IMPROVED GENERALIZATION BY ADDING BOTH AUTO-ASSOCIATION AND HIDDEN-LAYER-NOISE TO NEURAL-NETWORK-BASED-CLASSIFIERS

Hiroaki Inayoshi , Takio Kurita

National Institute of Advanced Industrial Science and Technology (AIST)
1-1-1 Umezono, Tsukuba, 305-8568 JAPAN

ABSTRACT

We propose a novel method for learning that improves generalization in classifiers based on neural networks. The proposed method consists of (1) adding auto-associative learning and (2) simultaneously adding independent noise to the hidden layer of the neural-network. We verify this method with the classification problem of faces under variable illumination. Considering the interpolation for untrained samples as the key aspect of generalization, we expect that in our method, neural-classifiers will (1) learn (nearly) principal components of trained samples by auto-association, and will (2) generate and learn the variated samples from trained samples (along the axes of nearly principal components) by added noise, which leads both to increased amount of trained samples and (hopefully) to improved generalization.

1. INTRODUCTION

How should you generalize when you only have a limited subset of samples? We hypothesize the following answer: First, find the principal directions (axes) of variations, then interpolate and/or extrapolate along those directions.

In this paper, we propose a novel method for learning that improves generalization in classifiers based on neural networks and show the plausibility of our hypothesis. The proposed method consists of (1) adding auto-associative learning and (2) simultaneously adding independent noise to the hidden layer of the neural-network (see Fig.1).

Following our hypothesis, we expect that in the proposed method, neural-classifiers will (1) learn (nearly) principal components, i.e. principal directions of variances, of given samples by auto-association, and will (2) generate and learn the variated samples from given samples (along the axes of nearly principal components) by added noise, which leads both to increased amount of (not given, self-generated) trained samples and (hopefully) to improved generalization. We apply the proposed method on the classification problem of faces under variable illumination.

The synergistic effect of adding both auto-association and neural-noise simultaneously has not been demonstrated before, although each of the two has separately been used: Kurita et al. [1] has added auto-associative learning to neural classifiers and Kurita et al. [2] has added independent noise to the hidden layer of the neural network. Related works concerning generalization are as follows: Bishop [3] showed that adding noise in the input-space is equivalent to Tikhonov regularization. Akaho [4] investigated the method of neural noise to interpolate the training samples. Murray et al. [5] studied the cases of synaptic weight noise. Tenenbaum and Griffiths [6] re-casted Shepard's theory in a more general Bayesian framework and showed how this naturally extends his approach to the more realistic situation of generalizing from multiple consequential stimuli with arbitrary representational structure. Our proposed method is based on different viewpoints from those works.

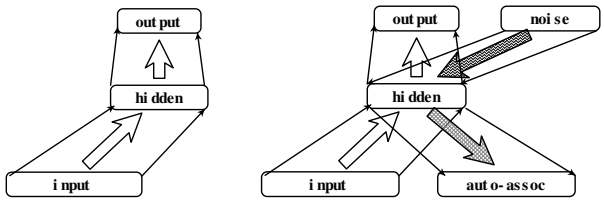

Fig. 1. Proposed (right) and conventional (left) architecture of neural-network-based-classifiers

The rest of the paper is organized as follows. In section 2, we show the learning algorithm of the proposed method. Section 3 gives an overview of the classification problem of faces under variable illumination and section 4 describes the experiments and the results. In section 5, conclusion and remaining tasks are presented.

2. LEARNING ALGORITHM

In this section we present the learning algorithm of the proposed method. (Since most of the algorithm is the same as [1] except for noise terms, we only show the important parts. For details, refer to [1].)

