use adaptive learning rate and momentum with basic BP, resilient back propagation (RPROP), and conjugate gradient algorithm for decreasing the training time and increasing the accuracy ratio of the faces classification.

The RPROP algorithm is an enhanced static NN and constitutes one of the best performing first-order learning methods for NN [12]. In the simple BP algorithm, the weights in the steepest descent direction (negative of the gradient), the direction in which the performance function is decreasing most rapidly [13], are adjusted. It turns out that, although the function decreases most quickly along the negative of the gradient, it does not unavoidably create the quickest convergence. In the conjugate gradient algorithm a search is done along conjugate directions, which creates normally quicker convergence than that of the steepest descent directions [13].

Then rest of this paper is organized as follows. DWT and NN learning using basic BP are initiated in Sections 2.1 and 2.2, respectively. Then, our proposed methods using adaptive learning rate and momentum, resilient BP, and conjugate

gradient algorithm are introduced in Section 3. Section 4 determines our experimental results on the ORL dataset. Finally conclusion is given in Section 5.

2. BAKGROUND KNOWLEDGE FOR THE PROPOSED METHOD

2.1. Feature Extraction

DWT has played a significant role to reduce the dimension of an image and extract the features by decomposing an image in frequency domain into sub-bands at different scales. The DWT of an image is created as follows: In the first level of decomposition, the image is split into four sub-bands, namely HH1, HL1, LH1, and LL1, as shown in Figure 1. The HH1, HL1 and LH1 sub-bands represent the diagonal details, horizontal features and vertical structures of the image, respectively. The LL1 sub-band is the low resolution residual consisting of low frequency components and it is this sub-band which is further split at higher levels of decomposition [14]. In Figure 2 an image from ORL face database with decomposed one-level wavelet and after three-level wavelet transform are shown, respectively.

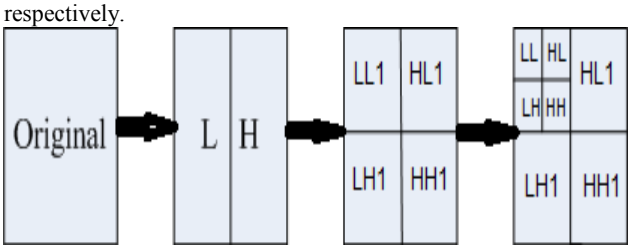

Fig 1: The process of decomposing an image.

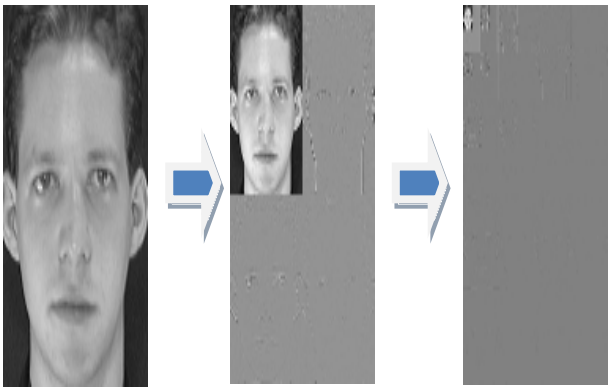

Fig 2: Original image and the DWT after one-level and three-level DWT respectively.

2.2. Neural Network Learning using basic Back Propagation

In a BP algorithm the weights are updated in two steps: forward and backward. The forward pass step computes feed-forward propagation of input pattern signals through network. For hidden units [15, 16]:

$$z_j(t) = f\left(\sum_{i=1}^{p} w_{ji}(t)x_i(t)\right) \quad (1)$$

and for output units:

$$y(t) = f\left(\sum_{j=1}^{q} w_j(t)z_j(t)\right) \quad (2)$$

The backward step propagates the error of the signal through network starting at output units (where the error is the difference between actual and desired output values). Cost function E is defined as sum of measure error squares:

$$E(t) = \frac{1}{2}e(t)^2 = \frac{1}{2}[y(t) - d(t)]^2 \quad (3)$$

where *d* and *y* are the target and output values, respectively. We want to know how to modify weights in order to decrease *E*. We utilize gradient descent as follows:

$$w_{ji}(t+1) = w_{ji}(t) - \eta\frac{\partial E(t)}{\partial w_{ji}(t)} \quad (4)$$

where the $\eta > 0$ is learning factor. Both for hidden units and output units, the partial derivative can be rewritten as product of three terms using chain rule for partial differentiation [15, 16]:

$$\frac{\partial E(t)}{\partial w_{ji}(t)} = \frac{\partial E(t)}{\partial y(t)}\frac{\partial y(t)}{\partial z_j(t)}\frac{\partial z_j(t)}{\partial w_{ji}(t)} \quad (5)$$

