New Approach to Face Recognition Using Co-occurrence Matrix and Bayesian Neural Networks

El houssaine HSSAYNI*
Modelling and Mathematical Structures Laboratory
Faculty of Sciences and Techniques
Fez, Morocco
elhoussaine.hssayni@usmba.ac.ma

Mohamed ETTAOUIL
Modelling and Mathematical Structures Laboratory
Faculty of Sciences and Techniques
Fez, Morocco
mohamedettaouil@yahoo.fr

***Abstract*—Faces represent complex multidimensional significant visual stimuli and developing a computational model for face recognition is difficult. In this paper we present a new approach to the face recognition problem by combining Co-occurrence Matrix and Bayesian Neural Networks.Firstly, we use Co-occurrence Matrix to extract the relevant information in a face image, which are important for identification. Using this we can represent face pictures with several coefficients instead of having to use the whole picture.Then, Bayesian Neural networks are used to recognize the face through learning correct classifcation of the coeficients calculated by the Co-occurrence Matrix. The experimental results on the ORL database illustrate that the proposed approach has better performance in term of accuracy compared to old approaches.**

***Index Terms*—Bayesian Neural Networks, Co-occurrence Matrix, Evidence Framework, Face Recognition, Machine Learning, Multilayer Perceptron.**

I. INTRODUCTION

As one of the most active and visible research topics in computer vision, pattern recognition and biometrics, face recognition has been extensively studied in the past two decades, yet it is still a challenging problem in practice due to uncontrolled environments, occlusions and pose variations, illumination, expression and aging, etc. Face recognition system would find countless applications, e.g. criminal identification and retrieval of missing children, workstation and building security, credit card verification and many more.
Among the many models proposed to solve the problem of face recognition, we have the neural networks [14], which occupy for about twenty years a significant place in difficult problems of classification and pattern recognition that are encountered precisely in face recognition [1].
Considerable efforts have been made by researchers to improve neural networks. Among these improvements are the integration of Bayesian methods into neural network learning [3], which have shown considerable potential in such problems. Bayesian learning to artificial neural networks consists of considering all network weights are random variables and their distribution depends on the inverse of the variance called hyperparameters. The posterior probability of weights is calculed by combining the prior information with the likelihood of data using Byes theorem [4].
The next section presents and describes the co-occurrence matrix and the relevant features of face, which are important for identification. Section III describes Bayesian learning for neural networks and precisely the approach based on the re-estimation of the hyperparametres nomed evidence framework (EF). In section IV we describe the used data base. Then a hybrid system based on Co-occurrence Matrix and Bayesian Neural Networks is proposed. And before concluding, experimental results are given in Section VI.

II. FEATURE EXTRACTION BY GRAY-LEVEL CO-OCCURRENCE MATRIX

The Gray Level Co-ocurrence Matrix (GLCM) method is a way of extracting second order statistical texture features [16]. A co-occurrence matrix measures the probability of appearance of pairs of pixel values located at a distance in the image. It is based on the calculation of the probability $P(i, j, \delta, \theta)$ which represents the number of times a color level pixel i appears at a relative distance δ and a color level pixel j and in a given θ orientation.
Most images are 256-grayscale, so the size of the co-occurrence matrices is 256×256. We realize that these matrices account for a very large amount of information difficult to exploit directly. Therefore we will define the six most important factors containing the maximum information on the image from the co-occurrence matrix:

a) **Energy:**

$$ENE = \sum_i \sum_j \Big(P_{ij}(\delta, \theta)\Big)^2 \tag{1}$$

This parameter measures the uniformity of the texture. It reaches strong values when the distribution of the levels of grey is constant or of periodic shape. In this last case, the high values of energy are obtained for matrices $P(\delta, \theta)$ when (δ, θ) corresponds to the period.