Consider a classifier based on a neural network, as de-

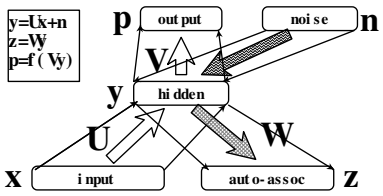

Fig. 2. A classifier based on a neural network with additions of auto-association constraint and noise upon each of the hidden layer neurons

picted in Fig.2. The main flow of information is:

$$\boldsymbol{x_j} \Longrightarrow^{U} \boldsymbol{y_j} \Longrightarrow^{V} \boldsymbol{p_j}$$

where $\boldsymbol{x_j}$ is the j-th input pattern ($j = 1, 2, ...N$) of M dimensional vector to be classified into K classes, $\boldsymbol{y_j}$ is a j-th internal representation of an H dimensional vector, and $\boldsymbol{p_j}$ is the j-th output of K dimensional vector. Note that $\boldsymbol{y_j}$ is both used for $\boldsymbol{z_j}$ of M dimensional vector (auto-association of $\boldsymbol{x_j}$), and is influenced by $\boldsymbol{n_j}$ of H dimensional vector, which is composed of H independent noise sources.

We use linear neuron model in the hidden layer and multinomial logit model [7] as the classifier. Multinomial logit model is a special case of the generalized linear model [7], and it can be regarded as one of the simplest neural network model for multi-way classification problems. Relations among the variables are as follows:

$$\boldsymbol{y_j} = \boldsymbol{U}\boldsymbol{x_j} + \boldsymbol{y_0} + \boldsymbol{n_j} \tag{1}$$

$$\boldsymbol{p_j} = \boldsymbol{f}(\boldsymbol{V}\boldsymbol{y_j} + \boldsymbol{p_0}) \tag{2}$$

$$\boldsymbol{z_j} = \boldsymbol{W}\boldsymbol{y_j} + \boldsymbol{z_0} \tag{3}$$

where $\{\boldsymbol{U}, \boldsymbol{V}$ and $\boldsymbol{W}\}$ are $\{H \times M, K-1 \times H$ and $M \times H\}$ matrices; $\{\boldsymbol{y_0}$ and $\boldsymbol{z_0}\}$ are bias terms of the same dimension as $\{\boldsymbol{y_j}$ and $\boldsymbol{z_j}\}$, respectively ($\boldsymbol{p_0}$ is a $K-1$ dimensional vector). Letting $K-1 \times H$ matrix $\boldsymbol{V}$ be composed of $(K-1)$ vectors of $1 \times H$ dimension: $\boldsymbol{v_k}(k = 1, 2, ...K-1)$ and $\eta_{jk} = \boldsymbol{v_k}\boldsymbol{y_j} + (\boldsymbol{p_0})_k$, the function $\boldsymbol{f}$ in eq. (2) is computed as the "softmax" as follows.

$$p_{jk} = \exp(\eta_{jk}) / \{1 + \sum_{i=1}^{K-1} \exp(\eta_{ji})\} \tag{4}$$

$$p_{jK} = 1 / \{1 + \sum_{i=1}^{K-1} \exp(\eta_{ji})\} \tag{5}$$

We also denote $\boldsymbol{r_j}$ as $(K-1)$ dimensional vector composed of $(K-1)$ elements of eq.(4) excluding the K-th element, i.e. eq.(5).

Now consider a classification problem with K classes $\{C_1, C_2, ...C_K\}$. Let $\boldsymbol{\tau} = (t_1, t_2, ..., t_K)^T \in \{0,1\}^K$ denote a binary vector composed of teacher signals with $t_k = 1$ if the input is C_k, otherwise $t_k = 0$. (Also let $\boldsymbol{t} = (t_1, t_2, ..., t_{K-1})^T \in \{0,1\}^{K-1}$.)

