Small Sample Size Face Recognition using Random Quad-Tree based Ensemble Algorithm

Cuicui Zhang, Xuefeng Liang, Takashi Matsuyama

Graduate School of Informatics, Kyoto University, Kyoto 606-8501, Japan
zhang@vision.kuee.kyoto-u.ac.jp, {xliang, tm}@i.kyoto-u.ac.jp

Abstract

Certain applications such as person re-identification in camera network, surveillance photo verification, forensic identification etc. suffer from a small sample size (SSS) problem severely. Conventional face recognition methods face a great challenge on SSS as the trained feature space is overfitted to the small training set. Interest in combination of multiple base classifiers to solve the SSS problem has sparked renewed research efforts in ensemble methods. In this paper, we propose a novel Random Quad-Tree based ensemble algorithm (R-QT) to address the SSS problem. In contrast to other methods confining the ideas on limited data, R-QT enlarges the training data to obtain more diverse base classifiers. Moreover, R-QT encodes not only discriminant features but also the geometric information across the face region, which further improves the recognition accuracy. Results on five standard face databases demonstrate the effectiveness of the proposed method.

Keywords: Small sample size, Random Quad-Tree, base classifier ensemble

1 Introduction

1.1 Small Sample Size face recognition

Appearance based methods have been extensively studied and acknowledged as one of the most popular approaches for face recognition. Representative and well-known algorithms include Eigenfaces, Fisherfaces, Bayes Matching and their weighted, kernelized and tensorized variants [7]. Conventional methods usually assume there are multiple samples per person (MSPP) available during the training phase for discriminant feature extraction. Figure 1 shows a manifold surface of a face space projected by amount of samples of a person (in red). However, in many practical applications such as person re-identification in camera network, surveillance photo verification, forensic identification etc., this assumption may not hold as it is usually difficult to collect adequate samples. Within insufficient samples, the learned feature space is very likely to overfit to the small training data. This is the *'small-sample-size'* (SSS) problem [10], interpreted by the fact that the number of samples per person in the probe set is much larger than that in the galley set. In Fig.1, the blue manifold surface represents the learned face space from limited number of samples of a subject, which suffers from overfitting severely. An extreme case of SSS is the *single sample per person* (SSPP), when only one sample per subject is enrolled or recorded in such systems as law enhancement, e-passport and ID card identification etc. Under such circumstances, the performance of conventional methods significantly degrade, because no enough samples can be used to shape an adequate face space.

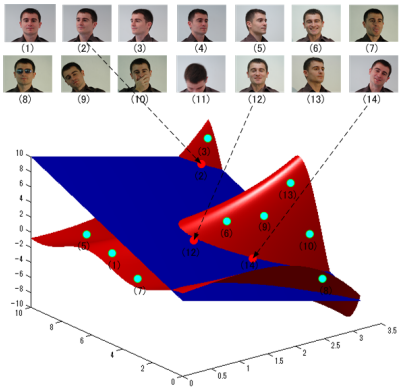


Figure 1: The demonstration of two manifold surfaces: the red one represents a full face space projected from all the samples of one person; and the blue one is a face space learned when only three samples are available.

1.2 Ensemble learning for SSS face recognition

To address the SSS problem, there have been many attempts in the literature, a machine-learning technique known as ensemble learning has received considerable attentions in pattern recognition community. A key benefit of ensemble is the base classifier collaboration based on voting where the overall classification ability is greater than a single classifier. Ensemble rules such as boosting, bagging, random forest, etc [14]. are based on this idea that a pool of different classifiers can offer complementary information for classification. Given a set of base classifiers $B = \{b_1, b_2, ..., b_i, ..., b_L\}$, the goal of these methods is to compose an ensemble E from B to achieve higher recognition accuracy. According to the definition of base classifiers, existing ensemble methods are mainly classified into three categories: (1) decision tree based ensemble [6, 17]; (2) random subspace, semi-random subspace [16, 19]; (3) image partitioning based sub-pattern ensemble [7, 8, 15].

