Descriptor Based Models for Face Recognition

Chapter 3

Descriptor Based Models for Face Recognition

3.1 Background

Every physical object in the universe has its own shape, texture and color The description of an object is always with reference to these attributes There may be more than one object which has same description based on shape, texture and color To give better description of an object for the purpose of discriminating with other objects, the attributes viz, shape, texture and color may not be sufficient The discriminating capability of human vision employs the method of identifying some unique "points" called key points or interesting points in some regions of an object For any object in an image, interesting points on the object can be extracted to provide a "feature description"

---

* Some parts of the material in this chapter appeared in the following research papers

1 **Discriminative Scale Invariant Feature Transform (SIFT-FLD) model for efficient representation and accurate recognition of faces** Proceedings of Indian International Conference on Artificial Intelligence-IICAI, Tumkur, India, December 16-18, pp 1914-1927 2009

2 **Monogenic Scale Space Based Region Covariance Matrix Descriptor for Face Recognition.** Proceedings of Bilateral Russian-Indian Scientific Workshop On Emerging Applications Of Computer Vision EACV-2011 , Moscow, Russia, November 1-5, pp 29-36,2011

of the object This description, extracted from a training image, can then be used to identify/recognize the object when attempting to locate the object in a test image containing many other objects To perform reliable recognition, it is important that the features extracted from the training image are detectable even under changes in image scale, noise and illumination Such points usually lie on high-contrast regions of the image, such as face edges Another important characteristic of these features is that the relative positions between them in the original scene may not change from one image to another For example, if only the four corners of a door were used as features, they would work regardless of the door's position, but if points in the frame were also used, the recognition would fail if the door is opened or closed Similarly, features located in articulated or flexible objects would typically not work if any change in their internal geometry happens between two images in the set being processed However, in practice the descriptors detects and uses a much larger number of features from the images, which reduces the contribution of the errors caused by these local variations in the average error of all feature matching errors

Local descriptors are commonly employed in a number of real-world applications such as object recognition (Lowe, 1999) and image retrieval (Mikolajczyk and Schmid, 2005) because they can be computed efficiently, are resistant to partial occlusion, and are relatively insensitive to changes in viewpoint There are two considerations for using local descriptors in these applications First, we must localize the interest point in position and scale Typically, interest points are placed at local peaks in a scale-space search, and filtered to preserve only those that are likely to remain stable over transformations Second, we must build a description of the interest point, ideally, this description should be distinctive (reliably differentiating one interest point from others), concise, and invariant over transformations caused by changes in camera pose and lighting While the localization and description aspects of interest point algorithms are often designed together, the solutions to these two problems are independent (Mikolajczyk and Schmid 2005) This chapter focuses on approaches to the second aspect – the construction and evaluation of local descriptor based representations for face

recognition applications Mikolajczyk and Schmid (2005) presented a comparative study of several local descriptors including steerable filters (Freeman and Adelson, 1991) differential invariants (Koenderink and van Doorn 1987), moment invariants (Van Gool, et al , 1996), complex filters (Schaffalitzky and Zisserman 2002), SIFT (Lowe, 1999), and cross-correlation of different types of interest points (Harris and Stephens, 1988) Their experiments showed that the ranking of accuracy for the different algorithms was relatively insensitive to the method employed to find interest points in the image but was dependent on the representation used to model the image patch around the interest point Since their best matching results were obtained using the SIFT descriptor, this chapter focuses on that algorithm and explores alternatives to this local descriptor representation The remainder of this chapter is organized as follows Section 3 2 reviews the related work in the field of face recognition using descriptors analysis and gives details about popular descriptor based models viz , DAISY, LBP, SURF and SIFT which are used for recognition of faces Section 3 3 gives details about our proposed methodologies for efficient representation and accurate recognition of faces

3.2. Local Descriptors Based Face Recognition: A Review

Feature detectors can be traced back to the Moravec's corner detector (Moravec 1977), which looks for the local maximum of minimum intensity changes As pointed by Harris and Stephens (1988) the response of this detector is anisotropic, noisy, and sensitive to edges To reduce these shortcomings, the Harris corner detector (Harris and Stephens, 1988) was developed However, it fails to deal with scale changes, which always occur in images Therefore, the construction of detectors that can cope with this scaling problem is important Lowe (2004) pioneered a scale invariant local feature, namely the scale invariant feature transform (SIFT) To deal with the viewpoint changes, Mikolajczyk and Schmid (2004) put forward the Harris (Hessian) affine detector, which incorporates the Harris corner detector (the Hessian point detector), scale selection, and second moment matrix based elliptical shape estimation Tuytelaars and Van Gool (2004) developed an edge-based region detector which considers both curved and straight edges to construct

parallelograms associated with the Harris corner points They also proposed an intensity-based detector (Tuytelaars and Van Gool, 2004), which starts from the local extrema of intensity and constructs ellipse-like regions with a number of rays emitted from these extrema Both the edge- and intensity-based methods preserve the affine invariance Matas et al (2002) developed the maximally stable extremal region (MSER) detector, which is a watershed-like method The last but not the least affine invariant detector is the salient region detector (Kadir et al , 2004), which locates regions based on an entropy function