For the training samples $\{(\boldsymbol{x_j}, \boldsymbol{\tau_j})\}_{j=1}^N$, the likelihood of the classifier is given by

$$P(\boldsymbol{\tau}|\boldsymbol{y}) = \Pi_{j=1}^N \Pi_{k=1}^K p_{jk}^{\tau_{jk}} \tag{6}$$

For auto-associative learning, the following averaged sum of squares error between input $\boldsymbol{x_j}$ and its auto-association output $\boldsymbol{z_j}$ is minimized.

$$\epsilon^2 = (1/N) \sum_{j=1}^{N} ||x_j - z_j||^2 \tag{7}$$

Suppose $\sum_{j=1}^N \epsilon_j = \sum_{j=1}^N (\boldsymbol{x_j} - \boldsymbol{z_j})$ as Gaussian with zero mean, then the above minimization is equivalent to the maximization of the following.

$$L_A = (-1/2) \sum_{j=1}^{N} \epsilon_j^2 \tag{8}$$

The learning algorithm or update equations of the parameters $\{\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}, \boldsymbol{y_0}, \boldsymbol{z_0}, \boldsymbol{p_0}\}$ that maximize the sum $L_C + L_A$, where $L_C = \log P(\tau|y)$ is the log-likelihood of eq.(6) and L_A (eq.(8)) is the log-likelihood of auto-association, can be written as eqs.(9)- (14).

In the fallowing, we denote α as the learning rate, concatenate each of N vectors $\{\boldsymbol{x_j}, \boldsymbol{y_j}, \boldsymbol{z_j}, \boldsymbol{t_j}, \boldsymbol{r_j}\}$ to form $(* \times N)$ matrix $\{\boldsymbol{x_j} \rightarrow \boldsymbol{X}, \boldsymbol{y_j} \rightarrow \boldsymbol{Y}, \boldsymbol{z_j} \rightarrow \boldsymbol{Z}, \boldsymbol{t_j} \rightarrow \boldsymbol{T}, \boldsymbol{r_j} \rightarrow \boldsymbol{R}\}$, and let $\boldsymbol{E} = (\boldsymbol{T} - \boldsymbol{R})$, $\boldsymbol{F} = (\boldsymbol{X} - \boldsymbol{Z})$, $\boldsymbol{G} = \boldsymbol{V}^T\boldsymbol{E} + \boldsymbol{W}^T\boldsymbol{F}$

$$\boldsymbol{V_{new}} = \boldsymbol{V_{old}} + \alpha(\boldsymbol{E}\boldsymbol{Y}^T) \tag{9}$$

$$\boldsymbol{W_{new}} = \boldsymbol{W_{old}} + \alpha(\boldsymbol{F}\boldsymbol{Y}^T) \tag{10}$$

$$\boldsymbol{U_{new}} = \boldsymbol{U_{old}} + \alpha(\boldsymbol{G}\boldsymbol{X}^T) \tag{11}$$

$$(\boldsymbol{y_0})_{new} = (\boldsymbol{y_0})_{old} + \alpha(sum(\boldsymbol{G})) \tag{12}$$

$$(\boldsymbol{z_0})_{new} = (\boldsymbol{z_0})_{old} + \alpha(sum(\boldsymbol{F})) \tag{13}$$

$$(\boldsymbol{p_0})_{new} = (\boldsymbol{p_0})_{old} + \alpha(sum(\boldsymbol{E})) \tag{14}$$

In eqs. (12), (13), and (14), we define $sum(A)$ as

$$sum(A)_i = \sum_j (A)_{ij}.$$

3. CLASSIFICATION PROBLEM OF FACES UNDER VARIABLE ILLUMINATION

First, we point out that the purpose of this paper is not to " propose the best method for the classification problem of faces under variable illumination" but we just have selected this classification problem as an example to prove the effectiveness of our proposed method. Now we give a brief overview of this classification problem.

It has been pointed out that "the variations between the images of the same face due to illumination and viewing direction are almost always larger than image variations due to change in face identity [8]". The set of images of an object in fixed pose but under all possible illumination conditions is a convex cone (termed the "illumination cone") in the space of images [9]. When the surface reflectance can be approximated as Lambertian, this illumination cone can be constructed from a handful of images acquired under variable lighting and additionally, if the object is convex then only three "basis images" are required to build the illumination cone [9].

Representative methods for the classification problem of faces under variable illumination are "Eigenfaces" by Turk et al. [10] and "Fisherfaces" by Belhumeur et al. [11].

