Proceedings of International Conference
on Artificial Intelligence

Facial Image Recognition Based on Connectionist Model and Coarse-to-Fine Pyramidal Decision Strategy[1]

Jia-Lin Chen * and Kun-Cheng Tsai
Department of Computer Science, Chung-Hua Polytechnic Institute
Hsinchu, Taiwan
* E-mail: jlchen@chpi.edu.tw

Abstract

In this paper, we propose a novel approach for facial image recognition. Coarse and fine resolutions respectively provide the global and locally detailed information; and we use one SOFM for automatically learning facial features for one resolution. A particular classification strategy, called the coarse-to-fine pyramidal decision strategy, is proposed to hierarchically integrate the recognition decision of each resolution with the order from the coarsest resolution to the finest. Potential candidates are sifted as the resolution increases, and fewer and fewer candidates are survived. The experimental results show 100% accuracy is achieved.

Keywords: *face recognition, self-organized feature map, coarse-to-fine pyramidal decision strategy, multi-resolution, wavelet transform.*

1. Introduction

The problem of facial image analysis is a popular research topic in many disciplines [16]. The related research topics to which have been paid attentions include face modeling, face recognition/categorization, facial expression understanding and so on [16]. Up-to-date, no complete solutions for all face related problems are present yet. In this paper, we are specifically interested in the problem of face recognition.

Geometrical feature extraction and template matching based approaches are two major paradigms for the facial image recognition. The feature based paradigm locates the structural features on a face such as eyes, eyebrow, nose, mouth and chin from the full facial view or silhouette profile [1,2,4,11,15]. The geometric properties such as shapes, positions and distances, or statistical properties of these features can then be used for the recognition purpose. The template based paradigm is based on the defined templates which may cover particular structural features such as eyes, noses and so on [1,11,17]. For some approaches, the selections of templates are based on the component analysis such as K-L procedure [5] and eigenfaces [12]. Distances and correlation are common metrics of measuring the similarity or matching. The detailed discussion and review of feature- and template-based paradigms can be found in [1,11]. The neural network based paradigm is also applicable to the face processing problems [8,13]. One advantage of using neural networks is that the facial information is automatically exploited statistically; therefore, the difficult problems of locating and computing facial features are avoided. A complete survey of neural network models on face processing is in [13].

In the paradigms described above, the face recognition problem is transformed to a traditional pattern recognition problem (feature based, template-matching based, or neural network based), where the clustering of feature vectors, similarity measure of matching and trained weights in neural networks respectively play very crucial roles on the recognition performance. However, computing the feature vectors or similarity measures obviously depends strongly on how the structural features of the face are defined and located [1]. On the other hand, the complicated details of numerous faces are usually hard to learned effectively in a simple neural network. Partitioning a face into subimages for less facial details and using more networks is an alternative [8]. Subsequently, some heuristic rules are usually used *a priori* such as the locations and shapes of structural

[1] This work was supported in part by National Science Council, Republic of China under Grant NSC86-2213-E-216-019.

features to simplify the system complexity

In this paper, we propose a novel approach based on various viewing distances, which mimics a human behavior on recognizing objects. No any *a priori* information regarding facial features is required in the proposed approach. A facial image is decomposed into subimages of various resolutions corresponding to gazing this face at various distances. A facial image of fine resolution corresponds to gazing a person at a short distance; and the one of coarser resolution corresponds to gazing a person at a longer distance. An image of coarse resolution gives the global information such as shape of face, and that of a fine resolution gives detailed and local information. This statement is illustrated in fig. 1.1. Subsequently, different modeling approaches or statistical properties should be applied to or exploited from different resolutions for their different extents of modeling difficulty. In our scheme, various statistical properties of faces associated with various resolutions are exploited via the self-organized feature map (SOFM) neural networks [3,6,8]. For learning facial features for individual resolution, one SOFM is created for one resolution. If there are L resolutions, there are L associated SOFM's.