These terms are:

$$\frac{\partial E(t)}{\partial y(t)} = e(t)\ ,\ \frac{\partial y(t)}{\partial z_j(t)} = w_j(t)f'\left(\sum_{j=1}^{q} w_j(t)z_j(t)\right)$$

$$\frac{\partial z_j(t)}{\partial w_{ji}(t)} = x_i(t)f'\left(\sum_{i=1}^{p} w_{ji}(t)x_i(t)\right) \quad (6)$$

actually in every stage $\eta\frac{\partial E(t)}{\partial w_{ji}(t)}$ is updated until this parameter reach to its desire. Also, the similar technique can be used for the bias vectors and $w_j(t)$ [15, 16].

3. PROPOSED METHOD

3.1. Adaptive Learning Rate

In order to increase the speed of the BP, the momentum, making weight changes equal to the sum of a fraction of the last weight change and the new change, can be added to BP. The magnitude of the effect that the last weight change is allowed to have is mediated by a momentum constant, m_c, which is often selected a number between 0.8 and 0.9. The learning rate that determines the scale of the increments of the weight at every updating step absolutely influences the performance of the learning. A very large learning rate value may cause unstable oscillations though a very small learning rate may slow down the learning procedure. In previous BP algorithm used in face recognition, the learning rate was assumed to be fixed and uniform for all the weights. Moreover, the learning rate was usually kept small to prevent oscillations and thus to ensure convergence. Often, the learning rate decreases during the iterations to enable faster convergence at the beginning and having less steady state error at the end. Therefore, we can improve the previous BP method using an adaptive learning rate. This can be by reducing the learning rate if the differential error goes below a threshold level [13].

3.2. Resilient Back Propagation

Resilient BP (RPROP) is a powerful algorithm to train an NN that has a high speed to converge and the ability that often cannot allow the responses trap to local minima in the defined space [17] as mentioned in equation (4):

$$\Delta w_{ji}(t) = w_{ji}(t+1) - w_{ji}(t) = -\eta \frac{\partial E(t)}{\partial w_{ji}(t)} \quad (7)$$

It is clear that the changes of the learning rate can considerably affect the performance of the training. The RPROP try to lessen the disadvantage of this problem by using adaptively computed parameters which change in every iteration. In fact, these parameters are adjusted during the learning process based on the direction of convergence. This is based on the sign of the respective partial derivative at the current and the previous epoch as follows [18]:

if

$$\frac{\partial E(t)}{\partial w_{ji}(t)} \cdot \frac{\partial E(t-1)}{\partial w_{ji}(t-1)} > 0$$

then

$$\Delta_{ji}(t) = \min\left\{\eta^{+} . \Delta_{ji}(t-1), \Delta_{\max}\right\} \quad (8)$$

else if

$$\frac{\partial E(t)}{\partial w_{ji}(t)} \cdot \frac{\partial E(t-1)}{\partial w_{ji}(t-1)} < 0$$

then

$$\Delta_{ji}(t) = \max\left\{\eta^{-} . \Delta_{ji}(t-1), \Delta_{\min}\right\} \quad (9)$$

else

$$\Delta_{ji}(t) = \Delta_{ji}(t-1) \quad (10)$$

$$\Delta w_{ji}(t) = \text{sig n}\left(\frac{\partial E(t)}{\partial w_{ji}(t)}\right) . \Delta_{ji}(t) \quad (11)$$

where η^{-} and η^{+} are the attenuation and increase factors, respectively, and they are set $\eta^{+} \in [1.1, 1.3]$ and $\eta^{-} \in [0.5, 0.8]$. Also, the step sizes are bounded by $\Delta_{\min}$ and $\Delta_{\max}$ [18].

Sigmoid function is usually used in the hidden layers of an NN. This function compresses an infinite input range into a finite output range. When we use steepest descent to train a MLP with sigmoid functions, since the gradient can have a very small magnitude, it causes small changes in the weights and biases, although the weights and biases are far from their desired values. The aim of the RPROP algorithm is to remove these harmful effects of the magnitudes of the partial derivatives [19].