Recently, Savvides et al. proposed "Corefaces" which, according to their experiments, achieves almost 100 percent classification accuracy [12].

4. CLASSIFICATION EXPERIMENTS

4.1. Classification task

Using "Yale Face Database B" [13], we have performed experiments to confirm the effectiveness of the proposed method. This database consists of $5850 = 10 \times 9 \times 65$ images, taken under 585 viewing conditions (9 poses times 65 illumination conditions) for 10 individuals. We used a subset of 650 images of the same pose (frontal view). A fixed window of 240x300 pixel is applied to cut face region and then resized to 36x45 pixel. Fig.3 shows the samples of 10 individuals. In this paper, a "suit" refers to a set of images (of 10 individuals) taken under the same illumination condition. (Each row in Fig.3 corresponds to one "suit".)

We normalized each image by

$$\boldsymbol{x} = \boldsymbol{x_{raw}}/||\boldsymbol{x_{raw}}||$$

where $\boldsymbol{x}$ is the normalized image and $\boldsymbol{x_{raw}}$ is the original 1620 (=36x45) dimensional vector of input image (pixel-value).

4.2. Settings of experiments

We compare the proposed neural classifiers with the conventional ones (proposed method = conventional method + 'auto-association' + 'neural-noise'). The common parameter is the number of hidden units (H) and we set

$$H \in \{20, 10, 3\}.$$

The proposed method has an additional parameter $\boldsymbol{s}$ which determines the magnitude of neural-noise, i.e. uniform noise in the range of $[-\boldsymbol{s}/2, \boldsymbol{s}/2]$ is applied as $\boldsymbol{n_j}$ in eq.(1) We choose

$$\boldsymbol{s} \in \{0, 0.25, 0.5, 0.75, 1.0, 1, 25, 1.5, 1.75, 2.0, 2.5, 3.0\}.$$

Note that when $\boldsymbol{s} = 0$ this corresponds to the case of 'no-noise' (i.e. adding only auto-association)

For the given 65 suits (total of 650 samples), trainings and tests are done by the following two conditions:

- 10 suits (100 samples) for training and remaining 55 suits (550 samples) for test
- 20 suits (200 samples) for training and remaining 45 suits (450 samples) for test

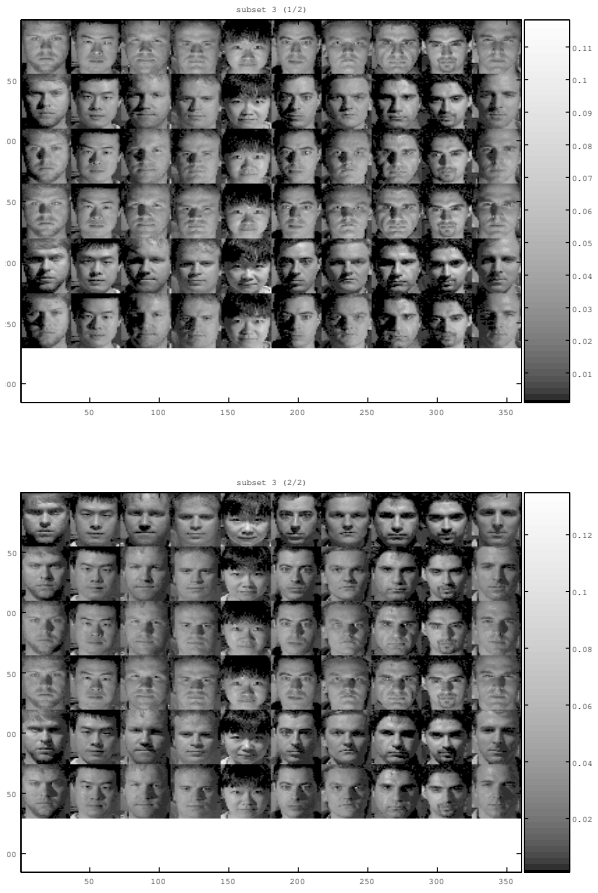

Fig. 3. Sample images of classification task

In each of these two conditions, trainings of the conventional and proposed neural classifiers are done using the same parameters (i.e. the learning rate and the number of training-iterations).

