Experimental Analysis of Face Recognition on Still and CCTV images

Shaokang Chen[†‡], Erik Berglund[†‡], Abbas Bigdeli[†‡], Conrad Sanderson[†‡], Brian C. Lovell[†‡]
[†]NICTA, PO Box 10161, Brisbane, QLD 4000, Australia
[‡]ITEE, University of Queensland, Brisbane, QLD 4072, Australia

Abstract

Although automatic identity inference based on faces has shown success when using high quality images, for CCTV based images it is hard to attain similar levels of performance. Furthermore, compared to recognition based on static images, relatively few studies have been done for video based face recognition. In this paper, we present an empirical analysis and comparison of face recognition using high quality and CCTV images in several important aspects: image quality (including resolution, noise, blurring and interlacing) as well as geometric transformations (such as translations, rotations and scale changes). The results show that holistic face recognition can be tolerant to image quality degradation but can also be highly influenced by geometric transformations. In addition, we show that camera intrinsics have much influence – when using different cameras for collecting gallery and probe images the recognition rate is considerably reduced. We also show that the classification performance can be considerably improved by straightforward averaging of consecutive face images from a CCTV video sequence.

1 Introduction

For surveillance applications, a potentially useful ability is the automatic inference on identity from face images. Although face recognition on well controlled still images (such as passport-quality photos) has achieved great success, automatic face recognition on uncontrolled images does not approach satisfying results. Furthermore, compared to recognition using still images, far fewer studies have been done for face recognition on video sequences [21]. Face recognition for CCTV is currently far from practical and many problems need to be solved before it can approach the capability of the human perception system.

The task is challenging as the overall similarity of all human faces accompanied by large differences between face images of the same person due to image capture variations such as changes in lighting, view point, head pose, and facial expression. Moreover, the differences between images of the same face due to these nuisance variations can be much greater than those between images of different faces [1], which makes it extremely difficult to compensate for common image capture variations. Most current face recognition systems only work well on controlled still images with images taken under very constrained conditions. Compared to still images, recognition using CCTV images is much more difficult due to the considerably worse quality of images. In addition, real-time recognition imposes time limits on the processing that can be done on each image [17].

In this paper, we empirically analyse several effects of image acquisition differences between still images and CCTV video sequences. In Section 2 we briefly describe Adaptive Principal Component Analysis for recognition with illumination, expression and pose variations. In Section 3, we explore the image quality differences between still images and CCTV images; we simulate these differences with corresponding image processing techniques. Section 4 is devoted to geometric transformations and simulation of alignment errors. Empirical results and analysis are given in Section 5, where we also introduce an averaging method which can improve the performance for CCTV based face recognition. Finally, we present conclusions and future work in Section 6.

2 Adaptive Principal Component Analysis

Recent research on face recognition has been focused on diminishing the impact of nuisance factors such as pose, illumination, and expression (PIE) variations on face recognition. Many approaches have been proposed for face recognition with PIE variations [2, 6, 13, 16]. The above approaches may be able to deal with certain kinds of face variation well but there are constraints restricting their application when multiple variations are involved and only one gallery image available per subject.

A holistic technique known as Adaptive Principal Component Analysis (APCA) was proposed in [3] for illumina-

978-0-7695-3341-4/08 $25.00 © 2008 IEEE
DOI 10.1109/AVSS.2008.15

tion and expression invariant face recognition with only one gallery image. The method was extended to deal with pose variations in [18]. We describe the method as follows.

APCA is a linear pattern classification algorithm that inherits merits from both Principal Component Analysis (PCA) and FLD (Fisher Linear Discriminant) by warping the face subspace according to the within-class and between-class covariance of samples. We first apply PCA on face images to extract eigenfaces. Consequently, every face image is projected into a face subspace with reduced dimensionality to form an m-dimensional feature vector $s_{j,k}$ with $k = 1, 2, ..., K_j$ denoting the kth sample of the class S_j . Then the face subspace is warped by the following three steps:

- Space Rotation: the feature space is rotated according to the overall within-class covariance. The rotation matrix R is a set of eigen vectors obtained by applying singular value decomposition to the overall within-class covariance matrix.