In the fist category, large number of decision trees (DT) are constructed by randomly selecting features from the feature space to define base classifiers. Ensemble rules are exploited on decision trees to make the final decision. A representative method [6] employs a random forest to learn discriminant features to deal with face recognition with image occlusions. In [17], five decision tree pruning methods outperform the Bagging, Boosting and Error-Correcting Output Code (ECOC). DT based methods perform well under MSPP face recognition but degrade in the SSS face recognition dramatically. The major reason is that base classifiers constructed from the feature space are highly depending on the original training sample set. In the SSS problem, insufficient samples supply limited discriminant information to generate diverse base classifiers.

In the second category, the idea of subspace is introduced to ensemble. Since [14] shows that the strong and stable base classifiers defined by subspace algorithms (e.g. LDA) are not suitable to ensemble rules such as boosting, bagging, etc. To overcome this problem, random subspace (RS) [16] was proposed to generate weak but diverse base classifiers. In [16], by random sampling on feature vectors in the PCA subspace, multiple Fisherface and N-LDA classifiers were constructed and the two groups of complementary classifiers were integrated using a bagging based fusion rule. Since RS just focuses on the global rather than local extraction of features, local discriminant information are not guaranteed. Motivated from this finding, a revised version named semi-random subspace (Semi-RS) was proposed in [19]. Different from RS, the Semi-RS randomly samples features on each local region partitioned from the original face image. Although Semi-RS has achieved success in dealing with local deformations e.g. facial expressions, it does not work well on the data with large scale variations in illumination, head poses etc. The reason is that local regions are too small to cover large variations. Furthermore, since the PCA subspace is also generated from the original training data, the feature selection of RS and Semi-RS also suffers from the SSS problem.

For the third category, each face image is first partitioned into several sub-patterns. Discriminant learning techniques are then applied on each of them. Finally ensemble rules are used for fusion. An early attempt [8] divided each face image into six elliptical sub-regions and learned a local probabilistic model for recognition. Topcu et al. [15] proposed an alternative way of partitioning face regions into equal-sized small patches. Features extracted from each patch are classified separately and the recognition results are combined by a weighted sum rule. The problem is that these methods ignore the geometric information during feature extraction, that is the coorelationship between local patches. Literature [7] shows it is unlikely to model the whole face region accurately by a simple distribution on separate local patches such as nose, mouth etc.

All the methods mentioned above share the same problem in the content of SSS that the species of base classifiers learned from insufficient training data are not diverse enough to represent an adequate face space. This suggests that sufficient various training samples play a significant role in recognition tasks. Diversity is an important parameter of ensemble which measures the disagreement of the outputs of base classifiers. High diversity assures that different base classifiers make different decisions to report different errors. Base classifiers generated from insufficient training data are more likely to be tightly correlated and make similar errors. They have limited discriminant power to deal with massive variations in the test data. The requirement for appropriately defined base classifiers significantly restricts the applicability of current ensemble algorithms for SSS face recognition.

1.3 The contributions of this work

This work proposes a novel Random Quad-Tree based ensemble algorithm (R-QT). Advantages over conventional ensemble methods for SSS face recognition are at two aspects: (1) the governing architecture yields more new samples per person to generate high diverse base classifiers which expands the face space; (2) the base classifiers generated from face images with overlapping regions preserve the geometric coorelationship between local patches.

The rest of this paper is organized as follows. In Section II, we briefly review the proposed R-QT algorithm. Then, in Section III, the theory and algorithm of R-QT based base classifier definition are stated in details. Section IV reports on a set of experiments using five widely-used databases to demonstrate the effectiveness of our method. Finally, conclusions and future directions are summarized in Section V.

2 Overview of the Random Quad-Tree based ensemble algorithm (R-QT)

The diagram in Fig.2 shows the skeleton of the proposed algorithm R-QT. From the original SSS training data, we first generate a template image to encode the distribution of discriminate features across the dataset. However, since the samples in the gallery set is insufficient, the template is of weak representativeness. Thus, we need to introduce new changes to make the template more diverse. We add a set of random matrices to the template to generate multiple templates of high diversity. Then we perform Quad-Tree decomposition on the new templates. The Quad-Tree decomposition is introduced based on literature observations [7, 18] that there is high overlap between different parts of face regions and the coorelationship between local regions plays an important role in classification. Instead of splitting templates into small non-overlapping regions, Quad-Trees decompose templates into overlapping patches. According to the decomposition results, face images in the original training data are reorganized to generate new samples. Lastly, we use a majority voting scheme to aggregate the base classifiers. Since the Quad-Tree decomposition is performed on a set of randomly generated templates. We call our method as Random Quad-Tree based algorithm.