4.3. Results

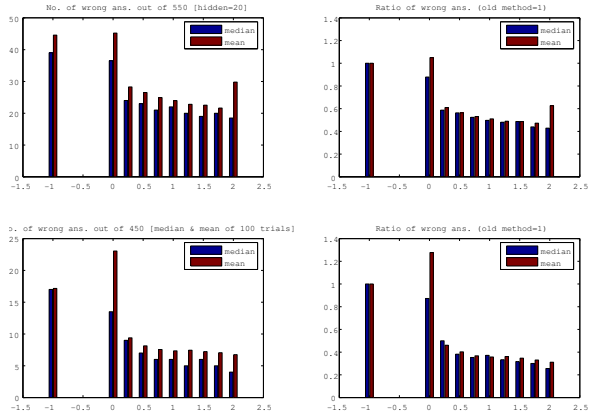

Fig. 4. Error counts (left) and ratio-to-conventional (right) when $H = 20$ (i.e. 20 hidden units)

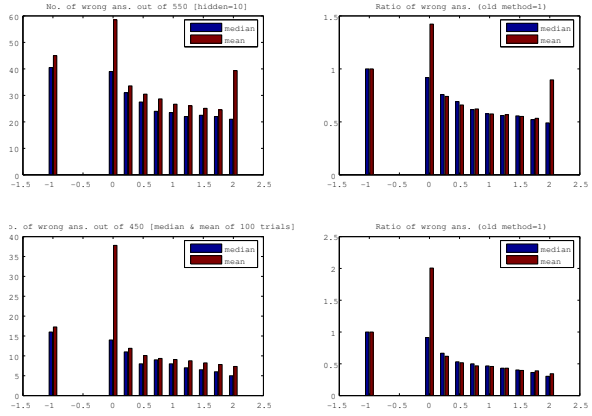

Fig. 5. Same as Fig.4 when $H = 10$

A trial of training and test is done as follows: First, for the given 65 suits, we divide randomly into '10-training-suits and 55-test-suits' (or into '20-training-suits and 45-test-suits'). Then after training by 10 (or 20) suits, remaining untrained 550 (or 450) samples are used to test (i.e. count the number of classification-errors). We perform 100 (and 100) trials and collect the number of classification-errors.

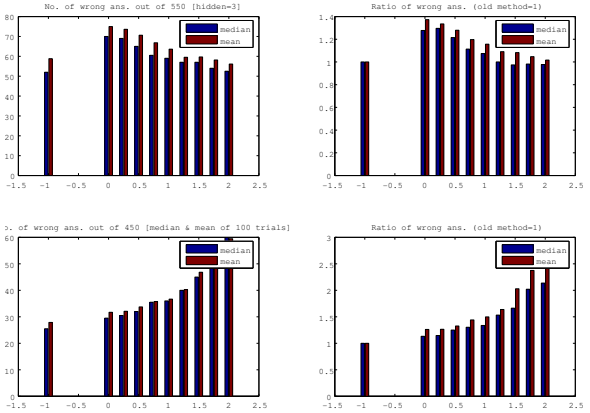

Fig. 6. Same as Fig.4 when $H = 3$

The left sides of Figs.4-6 show the median (bar-charts' left) and mean (right) of the error counts. (In those figures, the top shows 550 samples case and the bottom shows 450 samples case. Horizontal axis in each of those figures means parameter $\boldsymbol{s}$, except for the leftmost ($\boldsymbol{s} = -1$) that corresponds to conventional classifier)

The right sides of Figs.4-6 show the median (bar-charts' left) and mean (right) of 'the ratios of the error counts of proposed method to that of conventional method'.