- Whitening Transformation: the subspace is whitened according to the eigen values $\lambda_i (i = 1, 2, ..., m)$ of the features in rotated face subspace with a whitening power p. Consequently, the whitening matrix is:

$$Z = diag\{\lambda_1^p, \lambda_2^p, ..., \lambda_m^p\} \tag{1}$$

- Eigenface Filtering: eigen-features are weighted according to the identification-to-variation value $ITV_i (i = 1, 2, ..., m)$ with a filtering power q. The ITV is a ratio measuring the correlation with a change in person versus a change in variation for each of the eigenfaces. It is defined as follows:

$$\begin{aligned} ITV_i &= \frac{\frac{1}{M}\sum_{j=1}^{M}\frac{1}{K}\sum_{k=1}^{K}|s_{i,j,k} - \varpi_{i,k}|}{\frac{1}{M}\sum_{j=1}^{M}\frac{1}{K}\sum_{k=1}^{K}|s_{i,j,k} - \mu_{i,j}|}, \\ \varpi_{i,k} &= \frac{1}{M}\sum_{j=1}^{M} s_{i,j,k}, \\ \mu_{i,j} &= \frac{1}{K}\sum_{k=1}^{K} s_{i,j,k}, i = [1, ..., m], \end{aligned} \tag{2}$$

where $s_{i,j,k}$ denotes the i_{th} element of the face vector of the k_{th} sample for class (person) S_j. Then the filtering matrix Γ is defined by:

$$\Gamma = diag\{ITV_1^q, ITV_2^q, ...ITV_m^q\}. \tag{3}$$

The whitening power p and filtering power q are determined empirically by searching the two dimensional domain of the following cost function. We define the distance between two face vectors $s_{j,k}$ and $s_{j',k'}$ as the Euclidean distance of their transformed vectors:

$$d_{jj',kk'} = \|Z\Gamma(s_{j,k} - s_{j',k'})\|_2. \tag{4}$$

The cost function OPT is a combination of error rate and the ratio of between-class distance to within-class distance follows:

$$\begin{aligned} OPT &= \textstyle\sum_{j=1}^{M}\sum_{k=1}^{K}\sum_{n}(\frac{d_{jj,k0}}{d_{jn,k0}}), \\ &\forall n \in d_{jn,k0} < d_{jj,k0}, n \in [1...M]. \end{aligned} \tag{5}$$

Evaluations on various face datasets show that APCA performs considerably better than PCA, PRM [11] and FLD. More detailed descriptions of the algorithm are given in [3, 18].

3 Photo vs Video Sequence Image Quality

For face recognition in surveillance scenarios, identifying a person captured on image or video is one of the key tasks. This implies matching face images with normal passport quality photos. Thus, in this paper we mainly focus on this specific image based recognition using one gallery image. The gallery image is generally a neutral frontal view face image taken in normal lighting conditions. Furthermore, gallery images should have good quality with relatively high resolution, high signal to noise ratio (SNR), appropriate contrast and clear details. Figure 1 shows some passport-quality sample gallery images. As images can be taken in various conditions with various scales and backgrounds, we need to apply normalisation. We manually mark the left and right eye centers of each image and scale the image to align the left and right eyes. The face region is then segmented from the image to remove the background. Finally, images are transformed from colour to grayscale. Examples of normalised gallery images are shown in Figure 2.

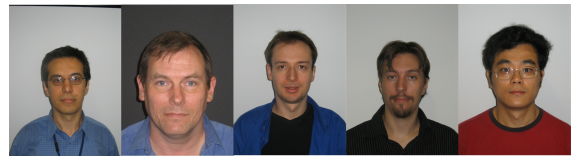

Figure 1. Sample gallery images.

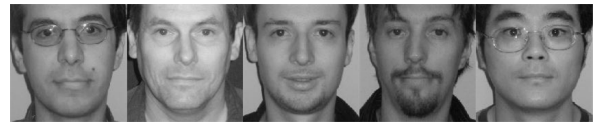

Figure 2. Normalised sample gallery images.