To represent points and regions, which are detected by the above methods, a large number of different local descriptors have been developed The earliest local descriptor could be the local derivatives (Koenderink and van Doorn, 1989) Florack et al (1994) incorporated a number of local derivatives and constructed the differential invariants, which are rotational invariant, for local feature representation Schmid and Mohr (2001) extended local derivatives as the local gray value invariants for image retrieval Freeman and Adelson (1991) proposed steerable filters, which are linear combinations of a number of basis filters, for orientation and scale selection to handle tasks in image processing and computer vision research Marcelja (1980) modeled the responses of the mammalian visual cortex through a series of Gabor functions (Daugman 1980), because these functions can suitably represent the receptive field profiles in cortical simple cells Therefore, Gabor filters can be applied for local feature description Wavelets, which are effective and efficient for multi-resolution analysis, can also represent local features Textons, e g , 2D textons, 3D textons (Leung, Malik 2001), and the Varma–Zisserman model (2002), have been demonstrated to have good performance for texture classification A texton dictionary is constructed from a number of textures and a clustering algorithm is applied to select a small number of models to represent each texture Texton representation is also a good choice for local feature modeling Van Gool et al (1996) computed Texton for moments up to second order and second degree based on the derivatives of x and y directions of an image patch Li and Xiangxin (2007) have studied the use of textons as a robust approach for face representation and recognition In

particular, they propose local Gabor textons extracted by using Gabor filters and K-means clustering algorithm in local regions

In the recent days, we have seen many variants of SIFT technique for face recognition problem which is initially designed for object recognition that detects and extracts local feature descriptors that are reasonably invariant to changes in illumination, image noise, rotation, scaling, and small changes in viewpoint The SIFT transforms image data into scale-invariant coordinates relative to local features that consists of four major phases (1) scale-space peak selection, (2) keypoint localization, (3) orientation assignment, (4) keypoint descriptor Bicego et al (2006) pioneered the use of SIFT in this field of face recognition and they apply SIFT keypoint extraction and comparison using different methods In their first approach, they extract the pair of descriptors that provides the minimum distance between the query and the template image This value of distance is then used to compute the similarity between faces In a second approach, they use a regular grid to extract and to compare the descriptors, the distances are then averaged in order to compute a similarity score for the given images A similar technique was proposed (Luo et al , 2007) where faces are segmented in 5 different areas (the two eyes, the nose, and the two sides of the mouth) The features belonging to each area on the query are compared with the ones present in the corresponding area on the template image Additionally a full match strategy is performed that does not consider the 5 regions constraint Both the scores for local and global analysis are then considered for the identification

Other approaches by (Kisku et al 2007, Ozkan et al 2006) exploit the SIFT features to build a connected graph that links all the features extracted from a face A graph matching algorithm is then introduced to compute the matching score Mohamed Aly (2006) proposed the new approach for face recognition based on SIFT matching features It was compared with earlier methods viz , PCA and LDA and the results of SIFT were better than PCA and LDA Wang and Miao (2008) proposed Scale Invariant Feature Extraction method which uses Bayesian Probabilistic Similarity matching measure The

performance was better for images of different resolutions Lanzarini et al (2010) proposed the strategy for face recognition based on SIFT descriptors of the various images In order to reduce the number of false positives and computation time, a selection of the most representative feature descriptor is carried out by applying a variation of the binary Particle Swarm Optimization (PSO) method

On the other hand, a more efficient and accurate local descriptor is proposed by Herbert Bay called Speed-Up Robust Feature (SURF) which is a scale and in-plane rotation invariant feature detector It contains interest point detector and descriptor The detector locates the interest points in the image, and the descriptor describes the features of the interest points and constructs the feature vectors of the interest points One of the main advantages of SURF is to be able to compute distinctive descriptors quickly In addition, SURF descriptor is invariant to common image transformations including image rotation, scale changes, illumination changes, and small change in viewpoint However, instead of difference of Gaussians (DoG) filter used in SIFT, SURF uses Hessian-matrix approximation operating on the integral image to locate the interest points, which reduces the computation time drastically As for the descriptor, the first-order Haar wavelet responses in x and y directions are used in SURF to describe the intensity distribution within the neighborhood of an interest point, whereas the gradient is used by SIFT In addition, only 64 dimensions are usually used in SURF to reduce the time cost for both feature computation and matching Because each of SURF feature has only 64 dimensions in general and an indexing scheme is built by using the sign of the Laplacian, SURF is much faster than the 128-dimensional SIFT at the matching step Furthermore, the method used to speed up the SIFT matching can also be applied to SURF

Dreuw et al (2009) analyzed the usage of SURF as local descriptors for face recognition The effects of different feature extraction and viewpoint consistency constrained matching approaches are analyzed Furthermore, a RANSAC based outlier removal for system combination was also proposed Their proposed approach is working for faces under partial occlusions, and even if they are not perfectly aligned or illuminated Du and

Cai (2009) successfully applied SURF for face recognition Just like in SIFT, SURF detectors are first employed to find the interest points in an image, and then the descriptors are used to extract the feature vectors at each interest point Yunqi et al (2009) proposed an approach for face feature extraction based on Speeded-Up Robust Feature They use Fisher Linear Discriminant (FLD) method to extract the quadratic features on the basis of SURF feature, and then measure the similarity of faces by calculating the Euclidean distance of the quadratic features

One more popular local image descriptor is the DAISY, which is very efficient to compute densely Unlike SURF, which can also be computed efficiently at every pixel, it does not introduce artifacts that degrade the matching performance when used densely DAISY consists of a vector made of values from the convolved orientation maps located on concentric circles centered on the location, and where the amount of Gaussian smoothing is proportional to the radii of the circles The authors proposed to use a circular grid instead of SIFT's regular one since it has been shown to have better localization properties In that sense, DAISY descriptor is closer to GLOH before PCA than to SIFT Also, the descriptor is naturally resistant to rotational perturbations as well by the use of isotropic Gaussian kernels with a circular grid The overlapping regions ensure a smooth changing descriptor along the rotation axis and by increasing the overlap we can further increase the robustness up to a certain point