3 Random Quad-Tree based base classifier definition

Random Quad-Tree based base classifier definition, demonstrated by Fig.3, mainly consists of three operations: (1) gen-

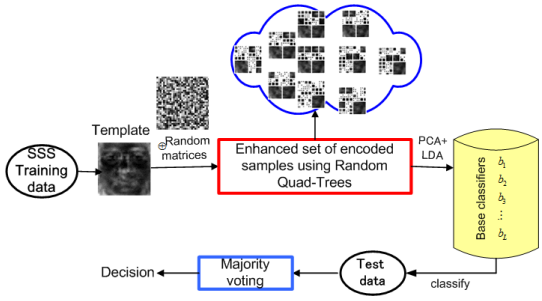


Figure 2: Illustration of the Random Quad-Tree based ensemble algorithm.

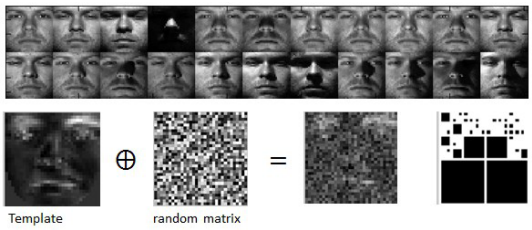


Figure 3: An example of Random Quad-Tree decomposition on a face database: Extended Yale database.

eration of a template image T; (2) Random Quad-Tree decomposition, and (3) the evaluation of each base classifier.

3.1 Template image

Motivated by the idea of LDA which encodes discriminant information by maximizing the between-class scatter matrix S_b and minimizing the within-class scatter matrix S_w (See Eq.(1)). We define a template face T by Eq.(2) to represent the distribution of discriminant information across the database. Thus, the total variance of the entire database is the variance of the template.

$$\begin{cases} S_b = \sum_{i=1}^{c} N_i(\mu_i - \mu)(\mu_i - \mu)^T, \\ S_w = \sum_{i=1}^{c} \sum_{x_k \in X_i} (x_k - \mu_i)(x_k - \mu_i)^T, \end{cases} \quad (1)$$

$$T = diag(\frac{S_b}{S_w}), \quad (2)$$

where c is the number of classes in the dataset, μ is the mean image of all classes, μ_i is the mean image of class X_i, N_i is the number of samples in class X_i, and x_k is the k-th sample of class X_i. Please note that the generation of T fails under the SSPP situation since no samples available to construct S_w. Thus, the template image is defined by S_b only under SSPP.

3.2 Random Quad-Tree decomposition

This part is at the very core of the definition of a set of base classifiers $B = \{b_1, b_2, ..., b_L\}$. It is done through performing Random Quad-Tree decomposition L times. Following describes the procedure in details.

Since the template image T generated from the original SSS training data is of weak ability to represent the whole face space, we create multiple new templates $T' = \{T'_1, T'_2, ..., T'_L\}$ by adding a set of random matrices $R = \{R_1, R_2, ..., R_L\}$ to T to expand the represented face space. Each random matrix R_i has the same size as T and the elements of R_i are randomly chosen from a even distribution. R_i is first normalized to the range of [0, 255] using Eq.(3).

$$R_i = normalize(R_i) = \frac{R_i - max(R_i)}{max(R_i) - min(R_i)} * 255, \quad (3)$$

$$T'_i = \frac{T + \alpha R_i}{T + R_i}. \quad (4)$$

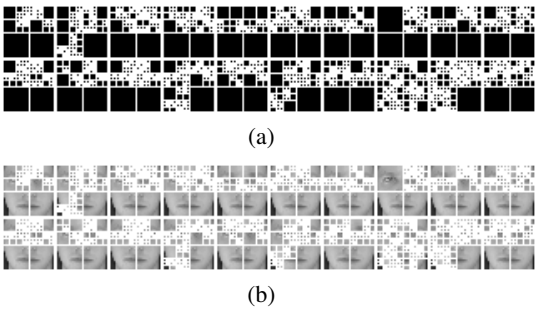
After normalization, R_i is added to T based on a weighted sum rule using Eq.(4) to generate a new template image T'_i, where α is the balance parameter in a range of [0, 2]. T'_i is also normalized to [0, 255]. Quad-Tree decomposition is implemented on each template image T'_i as below.