Figure 3. Normalised still face images, captured using a different camera than used in Figures 1 and 2.

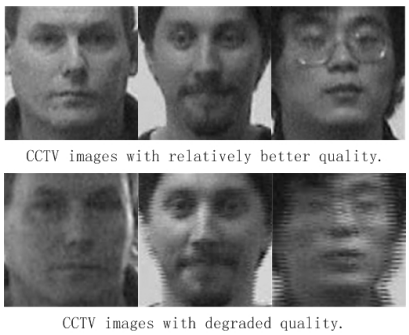

Figure 4. Normalised CCTV face images.

There are no readily apparent differences between the quality of still faces images (captured using a digital video camera) and reference passport photos. Figure 3 displays some normalised still face images captured by a digital video camera.

Images captured by CCTV cameras, on the other hand, are of poor quality. The image resolution is generally lower for CCTV cameras, the noise level is higher and images may be blurred due to movement or the subject being out of focus. These effects are illustrated in Figure 4. Images in the first row are CCTV images with relatively good quality. The second row shows degraded images, where the left hand side picture shows the effect of out of focus, the middle picture displays the effect of interlacing due to object movement and the right hand side one illustrates the combination of out of focus and interlacing. By comparing Figures 3 and 4, it can be seen that the image quality of CCTV cameras (even high-end ones) is much worse than still images. However, one advantage of video sequences over still images is that they contain motion and temporal information, which can be helpful for recognition in degraded environments [14].

3.1 Simulation of Image Degradation

In order to consistently compare still images and CCTV images, we attempt to simulate the above image degradations to find out how quality variations affect recognition. Below is a list of the degradations with our corresponding simulation:

- Resolution: images are down-sampled to various smaller sizes and up-sampled to the standard size by bi-linear interpolation between pixels.
- Noise: various levels of zero mean Gaussian noise are added.
- Blurring: circular averaging filters with various sizes are applied.
- Interlacing: images are interlaced with various movements.
- Combination: all of the above.

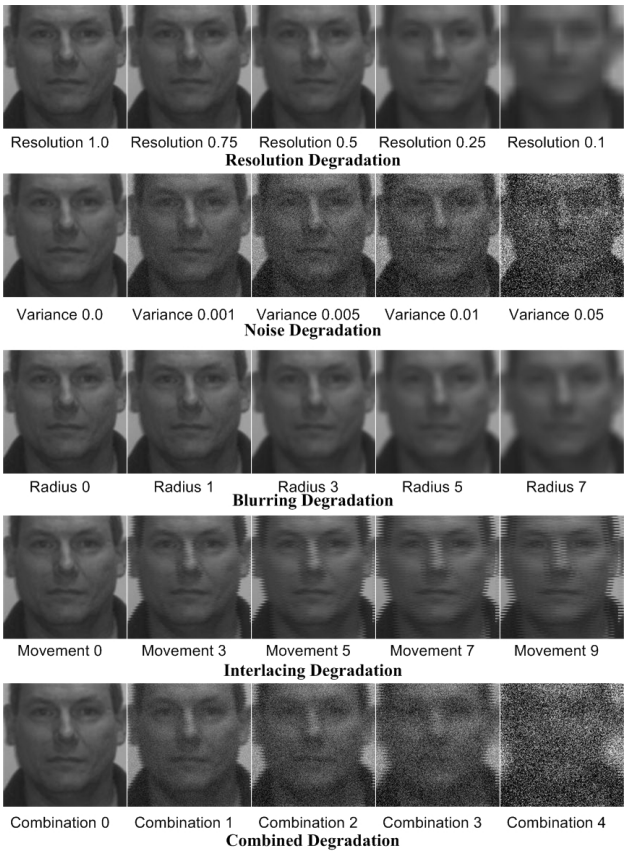

Figure 5. Simulated degraded sample images.

Figure 5 shows the corresponding simulated degraded images at five levels. The first row displays the resolution change effect. The second row is the effect of adding noise. The influence of blurring is illustrated in the third row. The fourth row shows the effect of interlacing due to movement.