Carmelo Velarado and Jean-Luc Dugelay (2010) proposed a new face recognition approach based on DAISY, a dense computed SIFT like descriptor This algorithm is designed to be fast for dense computation and useful for re-identification as it is able to distinguish pairs of images as belonging to the same subject or not The descriptors are computed densely and matched with a new strategy that represents an efficient tradeoff between accuracy and computational load, afterwards a Support Vector Machine is used to classify the output of the matching to recognize if the pair of images belongs to the same person The results show that DAISY gives better performance than SIFT techniques Simon Winder et al, (2009) explored DAISY and apply the same for

recognition of faces They developed a new training set of match/non-match image patches which improves on previous works Also they tested a wide variety of gradient and steerable filter based configurations and optimize over all parameters to obtain low matching errors for the descriptors

The Local Binary Pattern (LBP) is another popular local descriptor that is used for facial image analysis in recent years LBP has been exploited for facial representation in different tasks, which include face detection (Zhang and Zhao, 2004) face recognition (Tan and Triggs 2007), facial expression analysis (Shan et al , 2009), demographic (gender, race, age, etc) classification (Zhao et al 2006), and other related applications (Zhang et al , 2006)

Ahonen et al (2004) used LBP for face recognition where a face image is first divided into several blocks (facial regions) from which they extract local binary patterns and construct a global feature histogram that represents both the statistics of the facial micro-patterns and their spatial locations Then, face recognition is performed using a nearest neighbor classifier in the computed feature space We have also seen in the literature the works of Zang et al (2005), Shan et al (2006), Tan (2007) and Lei et al (2008) contributed towards face recognition using Local Binary Patterns

The RCM, proposed by Tuzel *et al* (2006) is a matrix of covariance of several image statistics computed inside a region of an image The RCM is considered as a feature descriptor of the region Classification is conducted based on these RCMs Pang et al , (2008) proposed a new way to further enhance the discriminating ability of RCMs They proposed a new method for human face recognition by utilizing Gabor-based region covariance matrices as face descriptors Both pixel locations and Gabor coefficients are employed to form the covariance matrices Experimental results demonstrate the advantages of this proposed method over other methods

It shall be observed from the above discussion that the local descriptors are the excellent choices for face representation and accurate recognition because of their inherent capabilities such as illumination invariance, robustness against noise and occlusion withstanding capabilities In this context, we have developed variants of SIFT and RCM

for efficient face representation and accurate classification The details of the proposed techniques are described below

3.3 Proposed Model-I: Discriminative Scale Invariant Feature Transform (SIFT-FLD) for Face recognition

The proposed discriminative SIFT is an integrated approached that combines SIFT with linear descriminant analysis (LDA) to have a compact and discriminative descriptor for human faces The first stage of SIFT model is scale-space extrema detection, which computes interest points that are invariant to scale and orientation are extracted by searching over all scales and image locations Interest points for SIFT features correspond to local extrema of difference-of-Gaussian filters at different scales The first step toward the detection of interest points is the convolution of the image with Gaussian filters at different scales, and the generation of difference-of- Gaussian images from the difference of adjacent blurred images The convolved images are grouped by octave by having same number of difference-of-Gaussian images per octave with a fixed number of blurred images per octave An efficient approach to construction of Gaussian pyramid is as follows

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y) \quad (3.1)$$

$$D(x, y, \sigma) = L(x, y, k\sigma) - L(x, y, \sigma) \quad (3.2)$$

Where $I(x, y)$ is the input image, * is the convolution operator in x and y, σ is a scale factor, k is a constant factor, D is the difference of Gaussian image and

$$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2)/2\sigma^2} \quad (3.3)$$

The second stage is keypoint localization, where keypoints are identified as local maxima or minima of the DoG images across scales Each pixel in the DoG images is compared

to its 8 neighbors at the same scale, plus the 9 corresponding neighbors at neighboring scales If the pixel is a local maximum or minimum, it is selected as a candidate keypoint All the candidate keypoints are localized to sub-pixel accuracy and eliminated if found unstable An accurate position fix on the key-points located in the previous step has been implemented by fitting a 3D quadratic function to the local sample points

$$D(x) = D + \frac{\partial D^T}{\partial X} X + \frac{1}{2} X^T \frac{\partial^2 D^T}{\partial X^2} X \qquad (3\,4)$$

The third stage is orientation assignment, where the keypoint is assigned an orientation It involves calculating the gradient vectors in a window around the SIFT feature on the scale at which the feature was detected To determine the keypoint orientation, a gradient orientation histogram is computed in the neighborhood of the keypoint using the Gaussian image at the closest scale to the keypoint's scale The contribution of each neighboring pixel is weighted by the gradient magnitude and a Gaussian window with σ that is 1.5 times the scale of the keypoint Peaks in the histogram correspond to dominant orientations Assign this orientation to the key point detected A separate keypoint is created for the direction corresponding to any other the histogram maximum All the properties of the keypoint are measured relative to the keypoint orientation which provides invariance to rotation

Finally, keypoint descriptors are computed by having keypoint orientation The feature descriptor is computed as a set of orientation histograms on 4 x 4 pixel neighborhoods The orientation histograms are relative to the keypoint orientation and the orientation data comes from the Gaussian image closest in scale to the keypoint's scale Histograms contain 8 bins each, and each descriptor contains an array of 4 histograms around the keypoint This leads to a SIFT feature vector of size (4x4) x 8

