Asian Journal of Information Technology 11 (6-12): 244-251, 2012
ISSN: 1682-3915

An Adaptive Recognition Technique Named SOMEF Based on Ear and Face Without Subject's Cooperation Using Neural Network Based Self Organizing Maps

[1]A.S. Raja and [2]V. JosephRaj
[1]Sathyabama University, Jeppiar Nagar, Chennai, Tamil Nadu, India
[2]Kamaraj College, Thoothukudi, Tamil Nadu, India

Abstract: Biometrics has been gaining attraction due to the ever-growing demand of this field of research on access control, public security, forensics and e-banking. However, there are still many challenging problems in improving the accuracy, robustness, efficiency and user-friendliness of these biometric systems. In this manuscript researchers propose a new adaptive multi-modal biometric framework based on self organizing maps for the recognition of individuals using face and ear. Researchers show that the proposed framework helps to improve the performance and robustness of recognition when compared to some standard methods in literature. The major focus of the approach is to keep the framework adaptive and robust, thereby, capable of being used in a wide variety of environments. Moreover, researchers also discuss some new directions on which SOM shall be effectively used in biometrics community. Researchers show all the findings with experimental results.

Key words: Biometrics, face, ear, Self Organizing Maps (SOM), public security

INTRODUCTION

The increase of terrorism and other kinds of criminal actions such as fraud in e-commerce, increased the interest for more powerful and reliable ways to recognize the identity of a person (Jain *et al*., 1999; Wechsler *et al*., 1997). To this end, the use of behavioral or physiological characteristics called biometrics is proposed. Biometrics is best defined as measurable physiological and or behavioral characteristics that can be utilized to verify the identity of an individual (Jain *et al*., 1999). Many physiological characteristics of humans, i.e., biometrics are typically invariant over time, easy to acquire and unique to each individual. Therefore, the biometrics traits are increasingly adopted for civilian applications and no longer confined for forensic identification.

The recognition of individuals without their full co-operation is in high demand by security and intelligence agencies requiring a Robust Person Identification System. Many face recognition algorithms have been proposed so far (Zhao *et al*., 2000; Turk and Pentland, 1991; Jain *et al*., 2008; Wiskott *et al*., 1997; Kotropoulos *et al*., 2000; Penev and Atick, 1996). Algorithms related to recognition of ear, hand geometry, iris, voice recognition have also been proposed (De Vel and Aeberhard, 1999). A Multimodal System is a combination of ear and face (for instance) or any other combination of biometrics. Multimodal biometrics can be used to overcome some of the limitations of a single biometric. For instance, it is estimated that 5% of the population does not have legible fingerprints (Jain *et al*., 1999) a voice could be altered by a Cold and Face Recognition Systems are susceptible to changes in ambient light and the pose of the subject.

A typical Biometric System usually consists of that specific biometric detection scheme followed by an extraction methodology (which shrinks the dimensionality of useful information) and then a classifier to make the appropriate decision (Chang *et al*., 2003; De Vel and Aeberhard, 1999) used PCA on face and ear with a manual land marking method. With a larger dataset of 111 subjects, they achieved a combined recognition rate of 90%. Mu *et al*. (2004) also used PCA for combining face and ear, they used profile images and manually extracted features. On a dataset of 18 subjects of profile face and ear, the recognition rate was 94.44%. Middendorff and Bowyer and Hurley *et al*. (2005) used PCA/ICP for face/ear, manually annotating feature landmarks. On a 411 subject dataset they were able to achieve a best fusion rate of 97.8%. Chang *et al*. (2003) used FSLDA (full-space linear discriminant analysis) algorithm on 75 subject database with 4 images each (USTB) and on the ORL database of 75 subjects, achieving a best recognition rate of 98.7%. Despite of all these advancements when it comes to practical usage in real life environment, there have been issues. Therefore, an adaptive framework

Corresponding Author: A.S. Raja, Sathyabama University, Jeppiar Nagar, Chennai, Tamil Nadu, India

which shall research in all scenarios is very much required. Here, researchers provide an adaptive framework for appearance-based multi-modal recognition based on self organizing maps. This framework shall be easily extended to address very interesting questions faced by the biometrics community.