The last row demonstrates the result of combining all of the four degradation operations, in which each bottom image is the effect of the above four operations in the same column. The first image of each row is the original image without degradation.

4 Image Transformation

Apart from image quality, image transformation due to face alignment errors also affects recognition. Holistic feature based face recognition methods, such as PCA and FLD, generally require accurate facial feature alignment. Thus these methods are very sensitive to feature location errors, or in other words, geometric transformations. We simulate three basic geometric transformations: translation, rotation and scale. The following is a list of the simulated transformations:

- X and Y axis translation: images are shifted either along X or Y axis, with various offsets.
- In-plane rotation: images are rotated with various angles.
- Scale difference: various scale scale changes are applied.
- Combination: all of the above geometric transformations are applied.

Figure 6 illustrates some samples of the corresponding transformations at five different levels. The last row shows the effect of combining all of the above four geometric operations within the same column (level). The total number of combinations for each level is 16. Here we only demonstrate one sample for each combination due to space limitations.

5 Empirical Analysis

The APCA algorithm has shown to achieve high accuracy on several public face datasets such as FERET [15] and Asian Face [19] datasets [3, 18]. Even though the above public databases cover all three kinds of the PIE variations, the range of variations they cover is limited due to the fact that each subject is represented by a limited number of images taken in a controlled environment. In this paper, we attempt to do the test based on real-life images. We therefore collected static face images from an Axis-211 network digital camera and video sequences from a Sony IPELA SNC RZ30 CCTV camera. We performed the test on the two data-sets separately. The APCA algorithm is is trained on images from the Asian Face dataset [19].

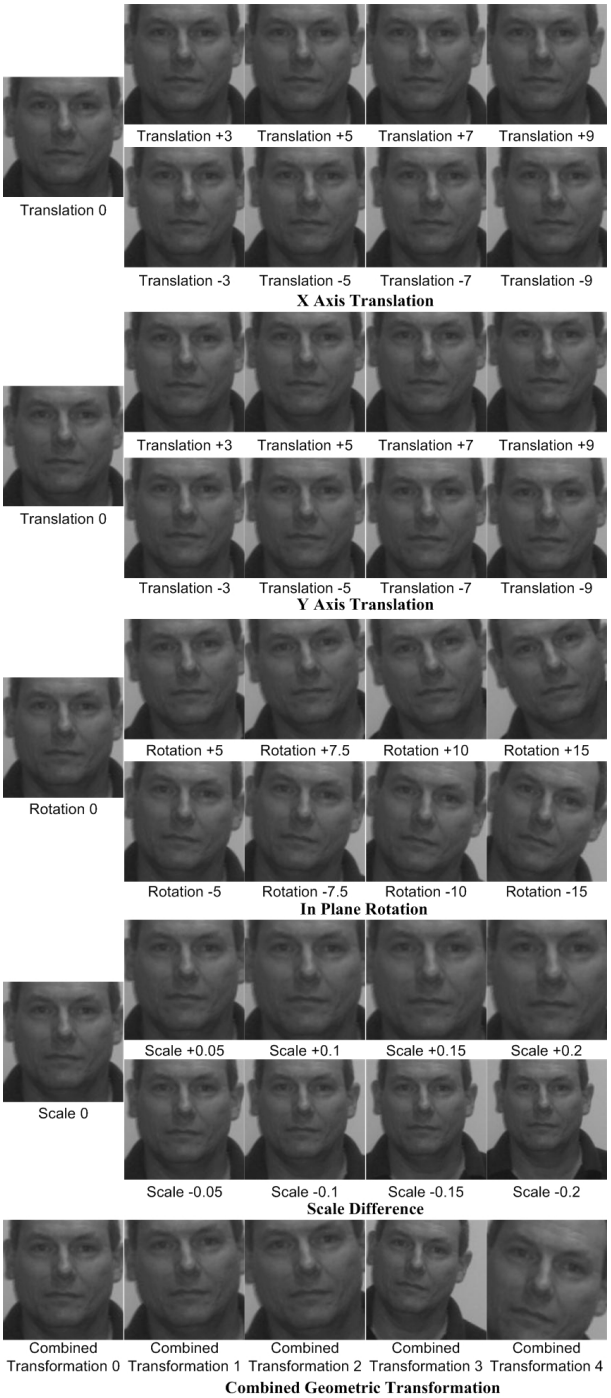

Figure 6. Simulated transformed sample images.

