

# Patch-based Probabilistic Image Quality Assessment for Face Selection and Improved Video-based Face Recognition

Yongkang Wong, Shaokang Chen, Sandra Mau, Conrad Sanderson, Brian C. Lovell

NICTA, PO Box 6020, St Lucia, QLD 4067, Australia

The University of Queensland, School of ITEE, QLD 4072, Australia

## Abstract

*In video based face recognition, face images are typically captured over multiple frames in uncontrolled conditions, where head pose, illumination, shadowing, motion blur and focus change over the sequence. Additionally, inaccuracies in face localisation can also introduce scale and alignment variations. Using all face images, including images of poor quality, can actually degrade face recognition performance. While one solution is to use only the ‘best’ subset of images, current face selection techniques are incapable of simultaneously handling all of the abovementioned issues. We propose an efficient patch-based face image quality assessment algorithm which quantifies the similarity of a face image to a probabilistic face model, representing an ‘ideal’ face. Image characteristics that affect recognition are taken into account, including variations in geometric alignment (shift, rotation and scale), sharpness, head pose and cast shadows. Experiments on FERET and PIE datasets show that the proposed algorithm is able to identify images which are simultaneously the most frontal, aligned, sharp and well illuminated. Further experiments on a new video surveillance dataset (termed ChokePoint) show that the proposed method provides better face subsets than existing face selection techniques, leading to significant improvements in recognition accuracy.*

## 1. Introduction

Video-based identity inference in surveillance conditions is challenging due to a variety of factors, including the subjects’ motion, the uncontrolled nature of the subjects, variable lighting, and poor quality CCTV video recordings. This results in issues for face recognition such as low resolution, blurry images (due to motion or loss of focus), large pose variations, and low contrast [14, 35, 40]. While recent face recognition algorithms can handle faces with moderately challenging illumination conditions [15, 17, 24, 28], strong illumination variations (causing cast shadows and self-shadowing) remain problematic [30].

One approach to overcome the impact of poor quality images is to assume that such images are outliers in a se-

quence. This includes approaches like exemplar extraction using clustering techniques (eg. k-means clustering [13]) and statistical model approaches for outlier removal [6]. However, these approaches are not likely to work when most of the images in the sequence have poor quality — the good quality images would actually be classified as outliers.

Another approach is explicit subset selection, where a face quality assessment is automatically made on each image, either to remove poor quality face images, or to select a subset comprised of high quality images [10, 21, 32]. This improves recognition performance, with the additional benefit of reducing the overall computation load during feature extraction and matching [19]. The challenge in this approach is finding a good definition for “face quality”.

Several face image standards have been proposed for face quality assessment (eg. ISO/IEC 19794-5 [1] and ICAO 9303 [2]). In these standards, quality can be divided into: (i) *image* specific qualities such as sharpness, contrast, compression artifacts, and (ii) *face* specific qualities such as face geometry, pose, eye detectability, illumination angles.

Based in part on the above standards, many approaches have been proposed to analyse various face and image properties. For example, face pose estimation using tree structured multiple pose estimators [38], and face alignment estimation using template matching [7]. Asymmetry analysis has been proposed to simultaneously estimate two qualities: out-of-plane rotation and non-frontal illumination [10, 29, 39].

Since face recognition performance is simultaneously impacted by multiple factors, being able to detect one or two qualities is insufficient for robust subset selection. One approach to simultaneously detect multiple quality characteristics is through a fusion of individual face and image quality measurements. Nasrollahi and Moeslund [21] proposed a weighted quality fusion approach to combine out-of-plane rotation, sharpness, brightness, and image resolution qualities. Rua et al. [26] proposed a similar quality assessment approach, by using asymmetry analysis and two sharpness measurements. Hsu et al. [16] proposed to learn fusion parameters on multiple quality scores to achieve maximum correlation with matching scores between face pairs. Another proposed fusion approach uses a Bayesian network to

model the relationships among qualities, image features and matching scores [22]. The main drawbacks of the above fusion approaches are:

- Fusion-based approaches only perform as well as their individual classifiers. For example, if a pose estimation algorithm requires accurate facial feature localisation, the whole fusion framework will fail in the cases where that pose algorithm fails (such as in low resolution CCTV footage) [34].
- As various properties are measured individually and have different influence on face quality, it may be difficult to combine them to output a single quality score for the purposes of image selection.
- As multiple classifiers as involved, they are typically more time consuming and hence may not be suitable for real-time surveillance applications.
- Since face matching scores are heavily dependant on system-specific details (including the input features, matching algorithms and training images), quality assessment approaches that learn a fusion model based on match scores end up being closely tied to the particular system configuration and hence need to be re-trained for each system.