The reason why means sometimes show bigger values than medians is as follows: In these cases, training did not converge and these cases resulted in high error counts (outliers) affecting the mean values. (The training errors were always 0 except for these divergent cases.)

From Figs.4 and 5, we can see that the proposed method with proper range of noise achieved reduction of errors by almost 50 percent to the conventional method, thus proving the effectiveness of the proposed method.

The reason why no improvements of generalization occurred when $H = 3$ (see Fig.6) is as follows: We need more than three bases to build the "illumination cone" (see section 3) and variations in only this 3 dimensional space are not supposed to be sufficient to make interpolation.

Fig.7 shows the result of a control experiment, where only neural-noise is added (no auto-association) with other conditions being the same as those of Fig.5. Note that the leftmost case (i.e. No auto-association nor neural-noise being added) is identical with the '$\boldsymbol{s} = 0$' case (adding only neural-noise of scale 0).

By comparing Figs. 5 and 7, adding both neural-noise and auto-association performs better than adding only neural-noise. More specifically, the minimum of median values for

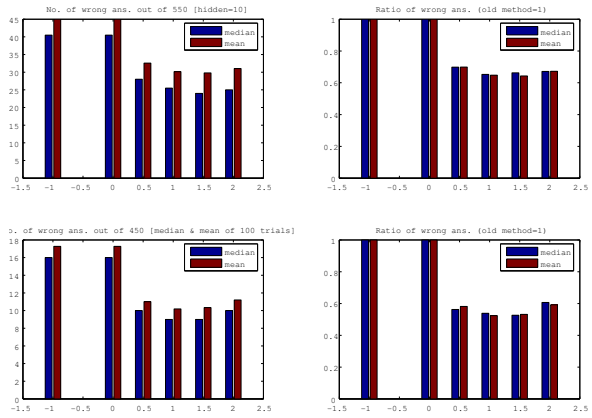

Fig. 7. Control experiment of adding only 'neural-noise' (i.e. without auto-association) when $H = 10$ (cf.Fig.5)

'ratios' are as follows: 0.4422 (both) vs. 0.6022 (noise-only) in 55-suits-tests and 0.2857 (both) vs. 0.5000 (noise-only) in 45-suits-tests.) Also by comparing the leftmost cases (i.e. conventional method) with the next-to-leftmost cases ($\boldsymbol{s} = 0$, i.e. adding only auto-association) in Figs. 4 and 5, we can see that adding only auto-association slightly improves the conventional method in those cases (excluding the divergent ones).

Fig.8 shows an example of the training result for a trial when $H = 10$. In this trial, the conventional method scored 34 errors and the proposed method with $\boldsymbol{s} = 1.75$ scored only 1 error. Each row of Fig.8 shows the following (from the top to the bottom):

1. matrix $\boldsymbol{U}$ by the conventional method
2. matrix $\boldsymbol{U}$ by the proposed method
3. matrix $\boldsymbol{W}$ by the proposed method
4. matrix $\boldsymbol{W}\boldsymbol{V}^{\boldsymbol{T}}$ and vector $\boldsymbol{z_0}$ by the proposed method

Finally, each of Figs. 9 and 10 shows the effect of added noise in the hidden layer as follows: The leftmost in the top row shows an input to the neural network, which is the average image of an individual over 65 illumination conditions. The leftmost in the second row shows the auto-association output under no noise added in the hidden layer. Other images in the top and second rows show the auto-association outputs under different combinations of uniform-noise, in the range $[-0.5, 0.5]$, added to all units in the hidden layer. Images in the third and bottom rows show the auto-association outputs when adding noise of scale 1 to only each of one unit (out of 10 hidden units) separately. From these figures, adding noise to the hidden layer seems being equivalent to