OBJECT DETECTION

Researchers extract the regions of interest using a Haar like features based object detector provided by the open source project Open CV library. This form of detection system is based on the detection of features that display information about a certain object class to be detected. Haar like features encode the oriented regions whenever they are found, they are calculated similarly to the coefficients in Haar wavelet transformations. These features can be used to detect objects, in this case the human ear and human face. The Haar like object detector was originally proposed by Viola and Jones (2001) and later extended by Lienhart and Maydt (2002).

Researchers used the dataset related to ear and face retrieved from IIT Delhi Ear Database Source Link (http://www4.comp.polyu.edu.hk/~csajaykr/IITD/Database_Ear.htm) and The Yale Face Database Source Link (http://cvc.yale.edu/projects/yalefaces/yalefaces.html Fischer, 2001) for the experiments (multimodal datasets are not readily available as a standard in literature. The dataset that we have used is a virtual database for the multimodal study. The reason for referring this as virtual is because of the fact that the original dataset does not contain the combined face and ear of every individuals. Instead, one dataset (IIT Delhi Ear Database Source Link, http://www4.comp.polyu.edu.hk/~csajaykr/IITD/Database_Ear.htm) had the ear images and the other dataset (The Yale Face Database Source Link, (http://cvc.yale.edu/projects/yalefaces/yalefaces.html) had the face images of individuals separately. Researchers had combined them by taking 3 face images of an individual from the face dataset with 3 ear images of an individual from the ear dataset. There were 107 ear subjects and 165 face subjects. Researchers had considered all the 107 ear subjects and we had chosen randomly 107 subjects from the face subjects). There are 107 subjects in this dataset. To create the ear detector and face detector researchers used several sample images. The positive images were scaled to a size of 60×60 for face and to a size of 18×65 to reflect the rectangular dimensions of the ear. The face detector and ear detector worked well with a few falsely detected face and ears, the problem was overcome by selecting the larger detected object. Some of the sample face and ear images detected from the database is mentioned in Fig. 1.

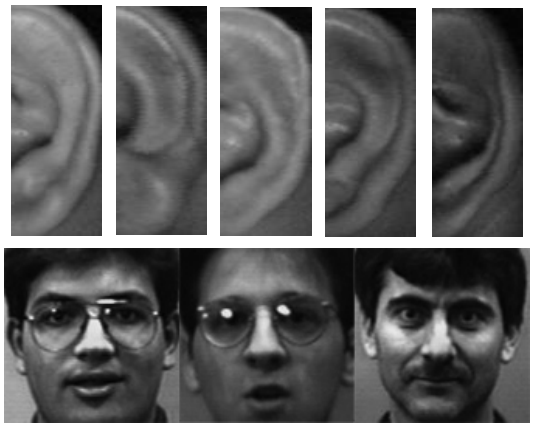

Fig. 1: Example images from the datasets (IIT Delhi Ear Database Source Link, http://www4.comp.polyu.edu.hk/~csajaykr/IITD/Database_Ear.htm; The Yale Face Database Source Link, (http://cvc.yale.edu/projects/yalefaces/yalefaces.html)

SOM AND METHODOLOGY USED IN THIS STUDY

Self Organizing Map (SOM) is a special kind of unsupervised computational neural network (Fischer, 2001) that combines both data projection (reduction of the number of attributes or dimensions of the data vectors) and quantization or clustering (reduction of the number of input vectors) of the input space without loss of useful information and the preservation of topological relationships in the output space (Fig. 2).