978-1-7281-6654-4/20/$31.00 ©2020 IEEE

b) **Contrast:**

$$CST = \sum_i \sum_j \Big((i-j)^2 P_{ij}(\delta, \theta) \Big) \tag{2}$$

The value is raised all the more as the texture presents a strong contrast. This parameter is strongly correlated in the energy.

c) **Entropy:**

$$ENT = \sum_i \sum_j \Big(log(P_{ij}(\delta, \theta))\, P_{ij}(\delta, \theta) \Big) \tag{3}$$

This parameter measures the disorder in the image. Contrary to the energy, the entropy reaches strong values when the texture is completely random (without visible structure). It is strongly correlated (by the opposite) in the energy.

d) **Variance:**

$$VAR = \sum_i \sum_j \Big((i-\mu)^2\, P_{ij}(\delta, \theta) \Big) \tag{4}$$

The variance measures the heterogeneousness of the texture. She increases when the levels of different grey of their average. The variance is independent from the contrast.

e) **Correlation:**

$$COR = \sum_i \sum_j \Big(\frac{(i-\mu)(j-\mu) P_{ij}(\delta, \theta)}{\sigma^2} \Big) \tag{5}$$

COR measure the linear dependance (with regard to (δ, θ)) levels of grey of the image.
The correlation is correlated neither in the energy, nor in the entropy.

f) **Inverse Difference Moment:**

$$IDM = \sum_i \sum_j \frac{P_{ij}(\delta, \theta)}{1 + (i-j)^2} \tag{6}$$

IDM (Inverse Difference Moment) Measure the homogeneity of the image. This parameter is correlated in a linear combination of variables ENE and CST.

III. Bayesian Neural Networkss

The Bayesian approach for neural networks is to consider all parameters of networks are random variables; then by using the Bayes'theorem and based on prior overall parameters and likelihood function, the posterior distribution can be computed [10]. So, instead of computing weights which are minimize the erreur function by the classical approach of maximum likelihood, Bayesian methods give a complete distribution for the Neural Network parameters. This posterior distribution can then be used, to deduce predictions of the network for new values of the input variables.

A. The prior distribution

As we have a priori little idea of what the weight values should be, the prior is, therefore, chosen as a rather broad distribution. This can be done by expressing the prior pdf as a Gaussian distribution with a large variance:

$$P(w|\alpha) = \frac{1}{Z_w(\alpha)} exp(-\alpha E_w) \tag{7}$$

Where $Z_w(\alpha)$ is a normalization factor given by

$$Z_w(\alpha) = \int exp(-\alpha E_w) dw \tag{8}$$

$$= \Big(\frac{2\pi}{\alpha} \Big)^{m/2} \tag{9}$$

In the Bayesian framework, α is called a hyper-parameter as it controls the distribution of other parameters, and represents the inverse of the variance on the set of weights and biases.

B. The likelihood function

Given a training dataset D of N examples of the form $D = \{x_i, d_i\}_{i=1}^N$, the goal of NN learning is to find a relationship R between x_i and d_i. Since there are uncertainties in this relation as well as noise or phenomena that are not taken into account, this relation becomes

$$d_i = R(x_i) + \varepsilon_i \tag{10}$$

where the noise ε_i is an expression of the various uncertainties. We suppose that the errors have a normal distribution with zero mean and variance $\sigma^2 = 1/\beta$, then the distribution of the noise is given by

$$P(\varepsilon_i|\beta) = \sqrt{\frac{\beta}{2\pi}} \Big(-\frac{\beta}{2} \varepsilon i \Big) \tag{11}$$

So the joint probability of noise is given by

$$P(\varepsilon|\beta) = P(\varepsilon_1, \varepsilon_2, ..., \varepsilon_N|\beta) \tag{12}$$

$$= \prod_{i=1}^N P(\varepsilon_i|\beta) \tag{13}$$

$$= \Big(\frac{\beta}{2\pi} \Big)^{N/2} exp\Big(-\frac{\beta}{2} \sum_{i=1}^N \varepsilon_i^2 \Big) \tag{14}$$

We have $\varepsilon_i = d_i - y_i$, so by remplacing in the previous expression we obtain

$$P(\varepsilon|\beta) = \Big(\frac{\beta}{2\pi} \Big)^{N/2} exp\Big(-\frac{\beta}{2} \sum_{i=1}^N (d_i - y_i)^2 \Big) \tag{15}$$