It is quite straightforward to realize that faces with obviously distinct features observed at a long distance would not look similar at a short distance. Namely, only those looked similar at a long distance need be moved to a shorter distance for a detailed examination. A particular classification strategy, coarse-to-fine pyramidal decision strategy, is proposed for this scenario accordingly. This novel decision strategy hierarchically integrates the decision of each resolution with the order from the coarsest resolution to the finest. As the resolution increases, potential candidates are sifted among the survived potential candidates from the previous resolution. These potential candidates are those who need be examined at the next fine resolution. Fewer and fewer candidates are survived as the resolution increases. The procedure stops when a unique candidate is achieved, which means no ambiguous candidates resulted, or the finest resolution is reached and the most likely candidate is selected. It is our specialty that the fuzzy decisions are allowed at each resolution due to the imperfect discriminative capability of statistical properties exploited; and the sifting procedure along the increasing resolutions constitutes a hierarchical classification based on the various statistical properties of a face.

This paper is organized as follows. Introduction is given in section 1. Section 2 describes the multi-resolution representation of faces. Sections 3 and 4 respectively describe the learning of facial features by SOFM neural networks and the proposed coarse-to-fine pyramidal sifting strategy. Experiments and conclusions are given in sections 5 and 6.

2. Multi-resolution representation of faces

The multi-resolution representation of a facial image is a natural representation in terms of viewing distance. The fine and coarse resolutions correspond to gazing a face at short and long distances, respectively. There are many techniques applicable to the multi-resolution decomposition purpose such as pyramidal Gaussian filtering [10], orthogonal and non-othogonal wavelet transforms [7,9]. The use of multi-resolution decomposition is motivated by the psychovisual study to mimic the visual channels of human beings. Therefore, this technique is quite commonly applied to the computer vision applications such as image coding [14], texture analysis [7] and so on [9]. In addition to the advantage of psychovisual background, the orthogonal wavelet transform is further with the advantage of perfect reconstruction from the decomposed subimages. Therefore, no information is lost or added during the transformation; and the total pixels of subimages are equal to those of the original image. The coarser is the resolution, the less size of subimage is. Subsequently, fast analysis can be achieved at the coarse resolution. Detailed discussion of the multi-resolution decomposition and its applications can be found in [9,10].

In this paper, we employ a quadrature mirror filter (QMF) bank to implement the orthogonal wavelet transform. The details of QMF bank can be found in [7,14]. Three layers of decomposition is used in our scheme as shown in fig. 2.1. The subimages in the high frequency subbands are not utilized here since they mostly provide edge or structure information which are not very discriminative while used alone. Subsequently, we only use the subimages of low frequencies.

3. Learning by self-organized feature map neural networks

The self-organized feature map (SOFM) neural network is a network for unsupervised learning; and may be used by itself or as a layer of another neural network [3,6,8]. We adopt the SOFM neural network in our proposed scheme for the following reasons:

1. The learning of SOFM neural network is unsupervised so as to reflect the clustering property of inputs. No feature extraction procedure is required *a priori*. For the face recognition problem, it is well known that the discriminative features are difficult to define or locate accurately. By using the SOFM, our proposed scheme can avoid the problems caused by improper or insufficient feature extraction techniques.
2. The SOFM neural networks use competitive learning to create low-dimensional topographic maps of higher

dimensional input data. The neurons can effectively infer the relationships among training patterns and map these patterns in the observation space onto the spatial extent of the map. Subsequently, a discriminative space is available for easier classification, which is beneficial to complicated patterns such as faces.

3.1 Self-organized feature map neural networks

The SOFM proposed by Kohonen is a computationally less intensive method for producing topographical ordering by adapting the synaptic weights of the neighbors of the firing neuron as well as those of the winning neuron itself [6]. The structure of a two dimensional SOFM neural network is shown in fig. 3.1, and its learning procedure is described as follows.