Simultaneously detecting multiple quality characteristics can also be accomplished by learning a generic model to define the ‘ideal’ quality. Luo [18] proposed a learning based approach where the quality model is trained to correlate with manually labelled quality scores. However, given the subjective nature of human labelling, and the fact that humans may not know what characteristics work best for automatic face recognition algorithms, this approach may not generate the best quality model for face recognition.

In this paper we propose a straightforward and effective patch-based face quality assessment algorithm, targeted towards handling images obtained in surveillance conditions. It quantifies the similarity of a given face to a probabilistic face model, representing an ‘ideal’ face, via patch-based local analysis. Without resorting to fusion, the proposed algorithm outputs a single score for each image, with the score simultaneously reflecting the degree of alignment errors, pose variations, shadowing, and image sharpness (underlying resolution). Localisation of facial features (ie. eyes, nose, mouth) is not required.

We continue the paper as follows. In Section 2 we describe the proposed quality assessment algorithm. Still image and video datasets used in the experiments are briefly described in in Section 3. Extensive performance comparisons against existing techniques are given in Section 4 (on still images) and Section 5 (on surveillance videos). The main findings are discussed in Section 6.

## 2. Probabilistic Face Quality Assessment

The proposed algorithm is comprised of five steps: (1) pixel-based image normalisation, (2) patch extraction and normalisation, (3) feature extraction from each patch, (4) local probability calculation, (5) overall quality score generation via integration of local probabilities. These steps are elaborated below:

1. For a given image  $I$ , we perform non-linear pre-processing (log transform) to reduce the dynamic range of data. Following [9], the normalised image  $I_{\log}$  is calculated using:

$$I_{\log}(r, c) = \ln[I(r, c) + 1] \quad (1)$$

where  $I(r, c)$  is the pixel intensity located at  $(r, c)$ . Logarithm normalisation amplifies low intensity pixels and compresses high intensity pixels. This property is helpful in reducing the intensity differences between skin tones.

2. The transformed image  $I_{\log}$  is divided into  $N$  overlapping blocks (patches). Each block  $b_i$  has a size of  $n \times n$  pixels and overlap neighbouring blocks by  $t$  pixels. To accommodate for contrast variations between face images, each patch is normalised to have zero mean and unit variance [36].
3. From each block, a 2D Discrete Cosine Transform (DCT) feature vector is extracted [11]. Excluding the 0-th DCT component (as it has no information due to the previous normalisation), the top  $d$  low frequency components are retained. The low frequency components retain generic facial textures [12], while largely omitting person-specific information. At the same time, cast shadows [36] as well as variations in pose and alignment can alter the local textures.
4. For each block location  $i$ , the probability of the corresponding feature vector  $x_i$  is calculated using a *location specific* probabilistic model:

$$p(x_i | \mu_i, \Sigma_i) = \frac{\exp \left[ -\frac{1}{2} (x_i - \mu_i)^T \Sigma_i^{-1} (x_i - \mu_i) \right]}{(2\pi)^{\frac{d}{2}} |\Sigma_i|^{\frac{1}{2}}} \quad (2)$$

where  $\mu_i$  and  $\Sigma_i$  are the mean and covariance matrix of a normal distribution. The model for each location is trained using a pool of frontal faces with frontal illumination and neutral expression. All of the training face images are first scaled and aligned to a fixed size, with each eye located at a fixed location. We emphasise that during testing, the faces do not need to be aligned.

5. By assuming that the model for each location is independent, an overall probabilistic quality score  $Q$  for image  $I$ , comprised of  $N$  blocks, is calculated using:

$$Q(I) = \sum_{i=1}^N \log p(x_i | \mu_i, \Sigma_i) \quad (3)$$

The resulting quality score represents the probabilistic similarity of a given face to an “ideal” face (as represented by a set of training images). A higher quality score reflects better image quality.

### 3. Face Datasets

In this section, we briefly describe the FERET, PIE and ChokePoint face datasets, as well as their setup for our experiments.

FERET [23] and PIE [31] are used to analyse how accurate the proposed quality assessment algorithm is for correctly selecting best quality images with several desired characteristics, compared to other existing methods. In total, there are 1124 unique subjects in the training phase and 1263 subjects in the test phase.

The ChokePoint dataset contains surveillance videos. It is used to study the improvement in verification performance gained from subset selection, using the proposed quality method as well as other approaches.