5.1 Still Image Test

In total there are 178 images taken from five subjects. Each subject attempts to stand still at three different posi-

tions (at specific distances to the camera) and tries to face the camera to take photos. However, there are still PIE variations as shown in Figure 7.

5.1.1 Image Degradation

For the image degradation test, we attempt to evaluate how degradation affects the final recognition performance. In order to avoid the effect of pre-processing, all the face images are manually aligned and normalised to the same size according to eye positions. We then divide this test into two sub-tests using different gallery images. In the first sub-test, we only register one passport quality identity photo per subject as gallery and use the still face images acquired from the networked digital camera as probe. In the second sub-test, we manually choose one image per subject from the networked digital still images as gallery and the remainder as probe. Tables 1 and 2 show the results of both sub-tests, respectively. As can be seen, the image degradation operations do not affect the recognition performance significantly. With the increase of the strength of degradation, the recognition accuracy only decreases slightly. Moreover, the combined degradation has more influence on the accuracy than the corresponding single degradation operation at the same level. An important finding is that recognition using chosen still images as gallery performs much better than using passport quality photos. We attribute this phenomenon to the differences between camera intrinsics. Passport quality photos are acquired from a different digital camera, which may induce appearance variation of face images due to the differences in capture conditions and the camera optical system. Thus, we observe that image degradations have less influences on sub-test 2 than sub-test 1, with no accuracy reduction under the first three levels of degradation in sub-test 2.

5.1.2 Image Transformation

We now estimate how image transformation affects recognition performance. Similar to the image degradation test, we divide this test into two sub-tests by using two sets of gallery images: passport quality photos and selected still images. Tables 3 and 4 show the results of both image transformation sub-tests, respectively. It is apparent that recognition performance is affected significantly by geometric transformations. With higher levels of each transformation, the recognition accuracy decreases. In addition, the combined geometric transformations have more influence on recognition performance than the corresponding single transformation at the same level. Similar to image degradation test, face recognition using chosen still images as gallery performs much better than using passport quality photos. The former is also less affected by image pre-processing operations than the latter.

Figure 7. Sample still face images with PIE variations

Table 1. Image degradation on still face image sub-test 1 result using passport quality images as gallery.

Resolution	1	0.75	0.5	0.25	0.1
Accuracy	62.3%	61.2%	60.7%	58.5%	55.7%
Noise	0	0.001	0.005	0.01	0.05
Accuracy	62.3%	62.3%	62.3%	61.7%	61.2%
Blur	0	1	3	5	7
Accuracy	62.3%	62.3%	61.2%	59.6%	56.8%
Interlacing	0	3	5	7	9
Accuracy	62.3%	62.3%	61.7%	61.2%	56.3%
Combination	0	1	2	3	4
Accuracy	62.3%	61.9%	61.6%	59.3%	51.0%

5.2 CCTV Image Test

300 images were captured from Sony IPELA SNC RZ30 CCTV camera for 14 subjects. The camera was located at a fixed position at the end of a corridor and images were taken while subjects were walking towards the camera along the

Table 2. Image degradation on still face image sub-test 2 result using chosen still images as gallery.

Resolution	1	0.75	0.5	0.25	0.1
Accuracy	89.3%	89.3%	89.3%	89.3%	85.4%
Noise	0	0.001	0.005	0.01	0.05
Accuracy	89.3%	89.3%	89.3%	89.3%	89.1%
Blur	0	1	3	5	7
Accuracy	89.3%	89.3%	89.3%	89.3%	88.8%
Interlacing	0	3	5	7	9
Accuracy	89.3%	89.3%	89.3%	89.3%	88.2%
Combination	0	1	2	3	4
Accuracy	89.3%	89.3%	89.3%	87.7%	81.6%

corridor. We manually aligned face images and normalised them to the standard size. Figure 4 shows some samples of the normalised images captured from the CCTV camera. Again, we perform the video image test by using two different sets of gallery images as before.

