Face Recognition from Images with High Pose Variations by Transform Vector Quantization

Amitava Das, Manoj Balwani[1], Rahul Thota[1], and Prasanta Ghosh[2]

Microsoft Research India. Bangalore, India
amitavd@microsoft.com

Abstract. Pose and illumination variations are the most dominating and persistent challenges haunting face recognition, leading to various highly-complex 2D and 3D model based solutions. We present a novel transform vector quantization (TVQ) method which is fast and accurate and yet significantly less complex than conventional methods. TVQ offers a flexible and customizable way to capture the pose variations. Use of transform such as DCT helps compressing the image data to a small feature vector and judicious use of vector quantization helps to capture the various poses into compact codebooks. A confidence measure based sequence analysis allows the proposed TVQ method to accurately recognize a person in only 3-9 frames (less than ½ a second) from a video sequence of images with wide pose variations.

1 Introduction

Pose and illumination variations are the most dominating challenges in face recognition and have been the focus of many studies in the past [3][9][8][7]. Holistic face recognition methods such as the PCA-based Eigenface method [6] perform nicely for image sets where pose variation is minimal and performs rather poorly when there are wide variations of pose (e.g. the MSRI-V1 database shown in Fig. 1).

Even if the pose variation is minimal, e.g. for frontal images as shown in Fig. 2, there may still be a wide variation of expressions, which also limits the performance of several traditional face recognition approaches such as PCA. To overcome the challenges of pose and expression variations, several 2D and 3-D pose-normalization methods have been proposed, as nicely surveyed in [3]. However, in the same survey, it has been noted that most of these methods (which can handle pose variations) are quite complex in terms of computation complexity and memory usage.

Several recent studies [5][11][4][12][13] have shown that face recognition performance improves dramatically if a face video, or a sequence of face images of a person, is used for recognition, as opposed to using a single image. Several spatial-temporal methods such as [10], also reported good results. A good review can be found in [3]. Once again, due to the complex modeling employed to handle pose variations, these methods also require high computational complexity and the processing of a reasonably large sets of image frames, before reaching a decision.

1) IIT-Madras MTech & 2) IISc-Bangalore MS students, doing internship at MSR-India.

P. Kalra and S. Peleg (Eds.): ICVGIP 2006, LNCS 4338, pp. 674 – 685, 2006.

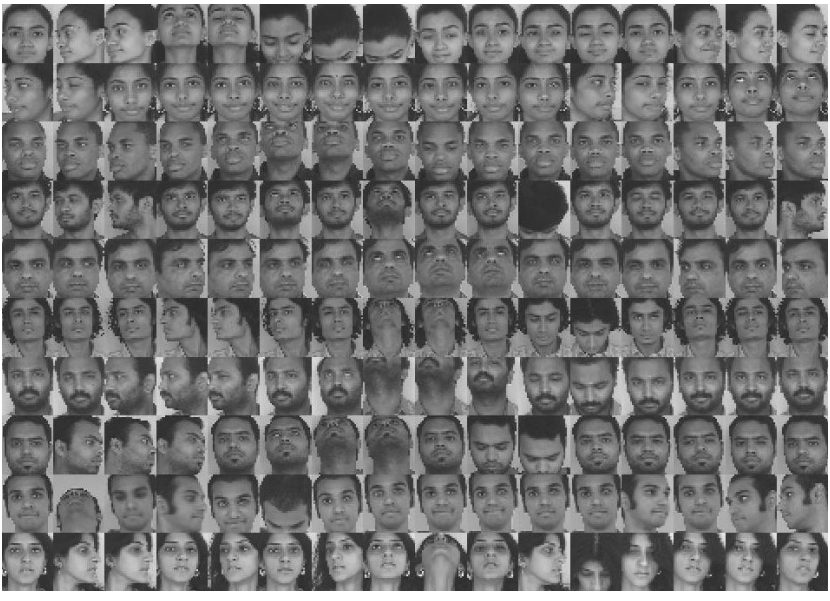

Fig. 1. Pose Variations in the MSRI-V1 Face Database

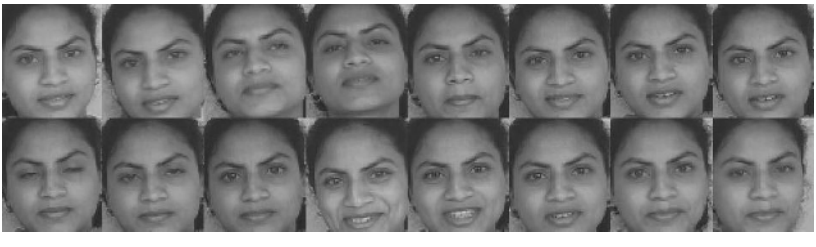

Fig. 2. Example of expression variations (MSRI V2 database of frontal face images)