A few concepts are useful to understand the workings of the technique. The input space (also called signal) is the set of input data researchers employ to feed the algorithm; the set of input data in the case refers to the set of images that researchers use for training typically, the observations are multidimensional and are thus expressed by using a vector for each of them. In the case the observations refer to the pixels present in each image (in the case the dimension/vector size of each ear image is 18×65 = 1170 and face image is 60×60 = 3600). On the contrary, the output space (trained network, network or SOM) refers to the low-dimensional universe in which the algorithm represents the input data. It usually has two-dimensions and is composed of a set of elements called neurons (or nodes) which are interconnected, hence the network. What the algorithm does is to represent the input space onto the output space, keeping all the relevant information and ordering observations in a way such that topological closeness in the output space implies statistical similarity in the input space.

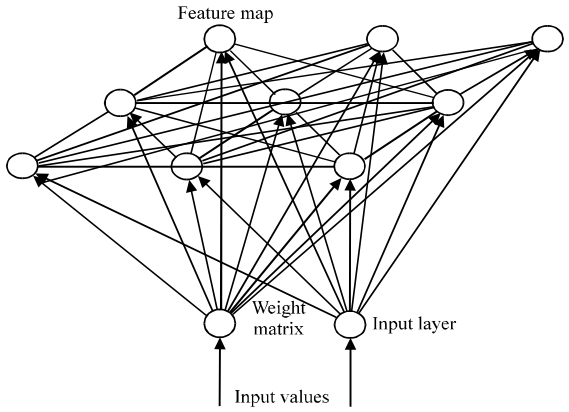

Fig. 2: Self Organizing Map (SOM)

The input space is composed of n-dimension vectors researchers want to visualize/cluster in a low-dimensional environment. Researchers can express the input vector t as:

$$x = [\xi 1(t), \xi 2(t), \ldots\ldots, \xi n(t)]T \in R^n$$

where, $\xi i(t)$ represents the value for each dimension. The output space is an array of p by q neurons (nodes) topologically connected following a kind of geometrical rule (the most common topologies being circles, squares and hexagons). In the case p = 11 and q = 11. Each of the nodes is assigned a parametric real vector of initially random values that researchers call model and express as:

$$mi = [\mu in, \mu in, \ldots\ldots, \mu in]T \in R^n$$

Last, researchers may also define as d (x, mi) any distance metric between two vectors x and mi. The most widely used is the Euclidean distance although, other specifications are also valid.

What we are looking for is a topologically-ordered representation of the signal space into the network. That is done by the SOM in an iterative process called training in which each signal vector is sequentially presented to the output space. The best matching unit (b.m.u.) for x is defined as the neuron minimizing the distance to x. When this is found, the b.m.u. is activated and an adaptive process starts by which such neuron and its topological neighbours are modified by the following scheme:

$$mi(t+1) = mi(t) + hci(t)[x(t) - mi(t)]$$

where, t and t+1 represent, respectively the initial and the final state after the signal has activated the neuron hci (t) is called neighbourhood function and expresses how the b.m.u. and its neighbours are modified when activated by a signal usually, the linear or Gaussian versions are used. This process is repeated over many cycles before the training is finished. The neighbourhood function depends on several parameters relevant for this stage: the distance between the b.m.u. and the modified neuron (so the further away the neuron is the smaller the adjustment); a learning rate α (t) that defines the magnitude of the adjustment and gradually decreases as the training cycles advance and the neighbourhood radius which decides which of the surrounding neurons of the b.m.u. are also modified and also decreases over the training stage and the self arranging (organization) of the input observations.

This procedure may be used as a visualization tool for multidimensional datasets as well as a clustering method. In the first case, researchers would want to see how the different observations are mapped into the SOM to discover (dis) similarities, making use of the topological preservation of the statistical characteristics and study how the different dimensions are distributed in the second one, the network would have a relatively small number of neurons (as many as clusters researchers would want to obtain) and researchers would focus on analyzing which observations are grouped with which. In the case, images which have similar ear characteristics gets grouped together within the respective nodes/maps.