As can be seen in Table 5, using CCTV images as gallery performs slightly better than using passport quality photos, which is consistent with the static image tests. However, comparing with the results of static image tests, we can see that the recognition accuracy decreases considerably, especially when using chosen CCTV images. The following three reasons may lead to the degradation of the performance. First, the image quality degradation of CCTV images compared to still images. Second, the change of the number of subjects. Third, the alignment errors due to the image quality degradation.

We now analyse these three main reasons individually. From the still image test, we can see that image degradation has only minor effects on performance. This result is also consistent with the results from [8, 12, 20], where the experiments also show that resolution changes and blurring have limited influence on recognition.

It may be argued that the number of subjects for the CCTV test is larger than for the still image test. But randomly choosing face images from the five subjects for testing gives the same average accuracy as the results in Table 5. Other tests on face recognition show that generalisation ability of appearance based subspace algorithms is only slightly affected by the numbers of subjects [5]. Our test of APCA on the FERET and Asian face datasets also indicate that this method can handle face recognition for more than 100 subjects.

The last but not the least reason leading to the reduction of accuracy is the alignment issue. Even though image quality degradation has little effect on recognition accuracy directly, it may affect alignment accuracy, which could lead to a significant reduction in performance. We evaluate the effect of quality degradation on image alignment in the following way. We manually align eye positions of still images and consider it as the ground truth. Then we apply combined degradation operations as shown in Figure 5 on still images. We do the manual alignment on the degraded images and compare the eye positions of the degraded images and original still images. Table 6 is the average alignment error due to image degradation. From Tables 1 and 2, we can see that the highest level of combined degradation reduces the accuracy by about 10 percentage points. The alignment error caused by the level 4 image degradation is similar to level 2 or 3 combined geometric transformation, which could reduce the accuracy by up to 30 percentage points. Therefore, for CCTV image based face recognition, alignment errors due to image degradations are more serious for recognition than the quality degradation itself.

5.3 Averaging CCTV Images

Based on the above observations, we propose an efficient straightforward method to improve the performance for video based face recognition. We need to reduce alignment error since it appears to be the main factor that influences performance. One way is to apply advanced image processing techniques such as Active Appearance Models [4] to align facial features. Unfortunately, these methods are time consuming and sensitive to image degradation. Currently, most feature alignment methods cannot work on heavily degraded images. Another way is to morph all images to the common shape and average them [9, 10]. Again, morphing is also time consuming and is highly affected by image degradation. We thus propose a direct averaging method that is suitable for CCTV based face recognition. The alignment errors generated due to image degradation are more apparent on x- and y-axis translations than on scale and rotation differences. Fortunately, translation errors are very likely to have a Gaussian distribution and can

Table 3. Image pre-processing on still face image sub-test 1 result using passport quality images as gallery.

X Translation	0	3	5	7	9
Accuracy	62.3%	53.3%	49.5%	47.3%	46.5%
Y Translation	0	3	5	7	9
Accuracy	62.3%	57.9%	51.9%	50.5%	47.0%
Rotation	0	5	7.5	10	15
Accuracy	62.3%	53.3%	47.0%	40.7%	34.4%
Scale	0	0.05	0.1	0.15	0.2
Accuracy	62.3%	48.9%	51.3%	43.5%	36.0%
Combination	0	1	2	3	4
Accuracy	62.3%	42.6%	37.2%	35.5%	31.7%

Table 4. Image pre-processing on still face image sub-test 2 result using chosen still images as gallery.

X Translation	0	3	5	7	9
Accuracy	89.3%	89.3%	84.7%	79.5%	73.6%
Y Translation	0	3	5	7	9
Accuracy	89.3%	89.1%	87.2%	83.3%	81.1%
Rotation	0	5	7.5	10	15
Accuracy	89.3%	81.1%	74.6%	64.2%	48.6%
Scale	0	0.05	0.1	0.15	0.2
Accuracy	89.3%	86.6%	80.3%	76.4%	68.0%
Combination	0	1	2	3	4
Accuracy	89.3%	78.7%	66.1%	56.8%	41.2%

Table 5. CCTV Image Test Results.