#### 3.1. Setup of Still Image Datasets: FERET and PIE

To study the performance of the proposed method in terms of correctly selecting images with desired characteristics, we simulated blurring as well as four alignment errors using images from the ‘fb’ subset of FERET. Experiments with pose variations (out-of-plane rotation) used dedicated subsets from FERET and PIE. Experiments with cast shadows used the illumination subset of PIE.

The generated alignment errors<sup>1</sup> are: horizontal shift and vertical shift (using displacements of 0,  $\pm 2$ ,  $\pm 4$ ,  $\pm 6$ ,  $\pm 8$  pixels), in-plane rotation (using rotations of  $0^\circ$ ,  $\pm 10^\circ$ ,  $\pm 20^\circ$ ,  $\pm 30^\circ$ ), and scale variations (using scaling factors of 0.7, 0.8, 0.9, 1.0, 1.1, 1.2, 1.3). For sharpness variations, each original image is first downsampled to three sizes ( $48 \times 48$ ,  $32 \times 32$  and  $16 \times 16$  pixels) then rescaled to the baseline size of  $64 \times 64$  pixels. See Fig. 1 for examples.

FERET provides the dedicated ‘b’ subset with pose variations, containing out-of plane rotations of  $0^\circ$ ,  $\pm 15^\circ$ ,  $\pm 25^\circ$ ,  $\pm 40^\circ$ ,  $\pm 60^\circ$ . PIE also provides a dedicated subset with pose variations, though with a smaller set of rotations ( $0^\circ$ ,  $\pm 22.5^\circ$ ,  $\pm 45^\circ$ ,  $\pm 67.5^\circ$ ).

The illumination subset of PIE was used to assess performance in various cast shadow conditions. In our experiments, we divided the frontal view images into six subsets<sup>2</sup> based on the angle of the corresponding light source. Subset 1 has the most frontal light sources, while subset 6 has the largest light sources angle ( $54^\circ - 67^\circ$ ). See Fig. 2 for examples.

<sup>1</sup>The generated alignment errors are representatives of real-life characteristics of automatic face localisation/detection algorithms [25].

<sup>2</sup>Subset 1: light source 8, 11, 20; Subset 2: light source 6, 7, 9, 12, 19, 21; Subset 3: light source 5, 10, 13, 14; Subset 4: light source 18, 22; Subset 5: light source 4, 15; Subset 6: light source 2, 3, 16, 17.

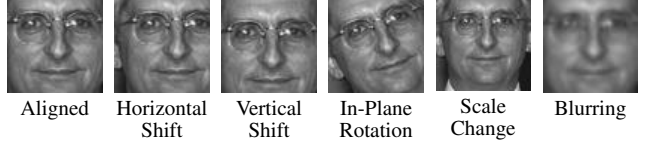


Figure 1. Examples of simulated image variations on FERET.

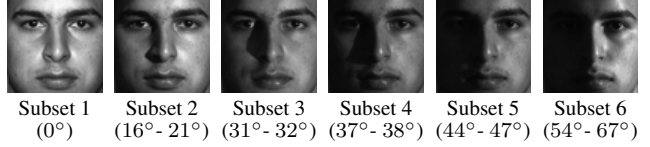


Figure 2. Examples from PIE with strong directed illumination, causing self-shadowing.

#### 3.2. Surveillance Videos: ChokePoint Dataset

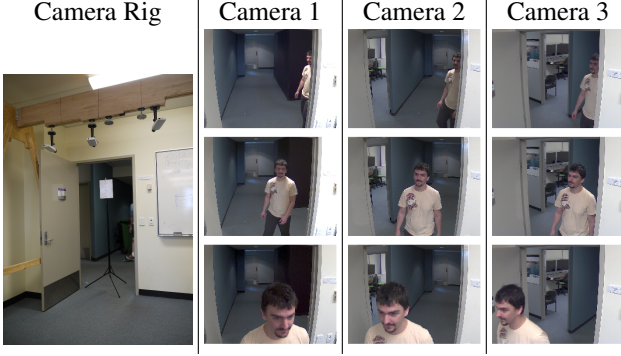
We collected a video dataset<sup>3</sup>, termed *ChokePoint*, designed for experiments in person identification/verification under real-world surveillance conditions using existing technologies. An array of three cameras was placed above several portals (natural choke points in terms of pedestrian traffic) to capture subjects walking through each portal in a natural way (see Figs. 3 and 4).

While a person is walking through a portal, a sequence of face images (ie. a face set) can be captured. Faces in such sets will have variations in terms of illumination conditions, pose, sharpness, as well as misalignment due to automatic face localisation/detection [25]. Due to the three camera configuration, one of the cameras is likely to capture a face set where a subset of the faces is near-frontal.