The description of SOM given above (also referred as unsupervised SOM in some literature) focuses on unsupervised exploratory analysis. However, SOMs can be used as supervised pattern recognizers, too. This means that additional information, e.g., class information is available that can be modeled as a dependent variable for which predictions can be obtained. The original data are often indicated with X the additional information with Y. An approach suggested by Kohonen for supervised SOM is to perform SOM training on the concatenation of the X and Y matrices.

Although, this research in the more simple cases, it can be hard to find a suitable scaling so that X and Y both contribute to the similarities that are calculated. Melssen proposed a more flexible approach where distances in X and Y-space are calculated separately. Both are scaled so that the maximal distance equals 1 and the overall distance is a weighted sum of both:

$$D(o,u) = Dx(o,u) + (1-\alpha)\, Dy(o,u)$$

where, D (o, u) indicates the combined distance of an object o to unit u and Dx and Dy indicate the distances in the individual spaces. Choosing α = 0.5 leads to equal

weights for both X and Y spaces. Scaling so that the maximum distances in X and Y spaces equal one takes care of possible differences in units between X and Y. Training the map is done as usual the winning unit and its neighborhood are updated and during training the learning rate and the size of the neighborhood are decreased.

One shall extend the principle used for supervised SOM to more than one layer as well the result of which is being referred in literature as super-organized SOM. This is the idea which is used in this framework. For every layer a similarity value is calculated and all individual similarities then are combined into one value that is used to determine the winning unit:

$$D(o,u) = \sum_i \alpha_i D_i(o, u)$$

where the weights i are scaled to unit sum. These weights are the only extra parameters (compared to classical SOMs) that need to be provided by us. The super-organized map accounts for individual types by using a separate layer for every type. When compared to other neural network based approaches, it shall be noted that in SOM, the neurons are arranged on a flat grid not as a multilayer perceptron (input, hidden and output).

PROPOSED APPROACH (SOMEF)

Researchers use SOM for ear and face recognition and hence researchers call the approach as SOMEF. The set of input data in the case refers to the set of images that is used the observations refer to the pixels present in each image. First researchers apply SOM to face and ear separately. In this case for face, the dimensionality of the input vector is 3600 (this is because of the normalized size of the face image that is used 60×60 size) and in the case of ear the dimensionality of the input vector if 1170 (this is because of the normalized size of the ear image that is used -18×65 size). The output space is an array (separate for both ear and face) of p by q neurons (nodes) topologically connected following a kind of geometrical rule (a rectangular topology has been used). In the case p = 11 and q = 11 for face and p =11 and q = 11 for ear. With the same setup, researchers do a supervised mode SOM analysis (where researchers use some images for training and some images for testing). In the end (SOMEF approach), researchers combine both the face and ear using super organized SOM. In other words, in SOMEF researchers get multiple layers (as opposed to supervised SOM where there are only two layers X and Y). Researchers play with the weight between face and ear layers and determine the optimum weightage for the recognition experiment under consideration. All these interesting experimental results obtained using SOM in unsupervised mode, supervised mode, super organized mode.

EXPERIMENTS AND RESULTS

As mentioned earlier, this study uses the Face and Ear dataset obtained from IIT Delhi Ear Database source link (http://www4.comp.polyu.edu.hk/~csajaykr/IITD/Database_Ear.htm) and The Yale Face Database Source Link (http://cvc.yale.edu/projects/yalefaces/yalefaces.html Fischer a, 2001). There are 107 subjects. Each subject has 3 images each for ear and 3 images for face.

The first experiment which was performed was to find the total number of output nodes which are required. Unsupervised SOM was ran over the given 107 subjects related to face and ear dataset. In the plot shown in Fig. 3, the background color of a unit corresponds to the number