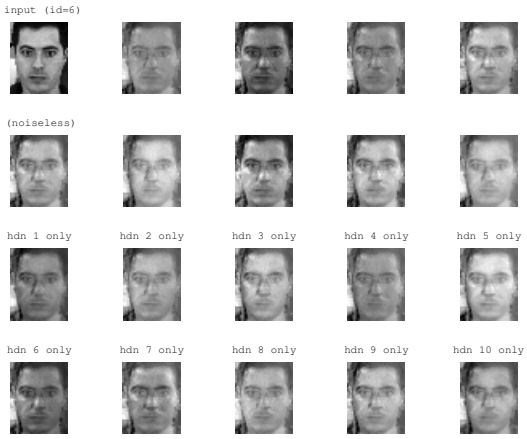

Fig. 9. Effect of added noise: examples of ARTIFICIALLY variated auto-association outputs by the proposed method using $H = 10$

Fig. 10. Same as Fig.9 for different individual

changing the illumination conditions, which is what we intended to demonstrate.

5. CONCLUSIONS

This paper proposed a novel method for learning that improves generalization in classifiers based on neural networks, i.e. method which adds both auto-associative learning and neural-noise to the hidden-layer of conventional neural classifiers. The effectiveness of the proposed method was demonstrated by the classification problem of faces under variable illumination.

The remaining tasks we intend to do are (1) to verify the proposed method by other classification problems and (2) to find out the detailed mechanism of how the proposed

Fig. 8. Example of training result when $H = 10$

method improves generalization.

6. REFERENCES

[1] Kurita T. Takahashi T. Ikeda Y., "A neural network classifier for occluded images," in *Proc. of International Conference on Pattern Recognition (ICPR2002) vol.III*, 2002, pp. 45–48.

[2] Kurita T. Asoh H. Umeyama S. Akaho S. and Hosomi A., "A structural learning by adding independent noises to hidden units," in *Proc. of IEEE Inter. Conf. on Neural Networks*, 1994, pp. 275–278.

[3] Bishop C. M., "Training with noise is equivalent to tikhonov regularization," *Neural Computation*, vol. 7, no. 1, pp. 108–116, 1995.

[4] Akaho S., "Regularization learning of neural networks for generalization," in *Algorithmic Learning Theory (ALT'92: Lecture Notes in Artificial Intelligence, Vol.743)*. 1992, Springer-Verlag.

[5] Murray A.F. and Edwards P.J., "Synaptic weight noise during mlp training : Fault tolerance and training improvements," *IEEE Trans. Neural Networks*, vol. 4, no. 4, pp. 722–725, 1993.

[6] Tenenbaum J.B. and Griffiths T.L., "Generalization, similarity, and bayesian inference," *BEHAVIORAL AND BRAIN SCIENCES*, vol. 24, pp. 629–640, 2001.

[7] McCullaph P. and Nelder J.A., *Generalized Linear Models*, Chapman and Hall, 1983.

[8] Adini Y. Moses Y. and Ullman S, "Face recognition: the problem of compensating for changes in illumination direction," Tech. Rep. Technical Report CS93-21, Mathematics & Computer Sience, Weizmann Institute Of Sience, 1993.

[9] Belhumeur P.N. and Kriegman D.J., "What is the set of images of an object under all possible illumination conditions?," *Int. Journal of Computer Vision*, vol. 28, no. 3, pp. 245–60, 1998.

[10] Turk M. and Pentland A., "Eigenfaces for recognition," *Journal of Cognitive Neuroscience*, vol. 3, no. 1, pp. 71–86, 1991.

[11] Belhumeur P.N. Hespanha J. and Kriegman D.J., "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE Trans. PAMI, Special Issue on Face Recognition*, vol. 19, no. 7, pp. 711–20, 1997.

[12] Savvides M. Kumar B.V.K.V. Khosla P.K., ""corefaces" - robust shift invariant pca based correlation filter for illumination tolerant face recognition," in *CVPR'04*, 2004, pp. 834–841.

[13] Georghiades A.S., Belhumeur P.N., and Kriegman D.J., "From few to many:illumination cone models for face recognition under variable lighting and pose," *IEEE Trans. Pattern Anal. Mach. Intelligence*, vol. 23, no. 6, pp. 643–660, 2001.