Then the likelihood function is given by:

$$P(D|w, \beta) = \frac{1}{Z_D(\beta)} exp(-\beta E_D) \tag{16}$$

Where $Z_D(\beta)$ is a normalization factor given by

$$Z_D(\beta) = \left(\frac{2\pi}{\beta}\right)^{N/2} \tag{17}$$

C. The posterior distribution

Using Bayes' theorem, the posterior distribution is given by

$$P(w|\alpha, \beta, D) = \frac{P(D|w, \beta)P(w|\alpha)}{P(D|\alpha, \beta)} \tag{18}$$

Then we obtain

$$P(w|\alpha, \beta, D) = \frac{1}{Z_S(\alpha, \beta)} exp[-S(w)] \tag{19}$$

Where $S(w)$ is the regularized error given by

$$S(w) = \alpha E_w + \beta E_D \tag{20}$$

And $Z_S(\alpha, \beta)$ is the normalization constant given by

$$Z_S(\alpha, \beta) = \int exp(-S(w))dw \tag{21}$$

In the general case this integral can not be analytically calculated. In this framework Makay [3] proposed an estimation of the probability a posteriori via the hypothesis of the Gaussian approximation.

D. Gaussian approximation to the posterior distribution

Gaussian approximation for the posterior distribution is obtained by considering the Taylor expansion of $S(w)$ around its minimum value w_{MP} and retaining terms up to second order so that

$$S(w) \approx \Big(-S(w_{MP}) - \frac{1}{2}(w - w_{MP})^t A(w - w_{MP}) \Big) \tag{22}$$

Where A is the Hessian matrix of the total (regularized) error function $S(w)$, with elements given by

$$A = \nabla\nabla S(w) \tag{23}$$

The Gaussian approximation of the posterior probability also makes it possible to calculate the normalization factor after the approximation $Z_S^{'}(\alpha, \beta)$ which is given by:

$$Z_S^{'}(\alpha, \beta) = (2\pi)^{m/2} * |A|^{-1/2} * exp(-S(w_{MP})) \tag{24}$$

Where $|A|$ is the determinant of the hessian matrix A.

E. The evidence framework

To calculate the normalization factor $Z_S^{'}(\alpha, \beta)$ and the normalization constant $Z_w(\alpha)$ we considered that the hyper-parameters α and β were known. But these hyperparameters must also be estimated.

Using Bayes' theorem, the posterior distribution of the weights can be written as

$$P(\alpha, \beta|D) = \frac{P(D|\alpha, \beta)P(\alpha, \beta)}{P(D)} \tag{25}$$

where $P(\alpha, \beta)$ is the prior over the hyper-parameters, called a hyper-prior, $P(D|\alpha, \beta)$ is the likelihood term called the evidence for α and β and $P(D)$ is a normalization factor.

Using Gaussian approximations for the posterior PDF of weights, we can write the term of evidence in the following form

$$P(D|\alpha, \beta) = \frac{Z_S^{'}(\alpha, \beta)}{Z_w(\alpha) Z_D(\beta)} \tag{26}$$

$$= \frac{(2\pi)^{m/2} * |A|^{-1/2} * exp[-S(w_{MP})]}{\left(\frac{2\pi}{\alpha}\right)^{m/2} \left(\frac{2\pi}{\beta}\right)^{N/2}} \tag{27}$$

The optimal values of the hyperparameters α_{MP} and β_{MP} correspond to a maximization of the evidence are given by

$$\alpha_{MP} = \frac{\gamma}{2E_w^{MP}} \tag{28}$$

$$\beta_{MP} = \frac{N - \gamma}{2E_D^{MP}} \tag{29}$$

The quantity γ denotes the number of parameters properly determined and given by:

$$\gamma = \sum_{k=1}^{m} \left(\frac{\lambda_k}{\lambda_k + \alpha}\right) \tag{30}$$

where the λ_k are the eigenvalues of the Hessian matrix of the un-regularized error E_D, i.e. $H = \beta\nabla\nabla E_D$.