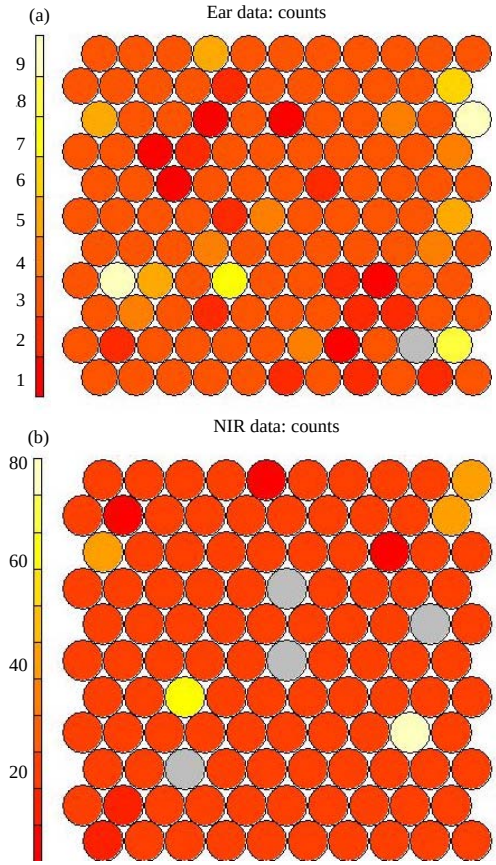

Fig. 3: Counts plot of the map obtained from the ear and the face dataset. Empty units are depicted in gray. The color in each cell represents the number of; a) ear and b) face images which went into that went into that cell

of samples mapped to that particular unit; one shall observe that they are reasonably spread out over the map (one unit is empty for ear; four units are empty for face; no samples have been mapped to them). The plot in Fig. 4 shows the mean distance of objects, mapped to a particular unit, to the codebook vector of that unit. A good mapping should show small distances everywhere in the map. These show that the number of output nodes which are chosen (11×11) are good enough for the purpose.

The second experiment which was performed was to do an exploratory analysis using unsupervised SOM. Figure 5a and b shows the mapping of images related to unsupervised SOM. Each color/shape in the figure is used to represent a particular subject. From the dataset, one shall infer that each subject has 3 ear images and 3 face images related to him which are more or less mapped into different unique cells. Figure 5 reveals this out clearly. For instance in Fig. 5, if one looks at the first cell, approximately 3 similar units are mapped onto that cell for ear and 3 similar units for face. The similar units indicate

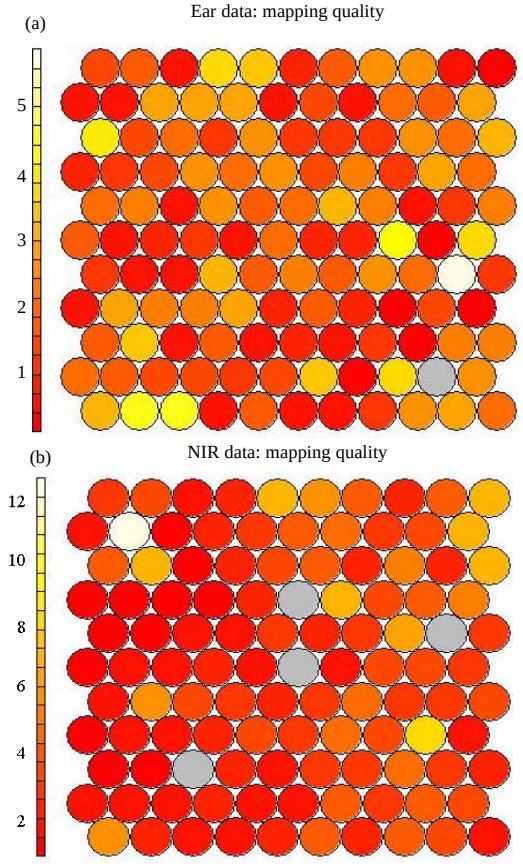

Fig. 4: The quality of the mapping; the biggest distance between x and mi vectors are found in the bottom left of the map for; a) Ear and b) Face

that they belong to the same subject. This explains that even without any training, unsupervised SOM was able to more or less grossly able to put the subjects into different cells. The error rate in grouping in this case was observed to be approximately 22% for ear (out of the 321 images of 107 subjects, 251 went into the appropriate cells which belonged to similar subjects and 70 images did not gets mapped properly) and 27% for face (out of the 321 images of 107 subjects, 234 went into the appropriate cells which belonged to similar subjects and 87 images did not gets mapped properly).