1. Initialize the weights of each neuron $\underline{w}_{ij}^k$, the adaptation neighborhood size $R^k = R_{\max}$, and the adaptation coefficient $\eta^k = \eta_{\max}$, where $k=0$.
2. Compute the winning neuron at iteration k, which is with minimum Euclidean distance to input vector $\underline{x}$, i.e., $y(i^*, j^*) = \min_{\text{all } i's, j's} \left\| \underline{x} - \underline{w}_{ij}^k \right\|$.
3. Updated the weights of neurons located at (i, j)

$$\underline{w}_{ij}^{k+1} = \begin{cases} \underline{w}_{ij}^k + \Delta \underline{w}_{ij}^k & \text{if } d(i,j) < R^k, \\ \underline{w}_{ij}^k & \text{otherwise} \end{cases} \quad (1)$$

where $\Delta \underline{W}_{ij}^k = \eta^k e^{-d/R^k} (\underline{x} - \underline{W}_{ij}^k)$, (2)

and $d(i,j) = \sqrt{(i-i^*)^2 + (j-j^*)^2}$

4. Decrease the size of adaptation neighborhood, $R^{k+1} = \alpha(R^k)$, and the adaptation coefficient, $\eta^{k+1} = \beta(\eta^k)$, where α and β are monotonically decreasing functions.
5. Stop if R^k and η^k reach their lower bounds or the learning cycle counter k reaches a predefined value T; otherwise $k=k+1$ and goto step 2.

The result is a topographically ordered map of neurons adjusted to the training patterns. In case that two inputs $\underline{x}_1$ and $\underline{x}_2$ are made more and more similar, $\underline{y}_1$ and $\underline{y}_2$, the corresponding spatial locations of maximum responses in the network, should get closer and closer and eventually coinciding. The learning procedure for higher dimension SOFM can be extended thereof.

3.2 Learning of facial features by SOFM's

Our proposed method uses two-dimensional map of identical neurons. The feature map possesses $P \times Q$ neurons and each neuron possesses $N \times M$ synaptic weights, which matches the order of the input vector. The input vector is the image intensity arranged in a column vector. The response of the network to an input vector is the two-dimensional pattern of excitation exhibited by the array of neurons. The response of each neuron is its level of correlation between the input vector and the synaptic weights of that neuron. There will be only one peak in this response function. In our problem, each neuron will adapt to represent a face. Consequently, similar faces will be located closely in the map and reflect themselves as a cluster; on the other hand, distinctive faces will be located distantly as distinct clusters.

Labeling the winning neuron

In our scheme, the SOFM itself is used for both learning and recognition. However, the learning procedure of SOFM is unsupervised which does not provide labels for the resulting clusters. Subsequently, labeling clusters is necessary for our recognition purpose. In our case, there is only one face used for training each person; therefore, the labeling procedure becomes very straightforward. The total amount of winning neurons is equal to that of N training faces or persons. What we need to do is re-input the training faces to the trained SOFM, and the corresponding winning neurons identify the faces.

4. Coarse-to-fine pyramidal sifting strategy

The proposed sifting strategy hierarchically integrates the decision of each resolution with the order from the coarsest resolution to the finest as shown in fig 4.1. For each resolution, potential candidates are selected from the survived candidates of previous resolution. Fewer and fewer candidates are sifted as the resolutions increase. The procedure stops when one candidate is left, or the finest resolution is reached. In the latter case, the most likely one will be selected.

4.1 Classification for one resolution

On one SOFM, the distance of the winning neuron of a tested facial image to the representative one of a trained face is used to compute the candidate membership, *CM*. The *CM* shows the likelihood of the tested image belonging to that trained face at that resolution. The larger is the *CM*, the more similar the tested facial image is to that trained face. Consequently, potential candidates are elected according to their associated *CM*'s.