3.3. Conjugate Gradient Algorithm

In the basic algorithm, the weights in the steepest descent direction are adjusted. Although the performance function reduces most rapidly along the negative of the gradient, it does not necessarily create the fastest convergence. Conjugate gradient algorithm has faster convergence than steepest descent directions along conjugate directions, which causes the convergence to considerably increase. The conjugate gradient approach here includes the following steps:

1. In the first step, algorithm is started by searching in the steepest descent direction.

$$P_0 = -g_0 , \; g_k = \nabla F(x)|_{x=x_k} \quad (12)$$

2. Conjugate gradient algorithm takes a step and selects the learning rate to minimize the function along the search direction.

$$x_{k+1} = x_k + \alpha_k . p_k \quad (13)$$

3. Then, conjugate gradient algorithm determines the next search direction. The general procedure for determining the new search direction is by combining the new steepest descent direction with the previous search direction according to:

$$p_k = -g_k + \beta_k . p_{k-1} \quad (14)$$

where β_k is defined as follows:

$$\beta_k = \frac{\Delta g_k^{T} . g_k}{g_{k-1}^{T} . g_{k-1}} \quad (15)$$

3.4. ORL Face Database

The ORL face database contains 400 images of 40 individuals (10 different images from each person) with various facial expressions and lighting conditions. A number of images from this database are illustrated in Figure 3.

Fig 3: Sample images from ORL face database.

4. PERFORMANCE EVALUATION

In order to extract features, at the first step we implemented the Haar wavelet with four-level decomposition and then features from the lower sub-band coefficients were chosen. In this paper we use three layers MLP with 65 neurons in hidden layer. It should be mentioned that we simulated resilient BP and conjugate gradient algorithm with neurons between 20 and 100 with an increment of 5, and understood that the best training occurred when we had 65 neurons in hidden layer.
The number of epochs is selected 700 and the sigmoid function is used for training the MLPs. Also, four algorithms to learn a MLP including BP with adaptive learning rate, BP with momentum and adaptive learning rate, resilient BP and conjugate gradient algorithm are used. We increase the number of features from 10 to 80. For all these ways, we train an MLP with 50% of the images in the database (from one to five for each person) and then used the rest for testing (from six to ten for each person). The momentum and initial learning rate are set to 0.85 and 0.05, respectively. Figure 4 shows a comparative analysis of different trainings for various features. Because of uncertain behavior of NNs, we run all algorithms 20 times, and the average of the results is presented.

The results demonstrate that three proposed approaches, namely, BP with momentum and adaptive learning rate, resilient BP and conjugate gradient algorithm have the best recognition rate with using 40 features. As illustrated, the BP with momentum and adaptive learning rate has better performance than the BP with adaptive learning rate. Also, the performance of the NN using resilient BP is better than BP with momentum and adaptive learning rate. Finally, conjugate gradient algorithm is the best algorithm to train an NN that classifies the faces from ORL database. Therefore, the best classification occurs when we train an MLP with conjugate gradient algorithm and by using 40 obtained features by DWT.

We may increase the number of iteration (say to 5000) to ensure about minimum steady-state error. The results are shown in Figure 5. As can be seen in this figure, when we increase the number of iteration, recognition rate doesn't change significantly. However, the training time increases considerably. Thus, we propose to use 500 iterations for training an MLP by using conjugate gradient algorithm.

5. CONCLUSION

In order to enhance the performance of the NN-based face recognition systems in this research three techniques are proposed for classification of the ORL database images. After decomposing the images and extracting their features, we train the NN by BP with adaptive learning rate, BP with momentum and adaptive learning rate, resilient BP and conjugate gradient algorithm. The BP with momentum and adaptive learning rate has better performance than the BP with adaptive learning rate. Also, the performance of the NN using resilient BP is better than BP with momentum and adaptive learning rate. Among these methods, the conjugate gradient algorithm method gave the best results with 65 neurons in hidden layer and 40 features. Although the number of iterations for an NN is very important subject, the results by conjugate gradient algorithm show that approximately 500 iterations is enough for training the NN using this algorithm (for classifying ORL database images).

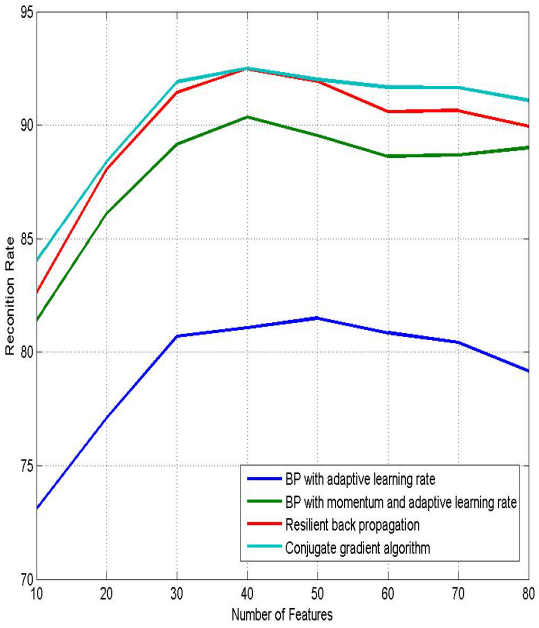

Fig 4: The performance of the proposed methods for various number features.