The third experiment that was used is to use the classifier information related to which image belonged to which subject using supervised SOM. In this experiment, the subject has been considered as the dependent variable (variable Y) and the pixel values of the image as the independent value (variable X). The 1 random image from each subject has been chosen for training and the rest of the 2 images of each subject has been used for testing. The weights for X and Y has been varied with supervised SOM and the following characteristics as mentioned in Table 1 has been observed (the weights in a way indicate the relative strength between X and Y for recognizing a subject).

The earlier Table 1 shows that if one uses the classification information also (using supervised SOM), then the recognition rate improves significantly (when compared to not using it as earlier seen with unsupervised SOM). This is true both for face and ear. The fourth experiment that was conducted was to vary the number of images used for training and testing. For this experiment, the weight of X has been chosen as 0.1 and

Table 1: Error rate with supervised SOM by varying X and Y weights for ear and weights for face

X weightage	Y weightage	Error rate
Weights for ear		
0.9	0.1	19.7
0.8	0.9	15.8
0.7	0.3	14.2
0.6	0.4	13.1
0.5	0.5	11.2
0.4	0.6	10.4
0.3	0.7	8.2
0.2	0.8	5.9
0.1	0.9	5.3
Weights for face		
0.9	0.1	23.5
0.8	0.9	19.8
0.7	0.3	17.2
0.6	0.4	16.6
0.5	0.5	14.7
0.4	0.6	12.9
0.3	0.7	11.2
0.2	0.8	8.6
0.1	0.9	7.9