We present a novel low-complexity transform vector quantization (TVQ) based face recognition method which handles pose variation quite efficiently and to some extent can handle illumination variation as well. The use of Discrete Cosine Transform [2] in our feature extraction delivers a significant amount of dimensionality reduction at much lower complexity than PCA. Our main contribution however is the application of vector quantization [1] in the transform domain, which allows the TVQ system to capture the pose variations of a person very effectively as multiple code vectors of a person-specific codebook (see Fig. 6). The proposed TVQ method gives a customizable platform, which can be tailored to work well with any kind of face data – with or without pose variations. For image data with little pose variation, TVQ demonstrates 100% identification accuracy with a very small codebook (size 4 and dimension 15) per person, compared to the 93% identification accuracy offered by PCA, using a 416 dimension feature vector per person. We also present a confidence-measure based fast face recognition from video, which delivers 100% accurate results with high confidence very quickly [processing only 3-9 frames

or in about 400 ms]. Finally, a set of illumination-variation trials indicate that the proposed TVQ is also quite robust to illumination variation. The computational complexity and memory requirements of TVQ is significantly less than traditional PCA based methods [6][7].

2 Face Recognition by Transform Vector Quantization

The proposed Transform vector quantization (TVQ) face recognition method exploits the de-correlating property of the Discrete Cosine Transform [2] to extract a low-dimension feature vector from the input face image. For classification, we use a Vector Quantization [1] based scheme. A confidence-measure based fast-selection process further reduces the complexity and speeds up the face recognition process, while guaranteeing 100% accuracy. Details of the pre-processing, feature extraction and classification steps are presented next.

2.1 Preprocessing, Feature Extraction and Face-Transform Space in TVQ

The input images frames $<\mathbf{X}_1, \mathbf{X}_2, .. \mathbf{X}_k, .. \mathbf{X}_M>$ from the video are converted into gray images. From each frame, a N1 x N2 size image is cut (in our experiments we kept N1=136 and N2 = 120) from the region of interest (ROI) to create images $<\mathbf{I}_1, \mathbf{I}_2, .. \mathbf{I}_k, .. \mathbf{I}_M>$. The ROI is automatically extracted by tracking the face contour in the image using vertical and horizontal projections of selected areas of the image. Figure 2 shows the pre-processing step and resulting processed gray images.

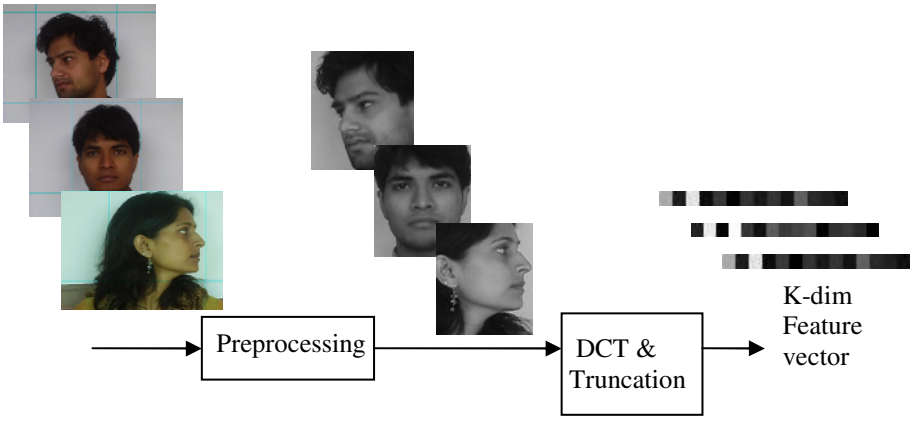

Fig. 3. Preprocessing and feature vector extraction in the proposed TVQ method

A 2-D DCT is then applied to the gray image $\mathbf{I}_k$ and the coefficients of the top left **m x m** corner of the transform coefficient matrix **C** is selected to form a feature vector $\mathbf{F}_k$ of size K=($\mathbf{m}^2$ -1). Note that all the components, except the 1st one (which is the DC coefficient C(0,0)), are chosen. Removal of the DC coefficient makes the proposed TVQ method somewhat immune to illumination variation.