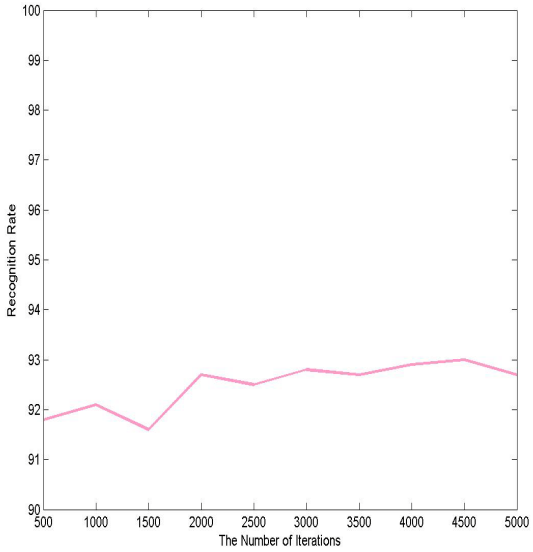

Fig. 5: Obtained recognition rate versus the number of iterations for training by conjugate gradient algorithm

6. REFERENCES

[1] M. R. M. Rizk and A. Taha, "Analysis of neural networks for face recognition systems with feature extraction to develop an eye localization based method", pp. 847-850, 2002.

[2] J. Daugman, "Face and gesture recognition: overview". IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 19, no. 7, pp. 675-676, July 1997.

[3] L. Wiskott, J. Fellous, N. Kruger and C. Malsburg, "Face recognition by elastic bunch graph matching", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 19, no. 7, pp. 775-779, 1997.

[4] M. Firdaus, "Face recognition using neural networks", International Conference on Intelligent System (ICIS), CD-ROM, 2005.

[5]. M. Firdaus, "Dimensions reductions for face recognition using principal component analysis", Proc. *11th International Symp artificial life and robotics (AROB 11th 06)*, CD-ROM 2006.

[6] L. sufen and G. junying, "Face recognition algorithm based on Local wavelet transform and DCT", vol. 22, no. 1, pp. 205-208, 2006.

[7] Y. A. Georghiades, P. Belhumeur and D. Kriegman, "From few to many: illumination cone models for face recognition under variable lighting and pose" *IEEE Transactions Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 643-660, 2001.

[8] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min and W. Worek, "Overview of the face recognition grand challenge," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 947-954, 2005.

[9] M. Ghazel, "Adaptive fractal and wavelet image denoising", Waterloo, Ontario, Canada, 2004.

[10] M. R. Mosavi, "GPS receivers timing data processing using neural networks: optimal estimation and errors modeling", *Journal of Neural Systems*, vol. 17, no. 5, pp. 383-393, 2007.

[11] M. R. Mosavi, "Precise real-time positioning with a low cost GPS engine using neural networks", *Journal of Survey Review*, vol.39, no. 306, pp. 316-327, 2007.

[12] C. Igel and M. Husken, "Empirical evaluation of the improved RPROP learning algorithms", *Neurocomputing*, vol. 50, pp. 105–123, 2003.

[13] F. Paulin and A. Santhakumaran, "Classification of breast cancer by comparing back propagation training algorithms", *International Journal on Computer Science and Engineering*, vol. 3, no. 1, pp. 327-332, 2011.

[14] D. L. Donoho, "Nonlinear wavelet methods for recovery of signals, densities, and spectra from indirect and noisy data", *in Proceeding. of Symposia in Applied Mathematics*, vol. 47, pp. 173-205, 1993.

[15] M. R. Mosavi and H. Azami, "Applying neural network ensembles for clustering of GPS satellites" *International Journal of Geoinformatics,* (accepted).

[16] D. J. Jwo and C. C. Lai, "Neural network-based GPS GDOP approximation and classification", *Journal of GPS Solutions*, vol. 11, no. 1, pp. 51-60, 2007.

[17] I. Ahmad, M. A. Ansari and S. Mohsin, "Performance comparison between backpropagation algorithms applied to intrusion detection in computer network systems", *International Conference on Neural Networks*, pp. 231-236, 2008.

[18] P. A. Mastorocostas, "Resilient back propagation learning algorithm for recurrent fuzzy neural networks", *Electronics Letters*, vol. 40, no. 1, 2004.

[19] K. Gupta and S. Kang, "Implementation of resilient backpropagation & fuzzy clustering based approach for finding fault prone modules in open source software systems" *International Journal of Research in Engineering and Technology (IJRET)*, vol. 1, no. 1, pp. 38-43, 2011.