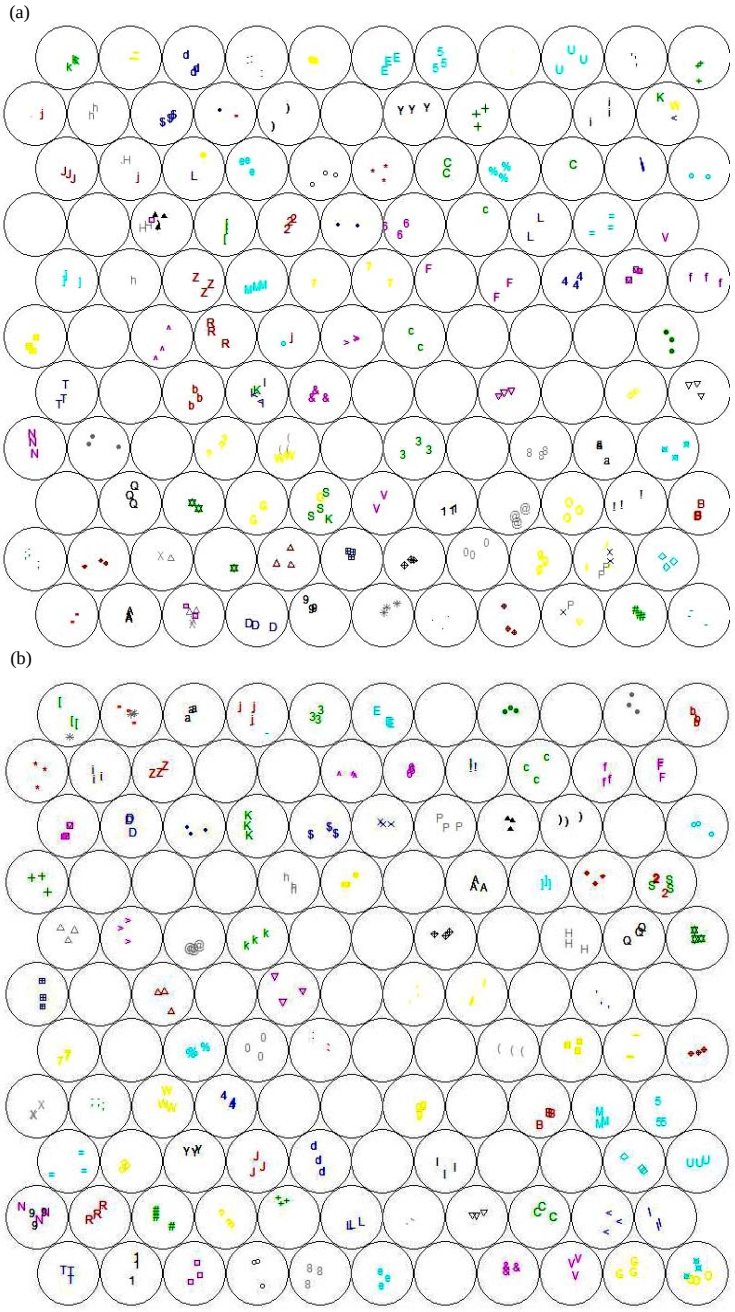

Fig. 5: a) Mapping of the 107 Ear subjects in a eleven by eleven SOM and b) Mapping of the 107 Face subjects in a eleven by eleven SOM

weight of Y has been chosen as 0.9. Table 2 shows the inferences of this experiment for face and ear.

The fifth experiment that was done was related to super-organized SOM. Researchers modeled the face related pixel values as one layer, ear related pixel values as the second layer and the class information as the third layer. A weight is associated to every layer to be able

Table 2: Error rate with supervised SOM by varying the number of images for training and testing with ear and testing with face

No. of images for training	No. of images for testing	Error rate
Testing with ear		
1	2	5.3
2	1	3.4
Testing with face		
1	2	7.9
2	1	5.4

Table 3: Recognition rates for different face/ear weights using super SOM

Ear weight age	Face weight age	Combined recognition error rate
0.9	0.1	5.1
0.8	0.9	4.8
0.7	0.3	3.9
0.6	0.4	2.4
0.5	0.5	4.7
0.4	0.6	4.9
0.3	0.7	5.2
0.2	0.8	5.9
0.1	0.9	6.4

Table 4: Test error recognition rate (%) with varying number of images per person

		Training phase			Testing phase		
No. of training images per person	No. of testing images per person	PCA	SFFS	SOMEF	PCA	SFFS	SOMEF
1	2	0	0	3.6	10.2	8.3	2.4
2	1	0	0	3.9	8.5	7.8	1.6

to define an overall distance of an object to a unit. Researchers pose an optimization problem to optimize the weights in such a way that the recognition rate is the maximum. Interestingly, this also allows one to easily find out the dominant metric (face or ear-based on the one which takes a higher weightage). To begin with researchers seeded the initial weights to be of extremely high (1) forface and extremely low (0) for ear. Researchers noted down the results. Researchers then optimized the weights for face and ear as explained above and observed the results. The experimental results are presented in Table 3. It seems that ear is a better metric when compared to face for the given standard dataset and a propositional weight of 6:4 seems to give a better recognition rate.

The sixth experiment that was done was a comparative analysis of SOMEF with other methods related to multimodal biometrics involving face and ear. Table 4 shows the comparative results between Face-Ear-PCA (Principal Component Analysis), Face-Ear-Sequential Float Feature Selection (SFFS) and Self Organizing Map for Ear and Face (SOMEF).

The size of the training set varied from 1-2 images per person and the remaining of the images for each subject form the test set. For the PCA and the SFFS, the experiments that were conducted showed that all the training images during the training phase are classified correctly (Table 4). On the other ear, the SOMEF could not classify correctly all the training images. Furthermore, Table 4 shows a greater improvement in the performed experiment with SOMEF than PCA or SFFS when using one number of training sample for each person. Using SOMEF with one image per person during training phase gives 1.6% error recognition rate (incorrectly classified 5 images of 214 test images) against 10.2% error recognition rate (incorrectly classified 32 images of 214 test images) using the PCA and 8.3% error recognition rate (incorrectly classified 26 images of 214 test images) using the SFFS Method.