We perform Quad-Tree decomposition on each T'_i to partition the template into smaller blocks recursively according to a function $doSplit(r)$ defined in Eq.(5) [18]. If the variance of a region r (begins with the template T'_i) is higher than a threshold variance ($T_v * totalVar$), then r is split into four sub-blocks with the same size. The threshold T_v is set to be a default value of 0.5 and the $totalVar$ is defined by the variance of T'_i. The partition carries on until no blocks satisfy the certification function in Eq.(5) or no blocks can be splitted. Usually, the template is split into less and bigger blocks when T_v is large, but into more and smaller blocks when T_v is small.

$$doSplit(r) = \begin{cases} true, while(var(r) > T_v * totalVar), \\ flase, otherwise, \end{cases} \quad (5)$$

$$totalVar = variance(T'_i),$$

After decomposition, we get a template face encoding pattern



(a)

(b)

Figure 4: An example of 20 Random Quad-Trees (a) Quad-Trees (b) Quad-Tree partitions on sample face images

as shown in Fig.(4) (a). Each Quad-Tree decomposition refers to such a pattern. Since larger blocks implies that the density of discriminate features in them are low, these blocks are of no need to keep their original sizes. We then resize them to $((d/2)\times(d/2))$ where d is the dimension of the block (in pixel). Quad-Tree decomposition and block resizing generate a new

face sample whose size is smaller than the original face image. In this way, all face images in the original training set are newly represented according to each encoding pattern to generate a new training sample set. A base classifier b_i is learned from this new sample set.

3.3 Evaluation of each base classifier

Individual base classifiers are trained on a set of virtually generated gallery data based on PCA+LDA. Each gallery set ($X^{\mathbb{G}} = \{x_i^{\mathbb{G}}; i = 1, 2, ..., N * M_G\}$) has the same number of face images as the original training data, where N is the number of subjects and M_G is the number of samples per subject in $X^{\mathbb{G}}$. Evaluation on the original probe set ($X^{\mathbb{P}} = \{x_j^{\mathbb{P}}; j = 1, 2, ..., N * M_P\}$) reports the performance of each base classifier. In the evaluation, we use the nearest neighbor classification with L_2-norm as the distance metric. The distance between two feature vectors, $X = (x_1, x_2, ..., x_n)$ and $Y = (y_1, y_2, ..., y_n)$ is defined in Eq.(6). Given an unknown sample j, a nearest neighbor classifier defined in Eq.(7) searches the feature subspace for the training sample i that is closest to the unknown sample, where $i = 1, ..., N*M_G, j = 1, ..., N*M_P$.

$$d_2(X, Y) = \|X - Y\|_2 = \sqrt{\sum_{k=1}^{n}(x_k - y_k)^2}. \quad (6)$$

$$i = argmin(d_2(X_i, Y_j)). \quad (7)$$

The accuracy of each classifier is defined by the number of correctly classified probe samples against the total number of probe samples, namely as the rank-one recognition rate.

Finally, we briefly analyze the computational complexity of R-QT, which involves L base classifiers. R-QT mainly includes a template image calculation using LDA, Quad-Tree decompositions on template images, and the evaluation of each base classifier using PCA+LDA. Suppose a training set contains m samples, and the size of each face image is d. Both the calculation of T and Quad-Tree decomposition can be performed in linear time $O(d)$. The classical PCA requires around $O(d^3+d^2m)$ computations [12]. And LDA needs $O(mnt+t^3)$ [2], where n is the number of features and $t = min(m, n)$. Since the feature dimension is usually smaller than that of the original face image, we have $d > n$. Thus, the total computational complexity of our method is $O(L(d^3 + d^2m))$.