For one tested facial image at a particular resolution γ, we compute the winning node (x^*, y^*) on its associated SOFM, and its Euclidean distance to every labeled winning node (x_i^*, y_i^*), denoted d_i^γ. The candidate membership of this tested image to a trained face i at resolution γ is defined as

$$CM_i^\gamma = \frac{m_i^\gamma}{m^\gamma} \tag{3}$$

where $$m_i^\gamma = \frac{1}{d_i^\gamma + \varepsilon} \tag{4}$$

$$d_i^\gamma = \sqrt{(X^* - X_i^*)^2 + (Y^* - Y_i^*)^2} \tag{5}$$

and $$m^\gamma = \max_{i=1,2,\ldots,N} m_i^\gamma \tag{6}$$

A threshold, CMT^γ, is set for selecting the potential candidates at resolution γ. In case that CM_i^γ is greater than CMT^γ for a tested face against a trained person i, this face is likely belonged to that person. The large value of CMT^γ means strictly likely candidates are allowed; and CMT^γ may vary from resolution to resolution. Less potential candidates being survived means a more discriminative classification is employed at this resolution. It should be noted that the topographical ordering of winning neurons of training faces may vary at various resolutions. This statement is applicable because the distinctions among statistical properties of faces at various resolutions may differ. The outputs of one SOFM is the survived potential candidates who will be examined at the next fine resolution.

4.2 Decision based on coarse-to-fine sifting

The final decision is made by integrating the individual decision from each resolution starting from the most coarsely resolution and increasing the resolution gradually. The procedure can be described as follows:

Step 1: Initialize with the coarsest resolution ($\gamma = 1$). All trained faces are potential candidates.

Step 2: For a tested facial image at resolution γ, compute CM_i^γ against all the potential candidates, and compare with the corresponding CMT^γ.

Step 3: If number of survived candidates is 1, stop; or if the resolution $\gamma = L$, the finest resolution, choose the neuron with the largest CM_i^γ, and stop; otherwise, increase the resolution to another finer level ($\gamma = \gamma + 1$) and goto step 2.

It should be noted that this proposed pyramidal sifting rule need not consider all trained persons at any resolution, except at the coarsest one. Only the survived candidates from the previous resolution will be considered at the next resolution, which subsequently results in a saving of computation and increase of decision accuracy.

5. Experiments

Ten persons with equal sexuality are selected for the experiment, where one is kid of age around 10 and the others are of ages around 24 to 30. As shown in fig. 5.1, five front view images are taken for each person: one is 'normal' with no expression (for training) and the others are expressive (for testing). The images are of size 128 × 128 and have been taken carefully so that significant facial features from forehead to chin are covered. Besides, neglecting the effect of skin color, we normalize the intensities of images to be with equal energy.

The 'normal' images are used for training, i.e., one image trained for one person. This assumption is reasonable for some practical applications where images with variations such as expressions for an ideal training purpose are not easily available. We do not take the oriented faces into account because usually this problem can be solved if more images with various orientations are used for training. The size of one SOFM is selected to be greater than the total amount of training images. The size of SOFM is 8 × 8 in our experiments, which is selected arbitrarily but to be large enough for good separation in-between clusters. We also try other size and dimension of SOFM's to assess the effects of SOFM's topology. The adaptation neighborhood size and adaptation coefficient are 7 and 1 initially, and linearly decreased to 1 and 0.01 at the end, respectively. To avoid the problem of converging to a local optimum during learning, we carefully select the initial guesses of the synaptic weights of the SOFM's. The average image of all training images is computed (as shown in fig. 5.2); and the multi-resolution representations of this average image are used as the initial guesses.