We have used SIFT descriptor to extract local features However it is identified that the dimensionality of SIFT descriptor is very high and hence it consumes much memory for representation and time for recognition To overcome these problems discriminative

subspace analysis technique called $(2D)^2$FLD (Nagabhushan et al , 2006) is employed on SIFT descriptors for their compact representation The dimension of the feature vector is significantly smaller than the standard SIFT feature vector

Our contribution lies in exploring FLD on key points in the context of face representation and recognition Unlike PCA-SIFT (Sukthankar) that employ PCA to the normalized gradient patch, in our proposed model FLD is employed on the key descriptors The proposed model is called SIFT-FLD, described in the following subsections

3.3.1 Training scheme

Formally, let there be T number of classes each with p_i, i=1 T, number of training images Therefore we have totally $N=\sum_{i=1}^{T} p_i$ number of training images Let A_i^j be an image of size m x n representing the j^{th} sample in the i^{th} class Let each image A_i^j has k_i^j number of keypoint's It shall be observed here that each keypoint is described by a key descriptor of dimension 16x8 and there is varying number of key points for each image Hence, associated with each image A_i^j, we are having k_i^j key descriptor matrices of size 16 x 8 Let $\overline{C_i}$ be the average key descriptor matrix of all k_i^j training images keypoint descriptor matrices of the i^{th} class Let **C** be the average key descriptor matrix all the N training images key descriptor matrices

To obtain, row-wise optimal projection axes, the descriptors between-class scatter matrix G_b and the descriptors within-class scatter matrix G_w are computed as follows

$$G_b = \frac{1}{N}\sum_{i=1}^{T} k_i (\overline{C}_i - C)^T (\overline{C}_i - C) \tag{3 5}$$

$$G_w = \frac{1}{N}\sum_{i=1}^{T} \sum_{j=1}^{k_i} (A_i^j - \overline{C}_i)^T (A_i^j - \overline{C}_i) \tag{3 6}$$

Once G_b and G_w are computed, it is recommended to find the optimal projection axis X so that the total scatter of the projected samples of the training images is maximized For this purpose, the Fisher's criterion given by

$$J(X) = \frac{X^T G_b X}{X^T G_w X} \tag{3 7}$$

is used It is a well-known fact that the eigenvector corresponding to the maximum eigenvalue of $G_w^{-1}G_b$ is the optimal projection axis which maximizes J(X) Generally, as it is not enough to have only one optimal projection axis, we usually go for *d* number of projection axes, say X_1, X_2, ,X_d, which are the eigenvectors corresponding to the first *d* largest eigenvalues of $G_w^{-1}G_b$

To obtain column-wise projection matrix, the descriptors between-class scatter matrix H_b and the descriptors within-class scatter matrix H_w are computed as follows

$$H_b = \frac{1}{N}\sum_{i=1}^{T} k_i(\bar{C}_i - C)(\bar{C}_i - C)^T \tag{3 8}$$

$$H_w = \frac{1}{N}\sum_{i=1}^{T}\sum_{j=1}^{k_i}(A_i^j - \bar{C}_i)(A_i^j - \bar{C}_i)^T \tag{3 9}$$

It shall be observed that H_b and H_w in Eqs (3 8) and (3 9) are obtained in this new formulation as outer products of column vectors unlike G_b and G_w (Eqs (3 5) and (3 6)) Using these two scatter matrices, we find the optimal projection axes Z (*m* x *c*) so that the total scatter of the projected samples is maximized using the same fisher's criterion given by

$$J(Z) = \frac{ZH_bZ^T}{ZH_wZ^T} \tag{3 10}$$

Thus, the eigenvectors of $H_w^{-1}H_b$ are computed and then *c* number of eigenvectors corresponding to the first *c* largest eigenvalues of $H_w^{-1}H_b$ are chosen

Finally, we recommend to project the key descriptor matrices on both directions simultaneously while extracting feature matrices Let X denotes *n* x *d* optimal projection matrix obtained and let Z denote the *m* x *c* matrix obtained During training, each key

descriptor matrix k_i^J of the training image A_i^J is projected onto both X and Z simultaneously to obtain the respective feature matrix F_i^J which is of dimension c x d as follows

$$F_i^J = Z^T k_i^J X \tag{3 11}$$

Each feature matrix F_i^J is represented in vector form for distance computational purpose Thus fisherface analysis can drastically reduce the dimension (16x8) of key descriptor matrix to the SIFT-fisherface dimension (c × d) while keeping several of the most effective features that summarize the original descriptors

3.3.2 Recognition

Let I be an image given for recognition Let I' be the key descriptor matrix which is projected onto the c and d number of optimal projection axes Z and Xs respectively that results in test image feature matrix say, I'' computed by $I'' = Z^T I' X$ This test image feature matrix is represented in vector for computation purpose Given two images, say v_{i_1} and v_{i_2} of any two face(s), represented by feature vectors $r = [r_1, r_2, \ , r_q]$ and $s = [s_1, s_2, \ , s_q]$, the $dist(r,s)$ is defined as

$$dist(r,s) = \sum_{j=1}^{q} \| r_j - s_j \|_2 \tag{3 12}$$

where $\|a-b\|_2$ denotes the Euclidean distance between the two vectors a and b

For classifying a given test image, we have considered a voting scheme which decides the class We have proposed voting scheme for classification as there are varying number of feature vectors in the knowledgebase associated with each image which are having varying number of keypoints The voting scheme is as follows The T-dimensional vote vector is initialized to zero (here, T is the number of classes) The similarity between each feature vector of test image and the feature vector of the knowledgebase is

computed and the i^{th} location of the VOTE vector is incremented if for a given test image feature vector I_t^*, if $dist(I_t^*, F_l) = \min_j dist(T_i, F_j)$ and $F_l \in T_i$, then the resulting decision is that I_t^* is the key point of T_i^{th} class Once the similarity of the entire feature vectors of test image are computed, the T-dimensional vote vector is analyzed The test image is said to belong to the i^{th} class if i^{th} vote array gets the maximum votes