4 Experiments

We evaluated our method (R-QT) on five widely-used databases, namely: ORL [11], Yale [1], Extended Yale (Yale2) [5], PIE [13], and color FERET [9]. We compared the proposed R-QT with existing several representative methods. The following are the details of the experiments and results.

4.1 Five databases used in the experiments

- ORL database: contains the images from 40 subjects, with 10 different images per subject. Images were taken at different conditions. They are various on facial expressions, open or closed eyes, smiling or non-smiling, with glasses or no glasses and scale changes (up to 10 percent). Moreover, the images were taken with a tolerance for tilting and rotation of the face of up to 20 degree.
- Yale database: consists of 165 faces images of 15 subjects, each providing 11 different images. The images are in upright, frontal position under various facial expressions and lighting conditions.
- Extended Yale database: contains more than 20,000 single light source images of 38 subjects with 576 viewing conditions (9 poses in 64 illumination conditions).
- PIE database: contains 41,368 images of 68 people, each person with 13 different poses, 4 different expressions, and under 43 different illumination conditions. That is, each person has 170 samples.
- FERET database: consists of 13539 images corresponding to 1565 subjects, which are various on facial expressions, ethnicity, gender and age.

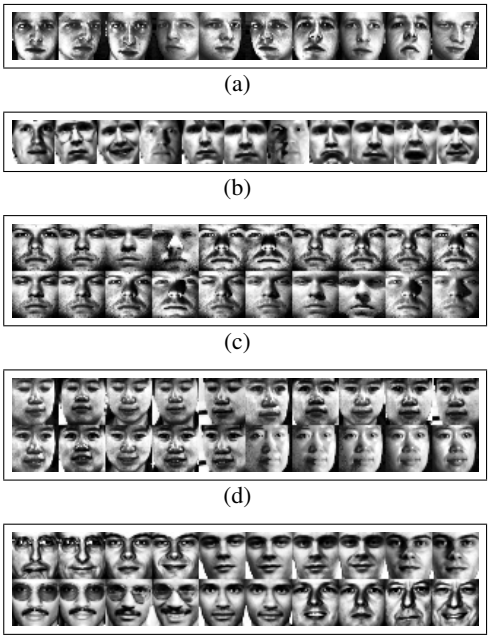


Figure 5: Five databases used in the experiments: (a) ORL, (b) Yale, (c) Extended Yale (Yale2), (d) PIE, (e) FERET.

Face images in the first four databases are aligned and normalized to 32×32 using the method in [3]. The FERET database is normalized and preprocessed to the same size using the *CSU Face Identification Evaluation System 5.1* [4]. Sample images of all databases after histogram equalization are shown in Fig.5.

4.2 Experimental setting

The SSS face recognition is conducted on all databases compared with four typical methods including: decision-tree based

ensemble (DT) [14], random subspace method (RS) [16], semi-random subspace method (Semi-RS)[19], and the decision fusion for patch based method (DF) [15]. We implemented DT and RS ourselves and tuned the best performance for a fair comparison. Recognition results of Semi-RS and DF on part of databases are referred to [19], [15] directly.

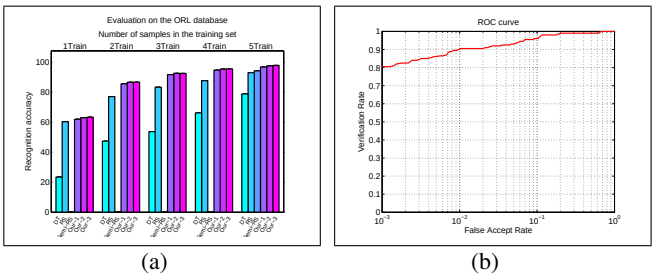
For SSS problem, a random subset with p images per subject is taken with labels to form the training set. For ORL and Yale databases, $p = 1, 2, 3, 4, 5$; For Extended Yale database, $p = 5, 10, 20$; and for PIE database $p = 5, 10$. Face the FERET database two cases are taken into account where $p = 1$ and $p = FaFb$ (each person has two types of images, Fa and Fb, Fa as training data, Fb as test data), respectively. For each given p, there are 10 randomly splits over each database: for each split, the number of p samples of each person are randomly selected as training data, and the rest as test data. The average recognition accuracy over 10 splits is reported on each database. In the experiments, three parameters of our method L, α, T_v are set to 20, 0.5, 0.5, respectively. And the feature dimension of PCA subspace is set to a smaller value between the number of classes in each database and 100.