As we need to find the optimum weight vector w_{MP}, as well as the optimal values of α and β, an iterative process is used to tune the network. This iterative procedure (nomed evidence framework (EF)) is described below:

1. Choose initial (small) values for α and β. Initialize the weights in the network using values drawn from the prior distribution. At iteration i, given the current estimates of the weights w^i, the hyper-parameters α^i and β^i, we can compute a current estimate of $S^i(w)$.
2. Find weights w_{MP}^{i+1} that minimizes $S^i(w)$ by using a standard non-linear training algorithm like Gauss–Newton or scale conjugate gradient. Given w_{MP}^{i+1}, compute E_w^{i+1} and E_D^{i+1}
3. Compute new values of α^{i+1} and β^{i+1} using the three successive sub-steps a, b and c:
   a) $\gamma^{i+1} = \sum_{k=1}^{m}(\frac{\lambda_k}{\lambda_k+\alpha})$ where λ_k is the kth eigenvalue of the Hessian matrix of the un-regularized error, i.e. $\beta^i\nabla\nabla E_D^{i+1}$.
   b) $\alpha^{i+1} = \frac{\gamma^{i+1}}{2E_w^{i+1}}$
   c) $\beta^{i+1} = \frac{N-\gamma^{i+1}}{2E_D^{i+1}}$
4. Iterate to step 2 and use parameters w^{i+1}, α^{i+1} and β^{i+1} to compute new values for the parameters.

In practice it is good to start with small values of α.

The convergence of the evidence framework is obtained when the regularized error $S(w)$ is equal to half the number of data points.

IV. Used data base

We evaluate our proposed approach with ORL database, which is made of 40 subjects having each one 10 different views [18]; the images in gray levels have the same size (92x112) pixels. Some images were taken at different times, containing variations of lighting, facial expressions, and details of the face. Thumbnails of all of the images are shown in Fig.1.

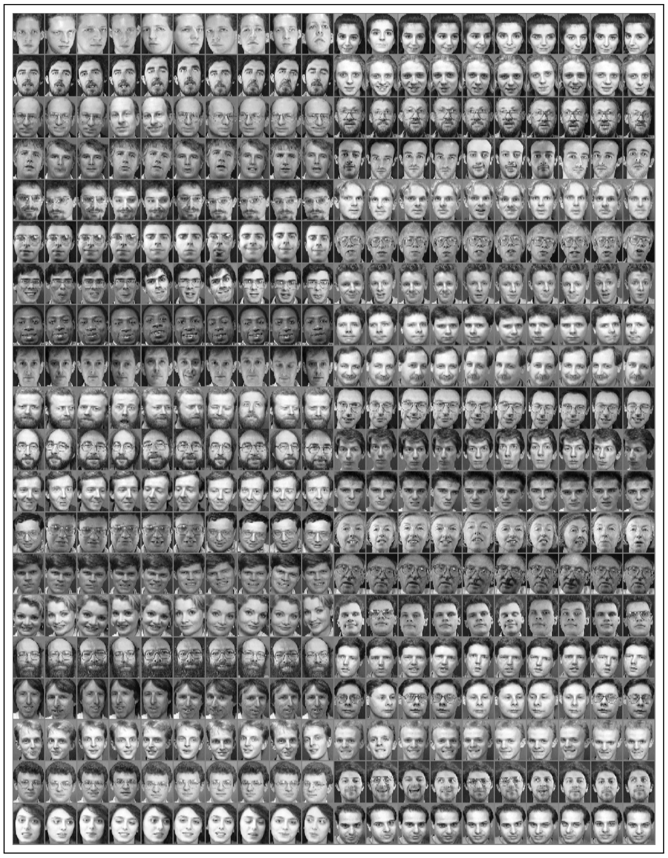

Fig. 1. ORL Database

V. Proposed Model

In this section a novel approach to face recognition is introduced. Our proposed method consists of combining the Co-occurrence Matrix and Bayesian Neural Networks for face recognition. Co-occurrence Matrix is used to feature extraction and then Bayesian Neural Networks are used to recognize the face via learning correct classification of the coefficients calculated by the Co-occurrence Matrix.