The dataset consists of 25 subjects (19 male and 6 female) in portal 1 and 29 subjects (23 male and 6 female) in portal 2. In total, it consists of 48 video sequences and 64,204 face images. Each sequence was named according to the recording conditions (eg. P2E.S1\_C3) where P, S, and C stand for *portal*, *sequence* and *camera*, respectively. E and L indicate subjects either *entering* or *leaving* the portal. The numbers indicate the respective portal, sequence and camera label. For example, P2L\_S1\_C3 indicates that the recording was done in Portal 2, with people leaving the portal, and captured by camera 3 in the first recorded sequence.

In this paper, all the experiments were performed with the video-to-video verification protocol. In this protocol, video sequences are divided into two groups ( $G1$  and  $G2$ ), where each group played the role of development set and evaluation set in turn. Parameters can be first learned on the development set and then applied on the evaluation set. The average verification rate is used for reporting results. In our experiments we selected the frontal view cameras (shown in Table 1). In each group, each sequence takes turn to be the gallery, with the the leftover sequences becoming the probe.

<sup>3</sup><http://itee.uq.edu.au/~conrad/chokepoint.html>



**Figure 3.** An example of the recording setup used for the *ChokePoint* dataset. A camera rig contains 3 cameras placed just above a door, used for simultaneously recording the entry of a person from 3 viewpoints. The variations between viewpoints allow for variations in walking directions, facilitating the capture of a near-frontal face by one of the cameras.



**Figure 4.** Example shots from the ChokePoint dataset, showing portals with various backgrounds.

**Table 1.** ChokePoint video-to-video verification protocol. Sequences are divided into two groups (G1 and G2). Listed sequences contain faces with the most frontal pose view. P, S, and C stand for *portal*, *sequence* and *camera*, respectively. E and L indicate subjects *entering* or *leaving* the portal. The numbers indicate the respective portal, sequence and camera label. For example, P2L.S1.C3 indicates that the recording was done in Portal 2, with people leaving the portal, and captured by camera 3 in the first recorded sequence.

G1	P1E.S1.C1	P1E.S2.C2	P2E.S2.C2	P2E.S1.C3
	P1L.S1.C1	P1L.S2.C2	P2L.S2.C2	P2L.S1.C1
G2	P1E.S3.C3	P1E.S4.C1	P2E.S4.C2	P2E.S3.C1
	P1L.S3.C3	P1L.S4.C1	P2L.S4.C2	P2L.S3.C3

## 4. Experiments on Still Images

In this section, we evaluate how well the proposed quality assessment method can identify the best quality faces when presented with both good and poor quality faces. The proposed method was compared with: (i) a score fusion method using pixel based asymmetry analysis and two sharpness analyses (denoted as *Asym\_shrp*) [26], (ii) asymmetry analysis with Gabor features (denoted as *Gabor\_asym*) [29], (iii) the classical Distance From Face Space (DFFS) method [5].

The ‘fa’ subset of FERET, containing frontal faces with frontal illumination and neutral expression, was used to train the location specific probabilistic models in the proposed method. The ‘fa’ subset was also used to select the decision threshold for rejecting “poor” quality images. The ‘fa’ subset was not used for any other purposes.

Based on preliminary experiments, closely cropped face images were scaled to  $64 \times 64$  pixels, the block size was set to  $8 \times 8$  pixels, with a 7 pixels overlap of neighbouring blocks. The preliminary experiments also suggested that using just 3 DCT coefficients was sufficient. This configuration was used in all experiments. The quality assessment methods were implemented with the aid of the Armadillo C++ library [27].

### 4.1. Quality Assessment of Faces with Variations in Alignment, Scale and Sharpness

In this experiment we evaluated the efficacy of each method to detect the best aligned images within a set of images that have a particular image variation. For example, out of the set of faces with rotations of  $0^\circ$ ,  $\pm 10^\circ$ ,  $\pm 20^\circ$ ,  $\pm 30^\circ$ , we measured the percentage of  $0^\circ$  faces that were labelled as “high” quality.

Results for variations in shift, rotation and scale, shown in Table 2, indicate that the proposed method consistently achieved the best or near-best performance across most of the variations. The results on the six PIE illumination subsets indicate that even in the presence of cast shadows, the proposed method can achieve good results, with the exception of images with scale changes. Averaging over all variations, the proposed method achieved the best results.