Gallery image dataset	Accuracy
Passport quality photos	55.0%
Chosen CCTV images	58.7%

Table 6. Alignment error due to image degradation.

Alignment Error	Combined Degradation			
	1	2	3	4
X Translation	0.89	2.75	4.43	6.94
Y Translation	0.86	1.68	2.37	5.91
Scale Difference	0.01	0.03	0.05	0.12
Rotation Difference	0.65	1.25	2.33	3.15

be efficiently reduced by averaging. Therefore, we simply average normalised face images, without the use of degradation sensitive operations such as morphing or deformation. This average operation not only reduces the alignment errors but also decreases noise and compensates for interlacing. One disadvantage of averaging is that it may induce blurring, however the still image tests done earlier suggest that blurring has only little influence on performance. Figure 8 illustrates the effect of averaging different number of consecutive CCTV face images.

It can be seen that the first image (without averaging) contains significant noise. After averaging, the noise is reduced and images are slightly blurred. We test the performance of averaging on CCTV image based recognition and present the results in Figure 9. As can be seen as the number of images is increased, recognition accuracy raises significantly for both gallery image datasets. This result is consistent with the test in [7], where temporal representation of face images are used and the video sequence lengths affect the performance.

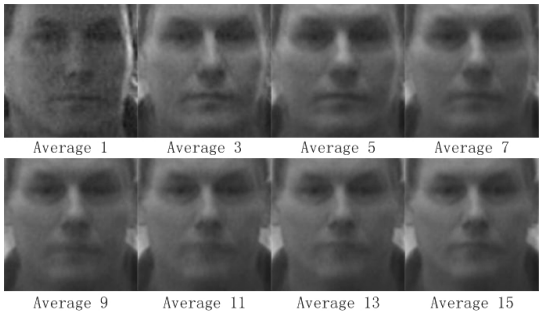

Figure 8. Averaged sample images.

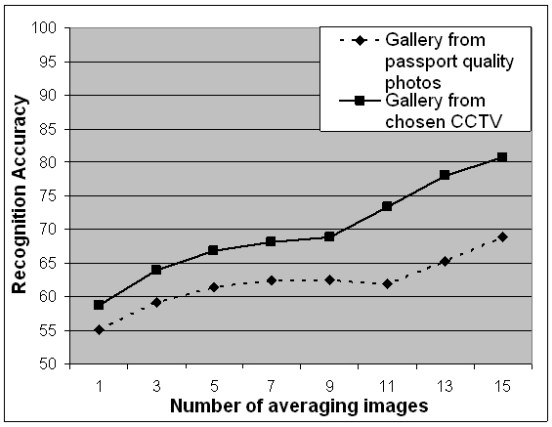

Figure 9. Test results on averaging CCTV images.

6 Conclusions

In this paper we evaluated the effect of image quality degradation and image transformation on image based face recognition. We simulated both degradation and transformation operations at various levels and tested on both static and CCTV face images. Empirical results show that holistic feature based face recognition is tolerant to image degradations while being sensitive to geometric transformations. Moreover, image degradations bring about alignment errors which lead to a more serious decrease in recognition accuracy than caused by the degradation itself. The results also show that recognition performance depends on which camera is being used: the performance of gallery and probe images collected from the same camera is much higher than the performance of gallery and probe images captured from different cameras. Finally, we proposed a straightforward averaging method, which can be applied easily on any image based face recognition system without adversely reducing processing speed. Empirical evaluation show that the averaging method can compensate for alignment errors and reduce noise, which considerably improves recognition performance. Our future work will focus on how to normalise face images to compensate for differences between cameras.

Acknowledgements

This project is supported by a grant from the Australian Government Department of the Prime Minister and Cabinet. NICTA is funded by the Australian Government's *Backing Australia's Ability* initiative, in part through the Australian Research Council.