Table 5.1 shows the experimental results for three candidate membership thresholds CMT^γ for 9000 learning cycles. The more candidates are survived at the resolution γ if CMT^γ is set smaller. The results show that all three CMT^γ's achieve 100% correct classification and the selection of CMT^γ seems not so crucial. Table 5.2 shows the experimental results for various learning cycles with CMT^γ set at 0.5 for all resolutions. The facial information is so complicated that insufficient learning may lead to insufficient separable clustering on the SOFM. The results also show that very stable recognition is achieved with incorrectly classified samples no more than 1 when the learning cycles are greater than 4000. Table 5.3 shows the experimental results for different size and dimension of SOFM's. The results show that the 12 × 12 SOFM also can achieve high classification accuracy, but it requires more learning cycles than 8 × 8 does. Though the increase of SOFM size can increase the separation in-between clusters on the map, it costs more training cycles. An alternative to this problem is to increase the dimension of the SOFM. The experimental results show that a 3-dimensional 8 × 8 × 8 SOFM with similar network configurations to those of a 2-dimensional SOFM have 100% classification accuracy for all the learning cycles under test. However, the shortcoming of this 3-D

configuration is the tremendous amount of computations required per learning cycle.

It is interesting to note that most of the tested images (more than 90%) are recognized during the first and second stages of sifting, 26 out of 40 recognized at the first stage, 11 recognized at the second stage. That is, the coarse resolution provides sufficient discriminative information which leads to unique survived candidate regardless of the value of CMT^{γ}. The sufficient learning cycles for various resolutions should be different due to the different complexity of facial images; and a coarse resolution ought to require less training cycles than a fine one. However, it is difficult to determine the 'optimal' training cycles for each resolution. The only criterion for justification is the classification accuracy which is obtained by integrating the classification from all resolutions, not simply an individual resolution. Besides, the analysis of the effect on recognition performance for each parameter is also difficult to perform. For instance, accurate recognition may be due to the good modeling of facial images, the sifting procedure of pyramidal decision strategy, or the suitable CMT^{γ} values. In fact, all the stages in our scheme are strongly related and contributed to the high classification accuracy.

6. Conclusion

In this paper, we propose a novel approach for facial image recognition. Various statistical properties associated with various resolutions of faces are exploited via the SOFM neural networks. The coarse-to-fine pyramidal decision strategy also is proposed to constitute a hierarchical classification based on various statistical properties of a face, which subsequently results in a saving of computation and increase of decision accuracy. The experiments show that our proposed scheme has very good performance and is able to achieve 100% classification accuracy and comparable to other schemes reported in the literature.

However, in our scheme, the huge amount of learning cycles become a burden of the training procedure. Therefore, a study on different neural networks for automatically and efficiently exploiting discriminative facial features for various resolutions, which subsequently leads a fast training, become a crucial issue for the future investigation.

Reference

[1] Brunelli, R and Poggio, T., Face Recognition: feature versus templates, *IEEE Trans. on Pattern Anal. and Mach. Intel.*, 15, 1042-1052, 1993.

[2] Campos, C., Linney, A.D. and Most, J.P., The analysis of face profiles using scale space techniques", *Pattern Recognition*, 26, 819-824, 1993.

[3] Freeman, J.A. and Skapura, D.M., *Neural network, algorithms, application, and programming techniques*, Addison-Wesley, 1991.

[4] X. Jia, X. and Nixon, M.S., Analysing front view face profiles for face recognition via the Walsh transform, *Pattern Recognition Letters*, 15, 551-558, 1994.

[5] Kirby, M. and Sirovich, L., Application of the Karhunen-Loeve procedure for the characterisation of human faces, *IEEE Trans. on Pattern Anal. and Mach. Intel.*, 12, 103-108, 1990.

[6] Kohonen, T., *Self-Organization and Associative Memory*, Springer-Verlag, Berlin, 1988.

[7] Kundu, A. and Chen, J.-L., Texture classification using QMF bank-based subband decomposition, *CVGIP: Graphical Model and Image Processing*, 54, 369-384, 1992.

[8] Luckman, A.J., Allinson, A.M., Ellis, A.W. and Flude, B.M., Familiar face recognition: A comparative study of a connectionist model and human performance, *Neurocomputing*, 7, 3-27, 1995.

[9] Mallat, S.G., A theory for multiresolution signal decomposition:the wavelet representation, *IEEE Trans. on Pattern Anal. and Mach. Intel.*, 11, 7, 674-693, 1989.