Table 4 shows that SOMEF can provide an improvement in error recognition rate when compared to the other approaches based on literature. Interestingly, self organizing maps shall also be used to address some interesting curiosities discussed in the biometrics community in a formal manner. For instance, there has been a curiosity/hypothesis which says that ear as a biometric does not change over age when compared to other biometrics like face. If one shall gather images of same subjects at different ages in a similar pose and background and do a supervised SOM across the different ages, one shall find out if ear has been consistently performing when compared to face or some other biometric. Most of the results used in this study are obtained using an opensource software framework named statistical R (R Project http://www.r-project.org/). The archive of results and code used related to this study is accessible at Archive of software code, dataset and experimental results (http://www.capeitech.org/research/asr/somef/).

CONCLUSION

Neural network based self organizing maps has been used in this study. The proposed approach SOMEF has been shown to perform well when compared to some standard methods from literature. This has been done by taking a standard dataset from literature.

REFERENCES

Chang, K., K.W. Bowyer, S. Sarkar and B. Victor, 2003. Comparison and combination of ear and face images in appearance-based biometrics. IEEE. Trans. Patt. Anal. Mach. Intel., 25: 1160-1165.

De Vel, O. and S. Aeberhard, 1999. Line-based face recognition under varying pose. IEEE Trans. Pattern Anal. Mach. Intell., 21: 1081-1088.

Fischer, M.M., 2001. Computational Neural Networks: Tools for Spatial Data Analysis. In: GeoComputational Modelling: Techniques and Applications, M.M. Fischer and Y. Leung (Eds.). Springer, Heidelberg, Germany, pp: 79-102.

Hurley, D.J., M.S. Nixon and J.N. Carter, 2005. Force field energy functionals for ear biometrics. Comput. Vision Image Understanding, 98: 491-512.

Jain, A.K., P. Flynn and A. Ross, 2008. Handbook of Biometrics. Springer, USA., ISBN: 9780387710419, pp:1-22.

Jain, A.K., R. Bolle and S. Pankanti, 1999. Biometrics: Personal Identification in Networked Society. Kluwer Academic Publishers, Boston/Dordrecht/London.

Kotropoulos, C.L., A. Tefas and I. Pitas, 2000. Morphological elastic graph matching applied to frontal face authentication under well-controlled and real conditions. Pattern Recogn., 33: 1935-1947.

Lienhart, R. and J. Maydt, 2002. An extended set of Haar-like features for rapid object detection. Proceedings of the International Conference on Image Processing, September 22-25, 2002, New York, USA., pp: 900-903.

Mu, Z., L. Yuan, Z. Xu, D. Xi and S. Qi, 2004. Shape and structural feature based ear recognition. Proceedings of the 5th Chinese Conference on Biometric Recognition, December 13-14, 2004, Guangzhou, China, pp: 663-670.

Penev, P.S. and J. Atick, 1996. Local feature analysis: A general statistical theory for object representation. Network Comp. Neural Syst., 7: 477-500.

Turk, M. and A. Pentland, 1991. Eigenfaces for face recognition. J. Cognitive Neurosci., 3: 71-86.

Viola, P. and M. Jones, 2001. Rapid object detection using a boosted cascade of simple features. Proceedings of the Computer Society Conference on Computer Vision and Pattern Recognition, Volume 1, December 8-14, 2001, Kauai, HI., USA., pp: 511-518.

Wechsler, H., J.P. Phillips, V. Bruce, Folgeman, F. Soulie and T.S. Huang, 1997. Face Recognition: From Theory to Applications. Springer, Berlin, Heidelberg, New York.

Wiskott, L., J.M. Fellous, N. Kruger and C. von der Malsburg, 1997. Face recognition by elastic bunch graph matching. IEEE Trans. Pattern Anal. Mach. Intell., 19: 775-779.

Zhao, W.Y., R. Chellappa, A. Rosenfeld and P.J. Philips, 2000. Face recognition: A literature survey. UMD CfAR Technical Report CAR-TR-948.