This way a large amount of the face image information is stored into a significantly low dimension feature vector. In our experiment, our input image size was 280x320 or 89600 pixels and we used feature vectors having dimensions ranging from 9 to 63. Fig. 4 shows images reconstructed after such truncation of DCT coefficients for various values of K. Note that here the objective is not to preserve the quality of the input image but to pack important discerning attributes of the face image into a small feature vector. Thus the images reconstructed from the truncated DCT coefficients (our feature vector) do not represent as much spatial detail as the original image, but as we will see later, they are quite successful in differentiating the face images of different persons. This feature vector dimension $\mathbf{K}= (\mathbf{m}^2 -1)$ will be a system parameter in the proposed TVQ face recognition method.

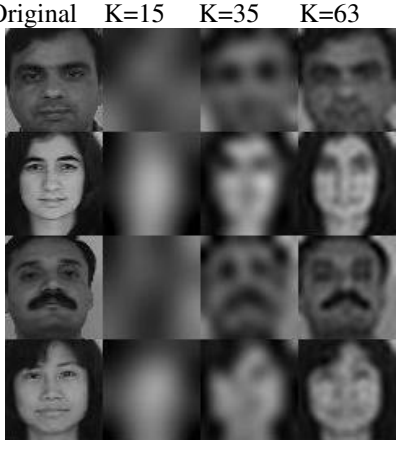

Fig. 4. Impact of the size of K in reconstructing back the image

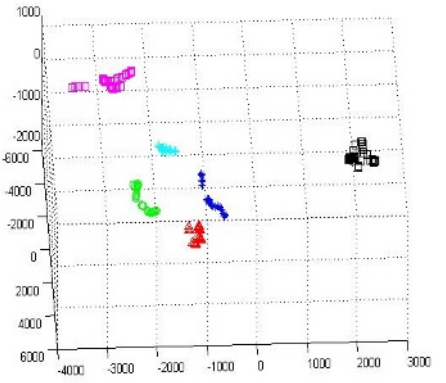

Fig. 5. Person clusters in the face transform space (K=3, 6 person system)

2.2 Recognition of Face by Transform Vector Quantization

In the TVQ method, the input face images are converted into the K-dimension feature vectors comprising of DCT coefficients. This can be interpreted as follows. The face images of all the T persons to be recognized, P_1, P_2, to P_T, are now lying in this K-dimension "face-transform" space. The face images (various poses) of each person will be occupying a certain "region" of this face-transform space. Vector quantization [1] can now be used to design a person-specific codebook, which by a proper Voronoi tessellations of this K-dimension face-transform space, will be able to capture and define the regions of various persons. Fig. 5 shows the results of a toy example, in which TVQ is applied to the face images of 6 different persons. Here K is chosen to be only 3 so that we can plot the "person regions" in the 3-D face transform space. Six distinct clusters, one for each person, are clearly seen in Fig. 5.

All the various face images of the i-th person, P_i, will now be lying within the i-th region in the face-transform space and can be represented by an N-size VQ codebook, $\mathbf{CB_i = [C_{i1}, C_{i2}, \ldots, C_{ij}, \ldots C_{iN}]}$, each $\mathbf{C_{ij}}$, being a K-dimensional code vector. The

Input set of face images of various pose

Same images encoded & clustered by TVQ

Fig. 6. Capture of the various poses of a person's face by the TVQ codebook. When a set of face images of various poses of a person (shown at the top collage) is encoded by the TVQ codebook of the same person, the different poses get automatically clustered as shown on the bottom collage. In the bottom collage, all the images in the i-th row is closest to the i-th vector of the TVQ codebook of the person. K=15 and N=8 are the codebook parameters for this example. This demonstrates that in TVQ, the different code-vectors of a person do capture the different poses of the person. Note that in the bottom collage, a maximum of 12 images per row is shown, i.e. some code-vectors (e.g. 5,6 and 8) has more than 12 images closest to them.

various poses of the i-th person will be captured by the various code vectors of this codebook C as shown in Figure 6. Here, the face images, which are "closest" to each of the 8 code vectors of the codebook of a certain speaker are shown. As seen here, the different code vectors do capture the pose variations of the same person. Such

codebooks, one for each person, can be designed by any VQ training algorithm such as LBG [1]. A large number of pose variations can be captured by a reasonably-sized codebook. Thus, representation of each person, in our proposed TVQ method, is much more "richer" than traditional PCA based approach [6] in which only one representative face data, the M-dimensional weight vector, represents the person.