3.3.3 Experimental Results

This section presents the results of the experiments conducted to corroborate the success of the proposed model We have conducted experimentation on AT&T and CALTECH face databases We have specifically chosen this database as these are used by many researchers as a benchmark database to verify the validity of their proposed face recognition models All experiments are performed on a P-IV 2 99GHz Windows machine with 504 MB of RAM

Experimentation on AT&T face database: The AT&T face database contains images from 40 individuals, each providing 10 different images of size 112x92 In our experiment, we have considered alternate five samples per class for training and the remaining samples for testing Similarly, we have conducted experiments considering 160, 120 and 80 faces as training faces of the AT&T database choosing 4, 3 and 2 faces respectively from each person and the recognition performance has been obtained considering the remaining faces as test faces The recognition performance of proposed model with varying dimension of feature vectors is given in Table 3 1 and also shown in Fig 3 1

Table 3 1 Recognition accuracy of SIFT-FLD for AT&T face database

Number of Training faces	Number of Test faces	Percentage of recognition accuracy for varying dimensions of feature vectors			
		9	16	25	36
200	200	91 25	96 25	98 00	98 25
160	240	83 25	94 50	95 75	98 00
120	280	82 25	89 00	94 00	95 00
80	320	62 00	75 00	81 00	84 00

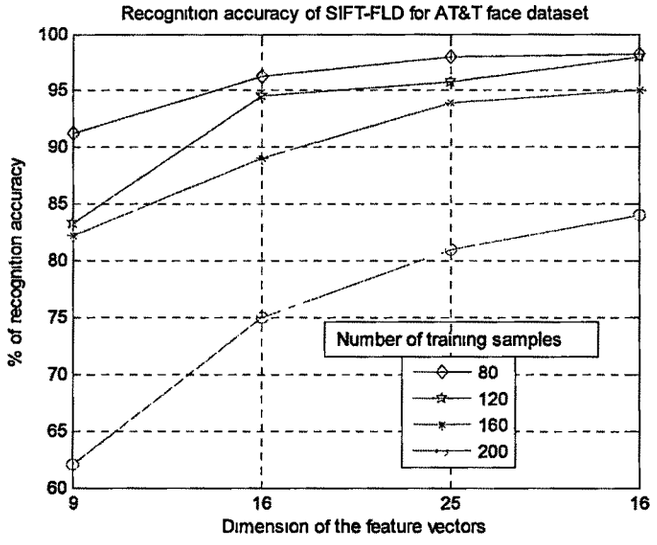

Fig 3 1 Recognition accuracy of SIFT-FLD for AT&T face database

Experimentation on CALTECH face database: The CALTECH face database contains images from 30 individuals, each providing varying number of images of size 130x100 In our experiment, we have considered alternate samples per class for training and the remaining samples for testing Similarly, we have conducted experiments considering 162 and 127 faces as training faces of the CALTECH database choosing 3 and 4 faces respectively from each person and the recognition performance has been obtained considering the remaining faces as test faces The recognition performance of proposed

model with varying dimension of feature vectors is given in Table 3 2 and also shown in Fig 3 2

Table 3 2 Recognition accuracy of SIFT-FLD for CALTECH face database

Number of Training faces	**Number of Test faces**	**Percentage of recognition accuracy for varying dimensions of feature vectors**			
		9	16	25	36
234	216	88 67	94 89	96 44	96 89
162	288	85 78	94 00	95 11	95 56
127	323	81 33	92 89	93 56	94 00