After reading the image that contains the face, and transforming it into a matrix A such that each component a_{ij} represents the gray level of the associated pixel in the image, the co-occurrence matrix associated B is calculated, then used to extract the relevant features $(x_1, x_2, ..., x_6)$, which are important for identification. which will subsequently be the input of Bayesian neural network trained using evidence framework.

The proposed procedure is summarized in Fig. 2

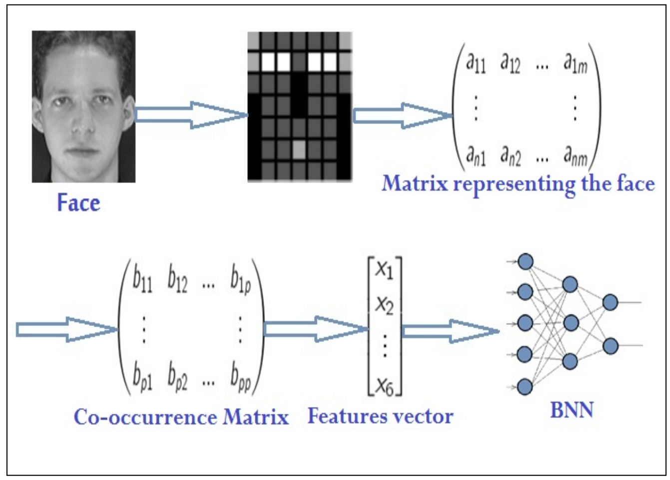

Fig. 2. Proposed Approach

VI. Experiments results

In this section, we evaluate the performance of our proposed approach on training bayesian neural network to recognize the ORL database.

Coding and Used Network Architecture

- The ORL database contains 40 different people, so we need 40 neurons in the output layers, in such a way that each neuron represents one person.
- The desired output associated with a vector of the characteristics of a face of the person i is the null vector except in the i th component contains 1.
- The size of feature vector is 6, so we need 6 neurons in the input layer, in such a way that each neuron represents one characteristic.

The used neural network is a MLP with 6 inputs, one hidden layer with 10 of hidden neurons and 40 output neurons. The nonlinear activation function used for the hidden units was the logistic sigmoid. For the output units, we used the SoftMax function. The architecture of the used neural network is represented in Fig. 3

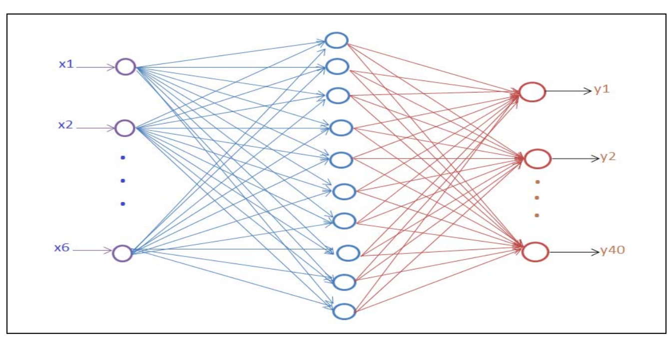

Fig. 3. Used Neural Network Architecture

Results

The problem of face recognition exists in the literature, but with different proposed models and different training methodology. We describe some of them and we compare it with our proposed approach:

- **FR-BP** This method allows training a fully connected MLP with Back-propagation algorithm such as the input of MLP in this way is an entire image after vectorization.
- **FR-EF** Consists to classify face images using a Bayesian neural network learned using Evidence Framework proposed by Mackay [3].

The obtained results are presented in the following table:

TABLE I
OBTAINED RESULTS

Aproach	Test Accuracy (%)
FR-BP	70.7
FR-EF	75.16
Proposed Aproach	81.23

As it is shown, experimental results show that combining the Co-occurrence Matrix and Bayesian Neural Networks could recognize the unseen face data with the highest classification accuracy of 81.23 in the last time in compares with other approaches. which shows the effectiveness of the proposed approach.