Given an input face image, the recognition task by TVQ then becomes the search for the best codebook, $\mathbf{CB_{j*}}$, which is "closest" to the feature vector **F**, extracted from the input image. We present 3 TVQ algorithms next: a) person identification from a single face image, b) person verification from a single image, and c) person identification from a video.

Algorithm 1: Person Identification by TVQ

a) Given an input image **X**, extract the transform feature vector **F**.
b) For each person, P_i, find D_i the closest distance of its codebook $\mathbf{CB_i}$ from the input feature vector **F**:

D_i = minimum of D_{ij}, where $D_{ij} = \|\mathbf{F} - \mathbf{C_{ij}}\|^2$, j=1,2,…,N, $\mathbf{C_{ij}}$ being the code vector of the codebook $\mathbf{CB_i}$

c) The identified person is person P_k, where D_k is the minimum of all D_i, i=1,2,3...T.

Algorithm 2: Person Verification by TVQ From a Single Image

For person verification, a face image **X** is presented along with an "identity claim" k. The task is now to verify whether the image belongs to the k-th person P_k, or not. The TVQ verification algorithm is given below:

1) Given an input face image **X**, extract the transform feature vector **F**.
b) Given the identity claim, k, compute two distances, d_trgt & d_bkgr, as follows:

d_trgt = minimum distance of the input feature vector from the codebook $\mathbf{CB_k}$ of claimed person P_k; In other words, d_trgt = minimum of D_{ij}, where $D_{ij} = \|\mathbf{F} - \mathbf{C_{ij}}\|^2$, j=1,2,…,N, $\mathbf{C_{ij}}$ being the code vector of the codebook $\mathbf{CB_k}$ of the target person P_k.

d_bkgr = minimum distance of the input feature vector from the collection of all codebooks of all other persons except person P_k. In other words, d_bkgr = minimum of D_{ij}, where $D_{ij} = \|\mathbf{F} - \mathbf{C_{ij}}\|^2$, j=1,2,…,N; i=1,2,3,..T ; i not equal to k, where $\mathbf{C_{ij}}$ is the i-th code vectors of j-th codebook.

c) Compute a confidence measure λ = d_trgt/d_bkgr, and if $\lambda < \theta$ (θ being a predetermined threshold during training) then the presented identity claim is accepted as person P_k; else it is rejected.

Algorithm 3: Person Identification by TVQ from a Video

In contrast to earlier methods of face recognition from video, which require processing of a reasonably large set of such frames before making a decision, TVQ offers a much faster and less complex method as described below:
a) Given input image sequence $\mathbf{X_1}$, $\mathbf{X_2}$,...$\mathbf{X_M}$, extract feature vectors, $\mathbf{F_1}$, $\mathbf{F_2}$,...$\mathbf{F_M}$
b) At each k-th sequence, for each person, P_i, compute an accumulated distance (AD) A_i as follows:

$A_i(k) = D_i(1) + D_i(2) + D_i(3) + \ldots\ldots + D_i(k)$, where $D_i(k)$ is the minimum distance of the feature vector $\mathbf{F_k}$ to the codebook $\mathbf{CB}_i$ of person P_i

c) If $\mathbf{A_j}$ is the least among all j=1,2,…T, then compute a confidence measure β_j, which is the difference ($\mathbf{A_j} - \mathbf{A_m}$) between the best candidate $\mathbf{A_j}$ and the next best candidate (say $\mathbf{A_m}$). If this confidence measure β_j is greater than a pre-determined threshold τ_j (learned during training), then the j-th person is chosen as the identified person. τ_j is calculated by computing the histogram of β_j with the training data and then setting τ_j as a fraction of the mean value of β_j .