[10] Rosenfeld, A. ed., *Multiresolutional Image Processing and Analysis*, Springer-Verlag, Berlin, 1984.

[11] Samal, A and Iyengar, P.A., Automatic recognition and analysis of human faces and facial expressions: a survey, *Pattern Recognition*, 25, 65-77, 1992.

[12] Turk, M. and Pentland, A., Eigenfaces for recognition, *Journal of Cognitive Neuroscience*, 3, 71-86, 1991.

[13] Valentin, D., Abdi, H., O'Toole, A.J. and Cottrell, G.W., Connectionist models of face processing: a survey, *Pattern Recognition*, 27, 1209-1230, 1994.

[14] Woods, J.W. and O'Neil, S.D., Subband coding of image, *IEEE Trans. on Acoust., Speech and Signal Processing*, ASSP-34, 1278-1288, 1986.

[15] Wu, C.J. and Huang, J.S., Human face profile recognition by computer, *Pattern Recognition*, 23, 255-259, 1990.

[16] Young, A.W. and Ellis, H.D., *Handbook of Research on Face Processing*, Elsevier Science, Amsterdam, 1989.

[17] Yuille, A.L., Deformable templates for face recognition, *Journal of Cognitive Neuroscience*, 3, 59-70, 1991.

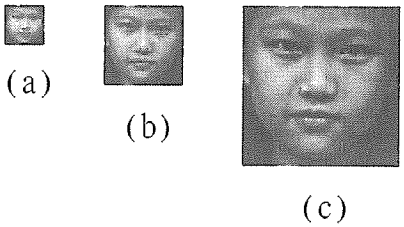

Fig. 1.1: Images of three resolutions corresponds to gazing a face at three viewing distances. (a) long distance, (b) medium distance, and (c) short distance.

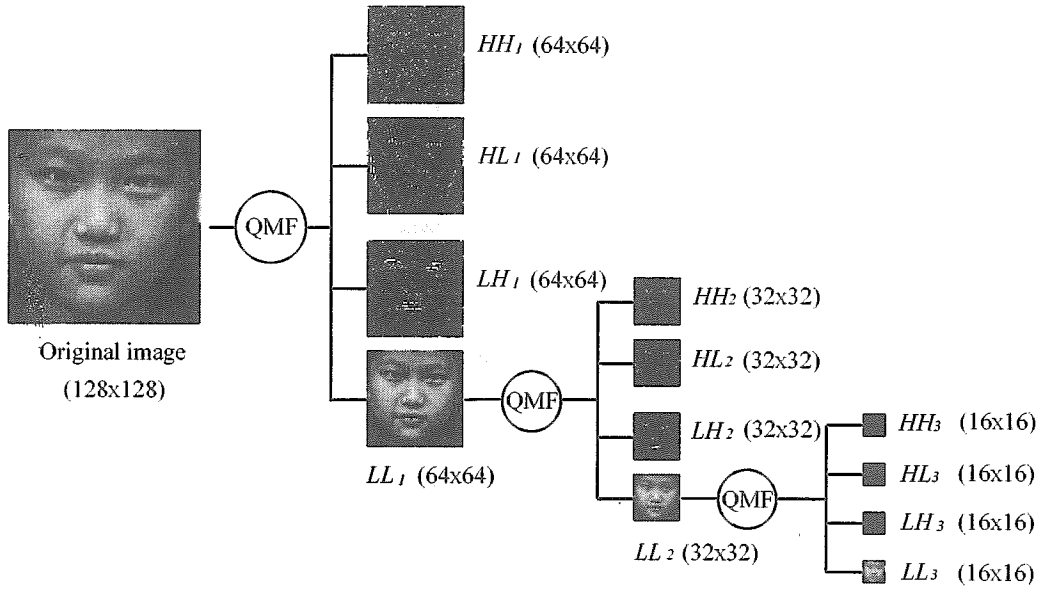

Fig. 2.1: Multiresolution decomposition of face of size 128×128 via a 3-layer pyramidal QMF bank.