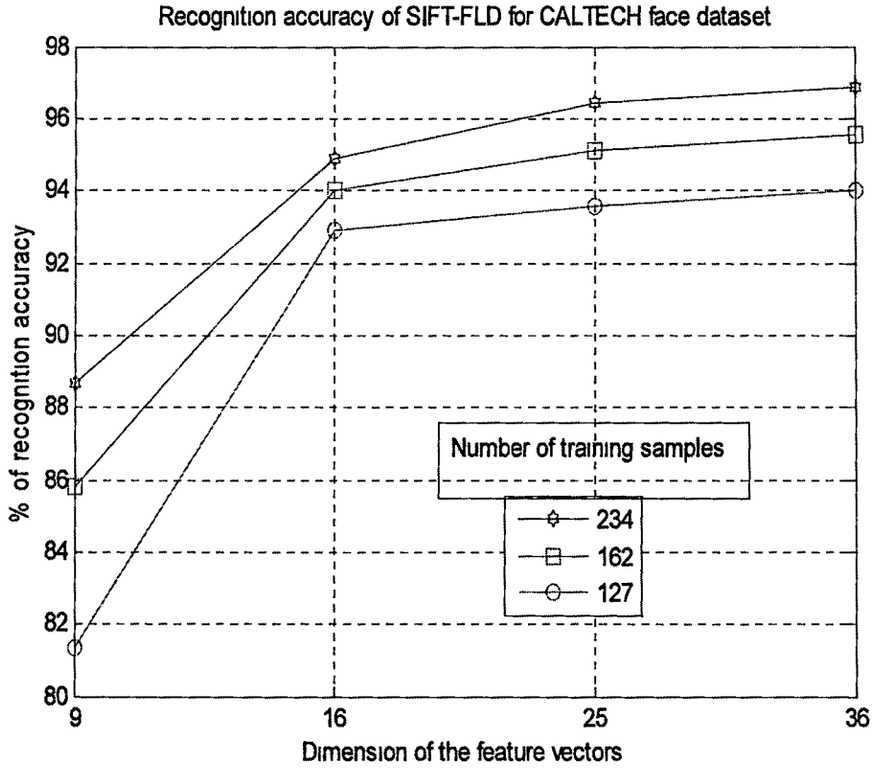

Fig 3 2 Recognition accuracy of SIFT-FLD for CALTECH face database

3.4 Proposed Model-II: Monogenic Scale Space Based Region Covariance Matrix Descriptor for Face Recognition

In this approach, we have proposed a new face recognition algorithm based on region covariance matrix (RCM) descriptor computed in monogenic scale space In this proposed model, energy information obtained using monogenic filter is used to represent a pixel at different scales to form region covariance matrix descriptor for each face image during training phase An eigenvalue based distance measure is used to compute the similarity between face images

Let I be an intensity image of size $W \times H$ Define a function ϕ that extracts d dimensional feature vector from a pixel at (x, y) of I , i e ,

$$\emptyset(I, x, y) = z_i \in \mathcal{R}^d \quad (3\ 13)$$

where $i = y \times W + x$ is the index of (x , y) Consider all the pixels (x , y) in region $\mathcal{R}$ i e , $(x , y) \in \mathcal{R}$

The number of pixels in the region $\mathcal{R}$ is n The region $\mathcal{R}$ can then be represented by the $d \times d$ covariance matrix of the feature points z_i inside the region

$$C_R = \frac{1}{n-1}\sum_{i=1}^{n}(z_i - u_{\mathcal{R}})(z_i - u_{\mathcal{R}})^T \quad .(3.14)$$

where $u_{\mathcal{R}}$ is the mean of z_i

$$u_{\mathcal{R}} = \frac{1}{n}\sum_{i=1}^{n} z_i \quad ...(3.15)$$

In (Tuzel et al, 2006) the feature mapping function is defined by pixel locations (x , y), color (RGB) values and the norm of the first and second-order derivatives of the intensities with respect to x and y Specifically, the function is shown in (3 16),

$$\emptyset(I, x, y) = z_i = [x\ y\ R(x,y)\ G(x,y)B(x,y)\ |Ix|\ |Iy|\ |Ixx||Iyy|]^T \quad (3\ 16)$$

where R, G and B are the RGB color values, Ix and Ixx are the first- and second-order derivatives, which can be expressed, respectively, as an

$$Ix \triangleq \frac{\partial I(x,y)}{\partial x} \quad and \quad Ixx \triangleq \frac{\partial^2 I(x,y)}{\partial x^2}$$

For gray level images (3 16) becomes

$$\emptyset(I,x,y) = z_i = [x\ y\ |Ix|\ |Iy|\ |Ixx||Iyy|]^T \tag{3.17}$$

For the human detection problem, Tuzel *et al* defined the mapping function

$$\emptyset(I,x,y) = [x\ y\ |Ix|\ |Iy|\ |Ixx|\ |Iyy|\ \theta(x,y)\ ]^T \tag{3.18}$$

where θ is the edge orientation

$$\theta(x,y) = arctan\left(\frac{I_y}{I_x}\right) \tag{3.19}$$

To investigate the role of intensity component$I(x,y)$, we propose the following two variants

$$\emptyset(I,x,y) = [x\ y\ |Ix|\ |Iy|\ |Ixx|\ |Iyy|] \tag{3 20}$$

$$\emptyset(I,x,y) = [x\ y\ I(x,y)\ |Ix|\ |Iy|\ |Ixx|\ |Iyy|\theta(x,y)] \tag{3.21}$$

Equation (3 20) is the counterpart of(3 17) Equation (3 17) has an intensity component, while (3 20) does not Equation (3 21) is the counterpart of (3 18). The former has an intensity component, while the latter does not

The RCM is a symmetric matrix Its diagonal entries represent the variance of each feature and the non-diagonal entries represent their respective correlations Thus, the RCM proposes a natural way of fusing multiple features without normalizing features or using blending weights Pixel locations (x,y) used as feature in equation (3.16) play an important role in the corresponding RCM Although the variance of s(x,y) is the same for all the regions of the same size, they are still important since their correlation with the

other features are used as the non-diagonal entries of the matrix Therefore, the constructed RCM capture both spatial and statistical properties

To perform classification, it is necessary to measure the distance/dissimilarity between the two RCMs C_1 and C_2 Since the covariance matrices lie on connected Riemannian manifold, common-used distance (e g , Euclidean distance) cannot be used Forstner and Moonen (1999) and Tuzel *et al* (2006) proposed to use eigenvalue-based distance measure

$$\rho(C_1, C_2) = \sqrt{\sum_{i=1}^{d} ln^2 \lambda_i(C_1, C_2)} \quad (3\ 22)$$

where $\lambda_1 \quad , \lambda_d$ are generated values of C_1 *and* C_2 computed from $\lambda_i C_1 u_i = C_2 u_i$ i=1, d