We call the number of frames TVQ takes to detect a person as the Time-of-Detection or ToD, which is used later in our trials as a performance metric.

3 Face Recognition Tasks, Databases and Description of Trials

At Microsoft Research India, we created our own Biometric Person Recognition Database called the MSRI database by recording face video as well speech from a wide set of people of various nationalities. For face recognition, we used two subsets (Table 1) of the MSRI database: a) MSRI-V1, consisting of face images with various pose variations (Fig. 1) and b) MSRI-V2, with only frontal face images (Fig. 2).

Table 1. Details of the MSRI-V1 & MSRI-V2 Face Databases

	MSRI-V1	MSRI-V2
No of Person in the database	65	52
Average no of Images/Person for Training	324	55
Average no of Images/Person for Testing	325	83
Total number of test trials	16884	5414

The performance metrics used for the tasks of identification, verification and verification from video are Percentage of Accuracy, Equal Error Rate (EER) and Time to Detect (ToD) respectively. Two system parameters, N (codebook size) and K (code-vector dimension), are varied during trials. The PCA based Eigenface [6] method was also ran for comparison and the dimension M of the pattern weight vector $\mathbf{\Omega}$ was varied as the system parameter.

4 Results and Discussions

The results, shown in Table 2&3 for MSRI-V1 and MSRI-V2 respectively, clearly shows the proposed TVQ method outperforms conventional PCA based method significantly when there is high pose variation.

Table 2. Identification accuracy of the TVQ and PCA based face recognition methods on the MSRI-V1 database having face images with high pose variations

TVQ		K=15	K=35	K=63	PCA	M=416	M=240	M=60
	N=4	83.7	86.2	87.0		68.8	68.7	67.6
	N=8	92.4	94.3	94.7				
	N=16	96.9	96.8	97.1				
	N=32	98.8	99.1	99.4				

Table 3. Identification accuracy of the TVQ and PCA based face recognition methods on the MSRI-V2 database having only frontal face images

TVQ		K=15	PCA	M=260	M=60	M=30	M=15
	N=1	90.9		93.9	92.8	90.4	85.7
	N=2	99.0					
	N=4	100.0					

The performance of TVQ increases dramatically if we increase N (number of code-vectors) as opposed to increasing K (feature dimension) as evident in these tables. For only frontal images, TVQ delivers 100% accuracy at N=4 and K=15 or storage of only 60 data points per person and 60*T multiply-add operations, as opposed to 93.9% performance by PCA which requires 260 data points to store per person, requiring 260*T multiply-add operations. The feature extraction process of TVQ is also much simpler (2*P*K multiply-add for a PxP image and K-dim TVQ) than PCA (P*P*M multiply add for a PxP image and M dimension PCA weight vector).

4.2 Person Verification

As seen in Table 4, TVQ is offering an EER of 2.1% for MSRI-V1, which is much better than the PCA figure of 48%. For the MSRI-V2 database, TVQ (N=16;K=15) delivered 1.5% EER. We did not run PCA for this trial.

Table 4. Performance comparison (in terms of Equal Error Rate) of the TVQ and PCA based face recognition methods on the MSRI-V1 database having high pose variations

TVQ		K=15	K=63	PCA	M=416
	N=4	5.2	4.4		48%
	N=16	2.1	2.0		

4.3 Person Recognition from Video

We ran these experiments only on the MSRI-V1 database, as for MSRI-V2 we are always getting the correct detection in the first frame itself.