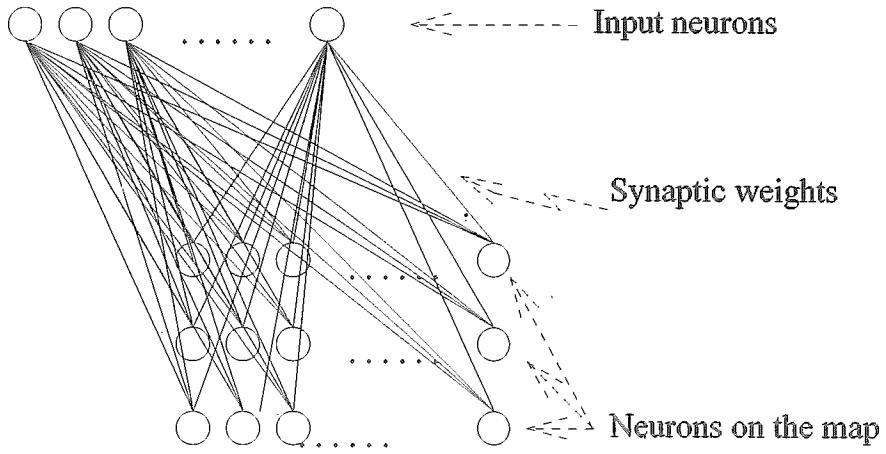

Fig. 3.1: Two dimensional SOFM neural network.

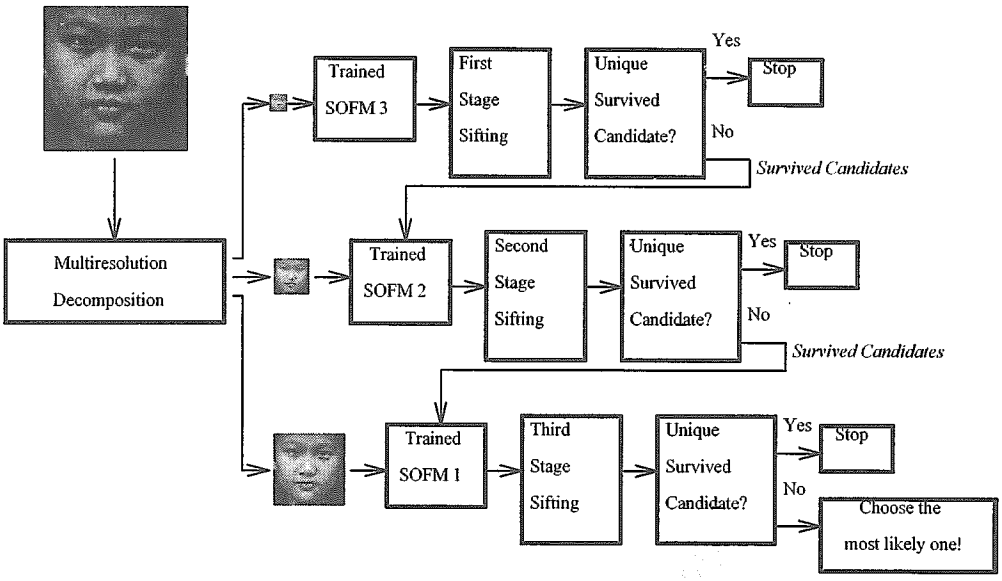

Fig. 4.1: Block diagram of proposed coarse-to-fine pyramidal sifting strategy.