4.3 Comparison with existing methods

Table $1 \sim 5$ tabulates the rank-one recognition rate of all methods on five databases. In these tables, the fist column lists the methods to be compared. The values in rest columns represent the recognition accuracy of these methods using p samples per person as training data. Due to the space limitation, we just adds a figure consisting of two sub-figures to the first table for a better illustration: (a) the accuracy histogram of each method, (b) the ROC curve of our method.

Table 1: Evaluation on the ORL database

Method \ p	1	2	3	4	5
DT	23.47	47.41	53.75	66.25	78.90
RS	60.39	77.13	83.39	87.75	93.05
Semi-RS	/	/	/	/	94.30
Our	**62.00**	**85.66**	**91.79**	**94.75**	**96.85**



(a) (b)

Figure 6: Evaluation on the ORL database: (a) the accuracy histogram of different methods, (b) ROC of our method.

Before the performance comparison, we first illustrate the effect of the number of samples in the galley set by measuring

Table 2: Evaluation on the Yale database

Method \ p	1	2	3	4	5
DT	22.07	32.44	40.25	45.14	44.67
RS	**34.80**	40.30	47.25	56.00	44.67
Semi-RS	/	/	/	/	69.1
Our	33.80	**50.22**	**60.67**	**68.48**	**70.22**

Table 3: Evaluation on the Yale2 database

Method \ p	5	10	20
DT	69.97	90.79	97.77
RS	84.28	95.76	**99.11**
Our	**85.25**	**95.26**	98.92

Table 4: Evaluation on the PIE database

Method \ p	5	10
DT	48.60	69.55
RS	62.87	79.42
Our	**70.40**	**84.68**

Table 5: Evaluation on the color FERET database

Method \ p	1	FaFb
DT	13.01	25.60
RS	34.99	78.62
DF	42.56	/
Our	**58.02**	**84.17**

the performance on all databases. As shown in Table $1 \sim 5$, the almost all gallery sets with relatively larger p give much better performance, especially the $p \geqslant 3$ in Table 1 and $p \geqslant 10$ in Table 3 give more than 90% recognition accuracy. However, the performance of almost all methods degrade obviously as the number of samples in the gallery set decreases, the results met our expectation that the SSS problem challenges face recognition algorithms.

Above observation also proves a fact that the decision tree (DT) based methods obtain the worst performance on all databases. The reason is that base classifiers generated by DT come from the original gallery set directly. The SSS problem makes them highly overfit to the training data. Inadequate base classifiers are unlikely to estimate the face space accurately, thus their performances drop rapidly (even lower than 20%) when probe sets have large variations. This observation implies that the key point using ensemble to solve the SSS problem is to enlarge the gallery set by introducing more variations so that the learned base classifiers are diverse enough to estimate the

face space.

In Table 1, 2, we find our method slightly outperforms RS under $p = 1$ while the probe set is small such as ORL and Yale databases. But in large databases such as Yale 2, PIE, and FERET with rather more variations, as shown in Table $3 \sim 5$ our method outperforms RS dramatically. For instance, the accuracy of our method on database PIE has $5\% \sim 10\%$ increases than RS. The reason is: RS projects face images to a low dimensional feature subspace and then randomly selects certain features as discrimination for recognition. The main advantage of RS is that the discrimination yielded based on random feature selection are more diverse than involving all features as DT does. However, the generation of the feature subspace is highly dependent on the variations in the gallery set. Thus, RS has good performance on small datsets where gallery data posses reasonable variations against probe set, but degrades on large datasets with much more variations in probe set.