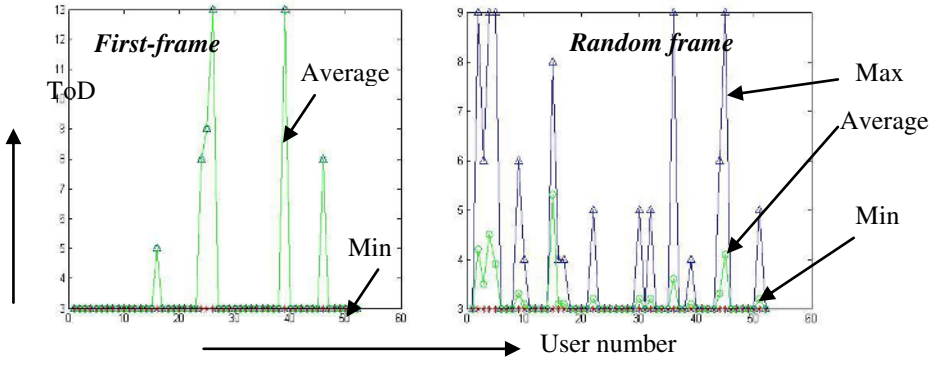

Fig. 7. Time to Detect Statistics of TVQ (n=16 K=15) for MSRI-V1

The goal is to determine how many frames (Time of detection or ToD) does TVQ need to recognize a person with total confidence. We ran two experiments: a) FIRST-FRAME, where the 1st frame is always the starting frame (this was ran 1 time per speaker), and b) RANDOM, here the starting frame is chosen at random, thereby allowing the processing to start from any pose variations (this trial is ran 10 times for each speaker) The results are shown in Fig. 7.

In case of FIRST-FRAME, the worst-case detection time is 13 frames or approx. 400 mill-second, although most persons were detected in 3 frames only (90 ms). For the RANDOM trial it was found that on average (green circle) the detection time is approximately 5 frames or 150 ms, where as the worst case was found to be 9 frames or 270 ms.

4.4 Illumination Variation Trials

In this experiment, we artificially changed the illumination of the test images from 80% to 120% in 3 different ways (as shown in Figure 8) : a) L-X: Change only left half (multiply by X%), b) C-X: change entire image, c) R-X: Change only right half (X=80% or 120%). This reflects somewhat what can happen in a real life situation.

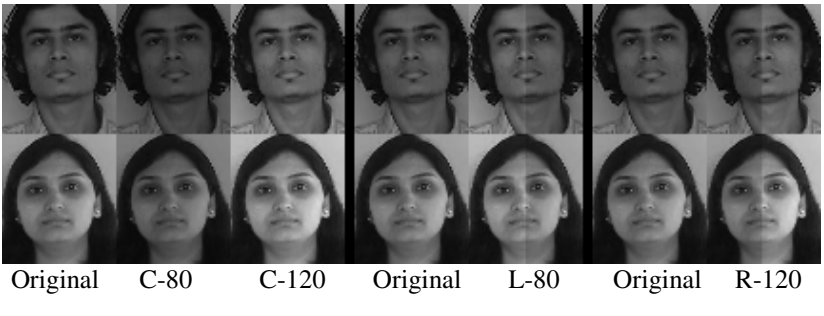

Fig. 8. Example images with 3 types of artificially-created illumination variation

Table 5 shows the performance figures (Identification Accuracy) for various illumination variations and Figure 7 shows the ToD figures for the L-80. For PCA the performance dropped from 68.8% (original) to 63% for R-120.

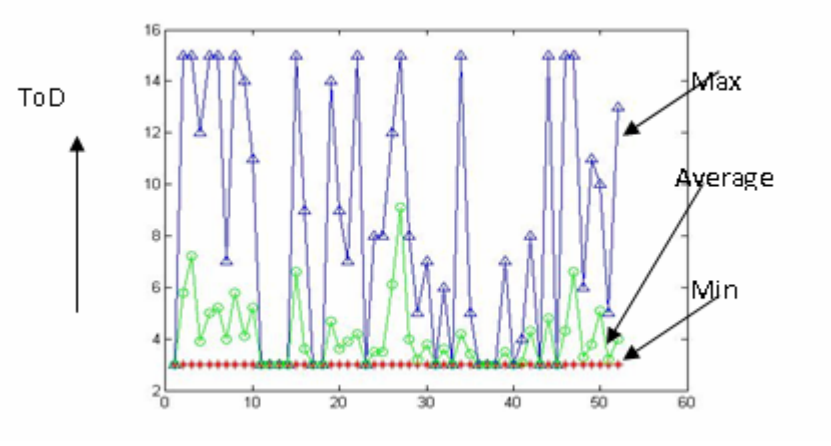

Fig. 9. Time to Detect results for the L-80 illumination variation trial (N=16; K=15)

Table 5. Identification accuracy of TVQ for various illumination-variation trials

	Left			Center			Right		
N x K	L80	Orig	L120	C80	Orig	C120	R80	Orig	R120
4 x 15	74	83.7	80	78	83.7	81	73	83.7	74
16 x 15	89	96.9	96	90	96.9	97	84	96.9	88

As seen in Fig. 9 and Table 5, illumination variation is impacting the identification accuracy of TVQ to some extent but not to a great extent and especially the ToD results are quite good. On average, people are getting detected in less than 7 frame (210 ms) and the worst case detection time is 15 frames or 450 ms for the L-80 condition.