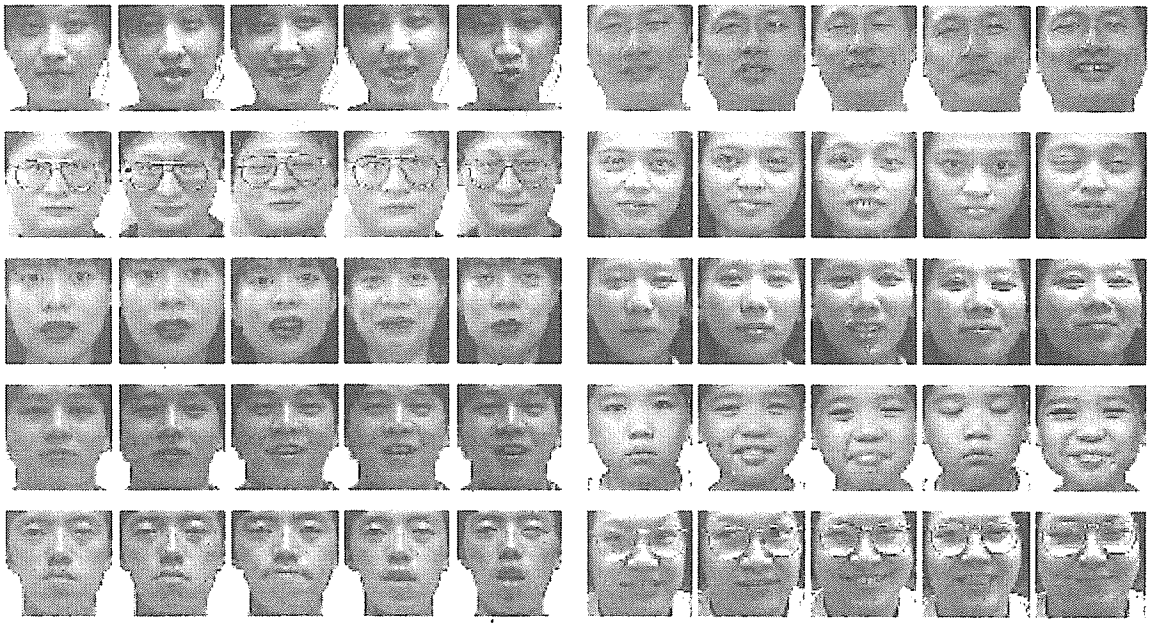

Fig. 5.1: Facial images of 10 persons for experiments. Five images are taken for each person and arranged in a row. The left image for each row is used for training; and the others are used as test samples.

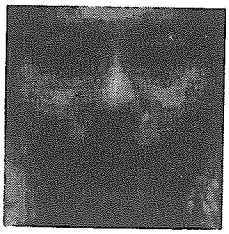

Fig. 5.2: The average image of training images. The multi-resolution representations of this average image are used as the initial guesses of the synaptic weights of the associated SOFM's.

	CMT^{γ} =0.5	CMT^{γ} =0.75	CMT^{γ} =0.9
8 × 8 SOFM	100% (error=0)	100% (error=0)	100% (error=0)

Table 5.1: The classification results for different candidate membership thresholds. The SOFM is trained with 9000 learning cycles and the CMT^{γ}'s are the same for all resolutions.

	T_L=3000	T_L=4000	T_L=5000	T_L=6000	T_L=7000	T_L=8000	T_L=9000
8 × 8 SOFM	95% (error=2)	100% (error=0)	97.5% (error=1)	97.5% (error=1)	97.5% (error=1)	97.5% (error=1)	100% (error=0)

Table 5.2: The classification results of different learning cycles with CMT^{γ} =0.5, γ=1,2,3.

	T_L=3000	T_L=4000	T_L=5000	T_L=6000	T_L=7000	T_L=8000	T_L=9000
12 × 12 SOFM	92.5% (error=3)	95% (error=2)	92.5% (error=3)	92.5% (error=3)	97.5% (error=1)	97.5% (error=1)	100% (error=0)
8 × 8 × 8 SOFM	100% (error=0)	100% (error=0)	100% (error=0)	100% (error=0)	100% (error=0)	100% (error=0)	100% (error=0)

Table 5.3: The classification results of different SOFM's, 12 × 12 and 8 × 8 × 8, for different learning cycles with CMT^{γ} =0.5, γ=1,2,3.