In addition, since RS selects features from global rather than local face images, local discriminant information are not guaranteed. For instance in Table 2, 4, the performance of RS is low on Yale and PIE database which contain amount of local deformations such as facial expressions, glasses or no glasses, etc. To overcome this, local sub-region based methods were proposed such as the semi-random subspace (Semi-RS) [19] and the decision fusion for patch based method (DF) [15]. From Table 2 and Table 5, we can see that Semi-RS and DF outperform RS, respectively. However, since these methods partition face image into small patches, and learn a base classifier from each single patch. The discrimination extracted from single patch is limited and the coorelationship between different patches are lost. In contrast, usage of the Quad-Tree structure in our method preserves these geometric information and explore them to contribute to recognition. Comparing the performance of Semi-RS, DF and our method in Table 1, 2, 5, R-QT always performs better.

5 Conclusion

To address the SSS problem, we propose a novel Random Quad-Tree based ensemble algorithm (R-QT) to estimate a set of base classifiers which encodes both meaningful discriminant features and geometric information across the face region and explore the possible face space by generating multiple new samples. Compared with conventional methods, more diverse and representative base classifiers are generated from the enlarged training data. As evaluated on five well-known databases, R-QT can estimate a more accurate and robust ensemble for SSS face recognition. Besides the SSS face recognition, our method can also be extended to other SSS scenario with limited training data, which appears to be another interesting direction of future work.

References

[1] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on PAMI*, 19(7):711–720, 1997.

[2] D. Cai, X. He, and J. Han. Training linear discriminant analysis in linear time. In *24th IEEE International Conference on Data Engineering (ICDE)*, pages 209–217, 2008.

[3] D. Cai, X. He, Y. Hu, J. Han, and T. Huang. Learning a spatially smooth subspace for face recognition. In *CVPR*, 2007.

[4] CSU. CSU face identification evaluation system, 2003.

[5] A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Transactions on PAMI*, 23(6):643–660, 2001.

[6] V Ghosal. Efficient face recognition system using random forests. Master's thesis, Indian Institute of Technology Kanpur, 2009.

[7] J. Lu, Y. P. Tan, and G. Wang. Discriminative multi-manifold analysis for face recognition from a single training sample per person. In *ICCV*, 2011.

[8] A. M. Martinez and A. C. Kak. PCA versus LDA. *IEEE Transactions on PAMI*, 23(2):228–233, 2001.

[9] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss. The FERET evaluation methodology for face-recognition algorithms. *IEEE Transactions on PAMI*, 22(10):1090–1104, 2000.

[10] S. J. Raudys and A. K. Jain. Small sample size effects in statistical pattern recognition: Recommendations for practitioners. *IEEE Transactions on PAMI*, 13(3):252–264, 1991.

[11] F. S. Samaria and A. C. Harter. Parameterisation of a stochastic model for human face identification. In *Proceedings of 2nd IEEE Workshop on Applications of Computer Vision*, pages 138–142, 1994.

[12] A. Sharma and K. K. Paliwal. Fast principal component analysis using fixed-point algorithm. *Pattern Recognition Letters*, 28(10):1151–1155, 2007.

[13] T. Sim, S. Baker, and M. Bsat. The CMU pose, illumination, and expression database. *IEEE Transactions on PAMI*, 25(12):1615–1618, 2003.

[14] M. Skurichina and R. P. W. Duin. Bagging, boosting and the random subspace method for linear classifiers. *Pattern Analysis and Applications*, 5(2):121–135, 2002.

[15] B. Topcu and H. Erdogan. Decision fusion for patch-based face recognition. In *ICPR*, pages 1348–1351, 2010.

[16] X. Wang and X. Tang. Random sampling for subspace face recognition. *IJCV*, 70(1):91–104, 2006.

[17] T. Windeatt and G. Ardeshir. Decision tree simplification for classifier ensembles. *Int. J. Patt. Recogn. Artif. Intell.*, 18:749–776, 2004.

[18] C. Zhang, X. Liang, and T. Matsuyama. Multi-subregion face recognition using coarse-to-fine Quad-tree decomposition. In *ICPR*, pages 1004–1007, 2012.

[19] Y. Zhu, J. Liu, and S. Chen. Semi-random subspace method for face recognition. *Journal of Image and Vision Computing*, 27:1358–1370, 2009.