VII. Conclusion and future work

In this paper, we have proposed a new approach to face recognition based on co-occurrence Matrix and Bayesian Neural Networks. As a first step we introduced the co-occurrence Matrix and we calculed the relevant features which are important for identification. After we have presented a training model for bayesian neural networks nomed evidence framwork and based on re-estimating hyperparameters. Depending on the ORL database, which is widely used in face recognition, the results obtained demonstrates that our approach outperform the other methods. In the follow-up work, we will improve the evidence framework algorithm and integrate other Bayesian methods for training neural networks and try to analyze more datasets to expand its applicability.

References

[1] C.M. Bishop: Neural networks for pattern recognition, Oxford: Oxford University Press, (1995).

[2] C.M. Bishop: Pattern recognition and machine learning, springer, (2006).

[3] D.J.C. MacKay: A practical Bayesian framework for back-propagation networks, Neural Computation, vol.4, pp.448-472, (1992).

[4] D.J.C. MacKay: Bayesian interpolation, Neural Computation, vol.4, pp.415-447, (1992).

[5] H. Ramchoun and M. Ettaouil : New prior distribution for Bayesian neural network and learning via Hamiltonian Monte Carlo, Evolving Systems, pp.1-11, (2019).

[6] H. Ramchoun and M. Ettaouil : Hamiltonian Monte Carlo based on evidence framework for Bayesian learning to neural network, Soft Computing, pp. 4815-4825, (2019).

[7] Z. En-Naimani, M. Lazaar, and M. Ettaouil : Architecture optimization model for the probabilistic self-organizing maps and classification, 9th International Conference on Intelligent Systems: Theories and Applications (SITA-14). IEEE, (2014).

[8] F.D. Foresee and M.T. Hagan: Gauss-Newton Approximation to Bayesian Learning, Neural networks, Vol.3, pp.1930-1935 , (1997).

[9] M.M. Kasar, D.Bhattacharyya and T. Kim: Face Recognition Using Neural Network: A Review, International Journal of Security and Its Applications, Vol.10, pp.81-100, (2016).

[10] P. Lauret, E. Fock, R.N. Randrianarivony and J.F. Manicom-Ramassamy: Bayesian neural network approach to short time load forecasting, Energy conversion and management, Vol.49, pp.1156-1166, (2008).

[11] P. Lauret, M. David, E. Fock, A. Bastie and C. Riviere: Bayesian and sensitivity analysis approaches to modeling the direct solar irradiance. Journal of solar energy engineering, 128(3), pp. 394-405, (2006).

[12] R.M. Neal: Bayesian learning for neural networks, Springer Science and Business Media, (2012).

[13] S. Lawrence, C. L. Giles, A. C. Tsoi and A. D. Back: Face Recognition: A Convolutional Neural-Network Approach, IEEE TRANSACTIONS ON NEURAL NETWORKS, VOL. 8, NO. 1, (1997).

[14] N. Jamil, S. Iqbal and N. Iqbal: Face Recognition Using Neural Networks, Multi Topic Conference, 2001. IEEE INMIC 2001. Technology for the 21st Century. Proceedings. IEEE International. IEEE, (2001).

[15] W. Yandong, Z. Kaipeng, and L. Zhifeng: A discriminative feature learning approach for deep face recognition. In: European conference on computer vision. Springer, Cham, pp. 499-515, (2016).

[16] A.H. Bishak, Z.Ghandriz, and T.Taheri: Face Recognition using Co-occurrence Matrix of Local Average Binary Pattern (CMLABP), Journal of Selected Areas in Telecommunications, p. 15-19, (2012).

[17] F. Yuan, X. Xia, J. Shi: Mixed co-occurrence of local binary patterns and Hamming-distance-based local binary patterns. Information Sciences, vol. 460, p. 202-222, (2018).

[18] http://www.face-rec.org [accessed October 12, 2019]