References

[1] Y. Adinj, Y. Moses, and S. Ullman. Face recognition: The problem of compensation for changes in illumination direction. *IEEE Transactions on PAMI*, 19(7):721–732, 1997.

[2] R. Basri and D. W. Jacobs. Lambertian reflectance and linear subspaces. *IEEE Transactions on PAMI*, 25(2):218–233, 2003.

[3] S. Chen and B. C. Lovell. Illumination and expression invariant face recognition with one sample image. In *Proc. of Intl. Conf. on Pattern Recognition*, 2004.

[4] T. Cootes, K. Walker, and C. Taylor. View-based active appearance models. In *IEEE Intl. Conf. on Automatic Face and Gesture Recognition*, 2000.

[5] K. Delac, M. Grgic, and S. Grgic. Generalization abilities of appearance-based subspace face recognition algorithms. In *Intl. Workshop on Systems, Signals and Image Processing*, Chalkida, Greece, 2005.

[6] W. Gao, S. Shan, X. Chai, and X. Fu. Virtual face image generation for illumination and pose insensitive face recognition. In *Intl. Conf. on Multimedia and Expo*, 2003.

[7] A. Hadid and M. Pietikainen. From still image to video-based face recognition: An experimental analysis. In *Proc. of the 6th Intl. Conf. on Automatic Face and Gesture Recognition*, Seoul, Korea, 2004.

[8] T. Heseltine, N. Pears, J. Austin, and Z. Chen. Face recognition: A comparison of appearance-based approaches. In *Intl. Proc. of Digital Image Computing: Techniques and Applications*, Sydney, Australia, 2003.

[9] R. Jenkins and A. M. Burton. 100% accuracy in automatic face recognition. *Science*, 319:435, 2008.

[10] R. Jenkins, A. M. Burton, and D. White. Face recognition from unconstrained images: Progress with prototypes. In *Proc. of the Intl. Conf. on Automatic Face and Gesture Recognition*, Southampton, UK, 2006.

[11] C. Liu and H. Wechsler. Probabilistic reasoning models for face recognition. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 1998.

[12] C. H. Liu, H. Seetzen, A. M. Burton, and A. Chaudhuri. Face recognition is robust with incongruent image resolution: Relationship to security video images. *Journal of Experimental Psychology: Applied*, 9(1):33–41, 2003.

[13] X. Liu, T. Chen, and B. V. Kumar. Face authentication for multiple subjects using eigenflow. *Pattern Recognition*, 36:313–328, 2003.

[14] A. J. O'Toole, D. A. Roark, and H. Abdi. Recognizing moving faces: A psychological and neural synthesis. *Trends in Cognitive Sciences*, 6(6):261–266, 2002.

[15] P. J. Phillips, H. Moon, P. J. Rauss, and S. Rizvi. The FERET evaluation methodology for face recognition algorithms. *IEEE Transactions on PAMI*, 20(10):1090–1104, 2000.

[16] C. Sanderson, S. Bengio, and Y. Gao. On transforming statistical models for non-frontal face verification. *Pattern Recognition*, 39(2):288–302, 2006.

[17] C. Sanderson, T. Shang, and B. C. Lovell. Towards pose-invariant 2D face classification for surveillance. In *Analysis and Modeling of Faces and Gestures, Lecture Notes in Computer Science (LNCS)*, volume 4778, pages 276–289, 2007.

[18] T. Shan, B. C. Lovell, and S. Chen. Face recognition robust to head pose from one sample image. In *Proc. of Intl. Conf. on Pattern Recognition*, 2006.

[19] Intelligent Multimedia Lab, Department of Computer Science and Engineering, Pohang University of Science and Technology. Asian face image database PF01.

[20] J. Wang, C. Zhang, and H.-Y. Shum. Face image resolution versus face recognition performance based on two global methods. In *Asian Conf. on Computer Vision*, Jeju Island, Korea, 2004.

[21] S. Zhou, V. Krueger, and R. Chellappa. Probabilistic recognition of human faces from video. *Computer Vision and Image Understanding*, 91:214–245, 2003.