3.4.1 Monogenic Scale Space: a Review

The monogenic signal (Felsberg and Sommer, 2004) is based on the Riesz transform which is used instead of the Hilbert transform The monogenic signal analysis is a framework to interpret images in terms of the local phase, local orientation and local energy The monogenic signal is an effective tool to analyze 2-D signals in a rotation invariant manner The signal is built upon the first order Riesz transform The spatial representation of the Riesz kernel in 2D space is

$$\left(R_x(X), R_y(Y)\right) = \left(\frac{x}{2\pi|x|^3}, \quad \frac{y}{2\pi|x|^3}\right), x = (x, y) \in \mathbb{R}^2 \quad ..(3\ 23)$$

and its transfer function in the Fourier domain is

$$(F_u(\boldsymbol{u}), (F_v(\boldsymbol{u})) = \left(-i\frac{u}{|\boldsymbol{u}|}, -\boldsymbol{i}\frac{v}{|\boldsymbol{u}|}\right), \boldsymbol{u} = (u, v) \in \mathbb{R}^2 \quad (3\ 24)$$

For any image, say I(x), the monogenic signal is defined as the combination of I and its Riesz transform

$$I_m(x) = \big(I(x), R_x\{I\}(x)\big) = \big(I, R_x * I, R_y * I\big) \tag{3.25}$$

where * stands for the convolution operation

And hence the local orientation is calculated as (Felsberg and Sommer 2004)

$$\theta = arctan\frac{R_y\{I\}}{R_x\{I\}}, \theta \in (0, \pi) \tag{3 26}$$

$$\theta = arctan\frac{R_y\{I\}}{R_x\{I\}}, \theta \in (0, \pi) \tag{3 26}$$

The local phase is defined as

$$\varphi = atan2\left(\sqrt{R_x^2\{I\} + R_y^2\{I\}}, I\right) \varphi \in [0, \pi] \tag{3.27}$$

The local energy is defined as

$$E = \sqrt{R^2\{I\} + R_y^2\{I\} + R_x^2\{I\}}, I \quad \text{... (3 28)}$$

where $R^2\{I\} \in I * F^{-1}\big(G(w)\big)$.

Here $\big(G(w)\big)$ is the log-Gabor filter in the Fourier domain Since log-Gabor filters are band-pass filters, usually multi-scale monogenic representation is required to fully describe a signal In figure 3 3, we have given the convolved images in monogenic scale space filter showing the log-Gabor transformed image with its energy and orientation images One can see that the local structure is well captured in monogenic components

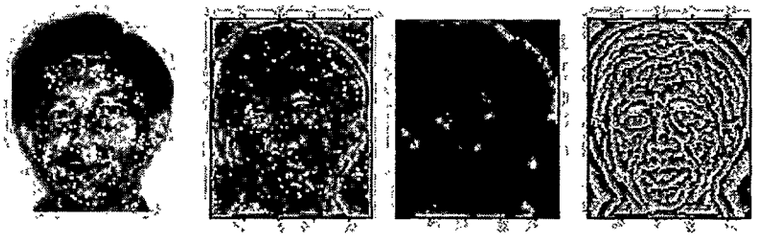

Fig 3 3 Original Image, log-Gabor Image, Energy image and orientation image

3.4.2 Construction of RCM in Monogenic Scale Space

The monogenic scale space based features are obtained by employing Riesz transform The monogenic filter is employed on an intensity image, say I at different scales to obtain the local energy, local phase and local orientation information In our work, we have considered local energy at five different scales for constructing RCM which is as follows

It shall be observed from Eq (3 28) that the energy can be computed at different scales Let A_1, A_2, ,A_5 be the energy computed at five different scales Hence, the feature mapping for each pixel (x, y) is defined as

$$\emptyset(I, x, y) = [A_1(x, y), A_2(x, y), A_3(x, y), A_4(x, y), A_5(x, y)] \quad . (3\,29)$$

It shall be observed here that the size of the RCM is $5 \times 5 = 25$ One can notice here that the proposed model requires only five convolutions with the Riesz transform where as Gabor wavelet based representation requires 40 convolutions (5 orientation and 8 scales) with Gabor filter and hence the feature mapping results in 40×40 dimension feature vector which is very massive and hence consume much memory to store the feature vector Obviously the computing time increases due to more number of faces in image databases

So far, we have considered whole image as an input It is possible to take upper half, lower half, left half or right half of an image and could construct RCM possibly to handle

occlusion As suggested in (Zabih, 1999), the eigen-value based distance measure (Eq (3 22)) is considered for finding the similarity between two faces images

3.4.3 Experimental Results

This section presents the results of the experiments conducted to corroborate the success of the proposed model We have conducted experimentation using two benchmark face image databases namely AT&T and YALE face databases All the experiments are conducted using P-IV machine on Windows-7 platform using MATLAB 7 8 tool

Experimentation on AT&T face databases: The AT&T face database consisting of 400 gray-scale images of 10 subjects, each covering a wide range of poses as well as race, gender and appearance In Fig 3 4, we have shown subset of one such subject of the AT&T database The experimentation consists of varying number of training samples and testing samples under each class We have chosen first five samples for training and the remaining samples for testing for each person and the recognition accuracy is obtained Similarly, two other test cases are generated which consists of first four and three samples for training and the remaining samples for testing The results obtained due to the proposed model are reported in Fig 3 5 The results obtained due to Gabor wavelet based RCM and basic RCM are also shown in Fig 3 5 It shall be observed that the recognition accuracy of the proposed model is on par with the Gabor wavelet based RCM, and quite high when compared to the basic RCM approach Over all, the performance of the proposed model is good as it consumes less time with good recognition results