5 Conclusions

We presented a novel transform vector quantization (TVQ) face recognition method which can be tailored to provide high performance for various extents of pose and illumination variations. Use of transform such as DCT in TVQ helps compressing the image data to a small feature vector and judicious use of vector quantization helps to capture the various poses into compact codebooks. The computational complexity of TVQ is significantly less than a conventional PCA based method as shown in Table 6.

Table 6. Comparison of TVQ with PCA for face images with high pose variation

	TVQ	PCA
System Parameters	N=32;k=15	M=416
Identification-Accuracy for MSRI-V1	98.80%	68.80%
EER for MSRI-V1	2.10%	48%
ToD for MSRI-V1-average value	3 frames	-
Feature extraction complexity - 160x160 image	4800	10649600
Classification complexity / per person	480	416
Overall Detection Complexity / per person	5280	10650016
Memory/user(float number to store) / per person	480	416

In the proposed TVQ method, high extent of pose variation can be handled by having more number of code vectors, while more image precision can be obtained by increasing the feature vector dimension. A confidence measure based sequence analysis allows the proposed TVQ method to accurately recognize a person in only 3-9 frames (less than ½ a second) from a video sequence of images with wide pose variations.

References

1. Gersho A. and Gray R.: Vector Quantization and Signal Compression. Kluwer Academic Publishers. (1992)
2. Rao K. and Yip P.: Discrete Cosine Transform – Algorithms, Advantages, Applications. Academic Publisher. (1990)
3. Zhao W., Chellappa R., Rosenfeld A., and Phillips P.J.: Face recognition: A literature survey. ACM Computing Surveys. Vol. 35. (2003) 399 – 458
4. Zhou, S., Krueger V., Chellappa R.: Probabilistic recognition of human faces from video. Computer Vision and Image Understanding. Vol. 91. (2003) 214 – 245
5. 5. Kuang-Chih Lee Ho, Ming-Hsuan Yang J., Kriegman, D.: Video-based face recognition using probabilistic appearance manifolds. Proc. IEEE CVPR (2003) 313 – 320
6. Turk M. and Pentland A.: Eigenfaces for recognition. Journal of Cognitive Neuroscience, 3(1). (1991) 71 – 86
7. Moghaddam B. and Pentland A. : Face recognition using view-based and modular eigenspaces. In : Automatic Systems for the Identification and Inspection of Humans, Vol. 2277. SPIE. (1994)
8. Phillips P., Grother P., Micheals R., Blackburn D., Tabassi E., and Bone J.: Face recognition vendor test 2002: Evaluation report. Technical Report NISTIR 6965, http://www.frvt.org (2003)
9. Yang M., Kriegman D., Ahuja N.: Detecting Faces in Images : A Survey. IEEE Trans. PAMI, vol. 24. (2002) 34 – 58
10. Li Y., Gong S., Liddell H. : Video-based face recognition using identity surfaces. Technical report, Queen Mary, University of London (2001)
11. Biuk Z. and Loncaric S.: Face Recognition from Multi-Pose Image Sequence. In: Proceedings of 2nd Intl. Symposium on Image and Signal Processing (2001) 319 – 324

12. Krueger V. and S. Zhou S.: Exemplar-based face recognition from video. In: Proc. ECCV (2002) 732–746
13. Aggarwal G., Roy-Chowdhury A., and Chellappa R.: A system identification approach for video-based face recognition. In: Proc. ICPR (2004) 23 – 26
14. Gong S., Psarrou A., Katsouli I., and Palavouzis P.: Tracking and Recognition of Face Sequences. In: European Workshop on Combined Real and Synthetic Image Processing for Broadcast and Video Production (1994)
15. Hafed Z. and Levine M.: Face recognition using the discrete cosine transform. Int. J. Comput. Vision Vol. 43(3) (2001) 167 – 188
16. Howell A. and Buxton H.: Towards unconstrained face recognition from image sequences. In: Proc. Intl. Conf. on Automatic Face Recognition (1996) 224 – 229