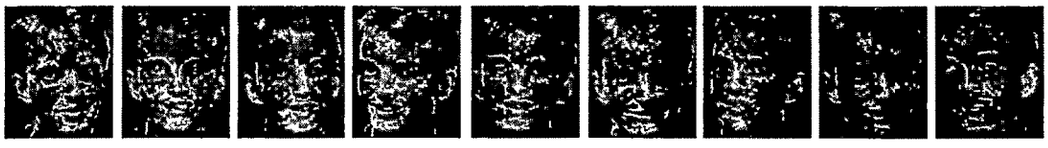

Fig 3 4 Subset of one subject of AT&T face database

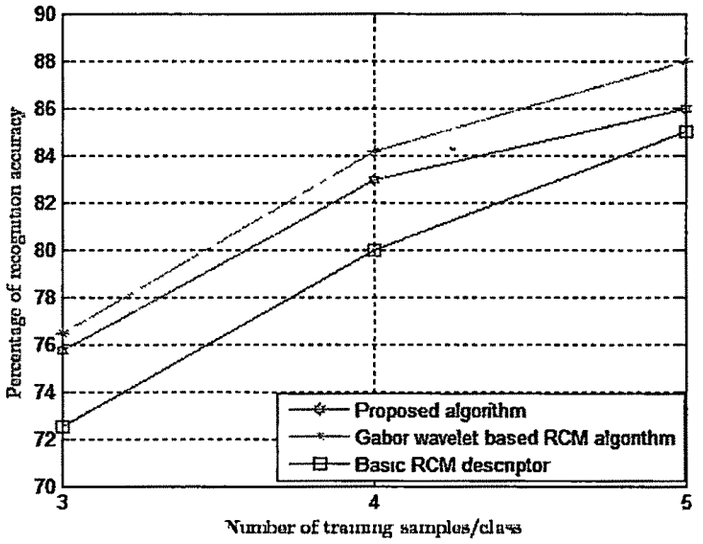

Fig 3 5 Recognition accuracy of the proposed model and other techniques on AT&T face database

Experimentation on YALE face databases: The YALE face database contains 165 images of 15 subjects that include variation in both facial expression and lighting In Fig 3 6, we have shown the closely cropped images which include internal facial structures Here, we have made the following type of testing The training set comprised of six images randomly chosen for each person with remaining number of face images for each person The recognition accuracy is computed for the proposed model, Gabor wavelet based RCM technique and the basic RCM technique Similarly, four and five face images are chosen randomly under each person during training and the remaining face images are used for testing The results are shown in Fig 3 7 It shall be observed from Fig 3 7 that the proposed model posses better performance

Fig 3 6 Subset of one subject of YALE face database

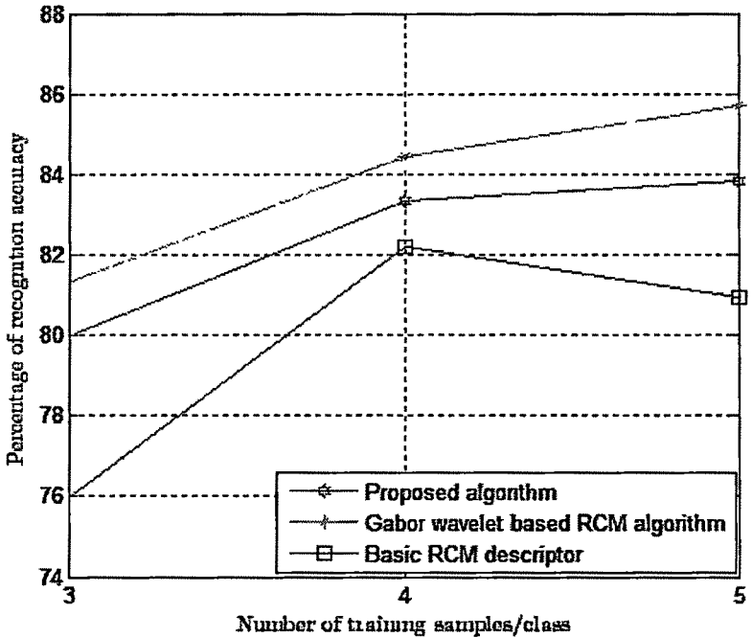

Fig 3 7 Recognition accuracy of the proposed model and other techniques on YALE face database

3.5 Conclusion

In this chapter, we have proposed two independent and integrated algorithmic models for efficient representation and accurate classification of human faces The first one is SIFT based and the second one is RCM based

Descriptor based paradigm is a well-known technique for face recognition and several attempts have been made by the research community to search for the best representation and as well for robust and efficient recognition We have made one such attempt for best representation and robust recognition of faces A more accurate and efficient SIFT-FLD is devised for the purpose of representing and recognizing face images The results are quite impressive and possess better recognition rate with least computing time for feature extraction The problem of huge memory requirement and hence much recognition time is resolved by employing FLD on the descriptor database

The monogenic scale space based region covariance matrix is also proposed for face recognition in this chapter The energy information contained in the monogenic signal at different scales is used to build the region covariance matrix descriptor for each face image Experimental results on the standard benchmark databases reveal the superiority of the proposed model for face recognition problems and its suitability in real environment