The asymmetry-based analysis methods (*Gabor\_asym* and *Asym\_shrp*) could not reliably detect vertical alignment errors and scale variations. *Gabor\_asym* also performed poorly for detecting images with various sharpness variations. *Asym\_shrp* addressed this by combining asymmetry analysis with two image sharpness measurements. Despite that, the overall performance of *Asym\_shrp* was still poor.

The performance of DFFS on alignment errors was consistent but generally lower than the proposed method. Notably, DFFS failed to detect images with the best sharpness.

**Table 2.** Quality assessment of alignment errors and sharpness variations on FERET ‘fb’ and all six PIE illumination subsets. Each value in the table indicates the percentage of the best aligned image in each variation type being assigned to have the highest quality score. For example, out of the set of faces with rotations of  $0^\circ$ ,  $\pm 10^\circ$ ,  $\pm 20^\circ$ ,  $\pm 30^\circ$ , the value indicates the percentage of  $0^\circ$  faces labelled as “high” quality. The variations included: horizontal shift (HS), vertical shift (VS), in-plane rotation (RT), scale (SC), sharpness (SH). The ‘overall’ columns indicate the average performance of the above variations. Best performance is highlighted in bold.

	FERET ‘fb’						PIE illumination					
	HS	VS	RT	SC	SH	overall	HS	VS	RT	SC	SH	overall
Asym_shrp [26]	44.4	7.7	79.8	7.4	<b>100.0</b>	47.9	10.3	4.0	40.4	2.4	<b>100.0</b>	31.4
Gabor_asym [29]	52.1	3.1	93.9	11.5	49.0	41.9	24.7	1.5	66.4	10.7	29.0	26.5
DFFS [5]	75.6	71.9	98.7	62.5	0.7	61.9	64.4	62.4	<b>99.6</b>	<b>44.4</b>	2.3	54.6
Proposed	<b>83.4</b>	<b>85.4</b>	<b>99.6</b>	<b>73.0</b>	99.8	<b>88.2</b>	<b>65.9</b>	<b>62.6</b>	98.8	37.0	95.9	<b>72.0</b>

**Table 3.** Quality assessment of pose variations on the pose subsets of FERET and PIE. Each value in the table indicates the percentage of images with a particular pose angle that were assigned to have the highest quality score. Best performance is highlighted in bold.

	FERET pose subset								
	−60°	−40°	−25°	−15°	0°	+15°	+25°	+40°	+60°
Asym.shrp [26]	0	0	0.5	30.5	68.0	1	0	0	0
Gabor.asym [29]	2	5.5	7.5	24.5	58.0	2.5	0	0	0
DFFS [5]	0	0	0	5	<b>92.0</b>	3	0	0	0
Proposed	0	0	0.5	28	68.5	3	0	0	0

	PIE pose subset								
	−67.5°	−45°	−22.5°	—	0°	—	+22.5°	+45°	+67.5°
Asym.shrp [26]	0	0	2.94	—	<b>94.1</b>	—	1.5	1.5	0
Gabor.asym [29]	0	8.8	10.3	—	73.5	—	5.9	1.5	0
DFFS [5]	0	1.5	11.8	—	79.4	—	7.4	0	0
Proposed	0	0	4.4	—	91.2	—	4.4	0	0

## 4.2. Quality Assessment on Pose Variations

In this experiment we evaluated the ability of each method to detect the most frontal faces in a set that included frontal and non-frontal (out-of-plane rotated) faces. The results, shown in Table 3, indicate that the proposed method consistently achieves second best performance on both FERET and PIE, with its performance on PIE being quite close to the top performer (Asym\_shrp).

We note that on FERET the visual differences between faces at  $0^\circ$  and  $\pm 15^\circ$  are minimal, which can explain why a significant proportion of faces at  $-15^\circ$  was classified as “frontal” by the proposed method.

While DFFS gave the best performance on FERET, its performance dropped on PIE. As there is an overlap between the subjects in the ‘fa’ and pose subsets in FERET (where ‘fa’ was used for training), the inconsistency in performance across FERET and PIE suggests that DFFS might be over trained to the training dataset.

The performance of Asym\_shrp and the proposed method is considerably better on PIE than on FERET. We conjecture that this is due to the larger pose variation between frontal faces and faces with the smallest pose angle ( $\pm 22.5^\circ$ ), in contrast to  $\pm 15^\circ$  on FERET.

**Table 4.** Quality assessment of images with cast shadows from the PIE dataset. Each value in the table indicates the percentage of images with a particular illumination direction that were assigned to have the highest quality score. The illumination ranged from frontal (subset 1) to strongly directed (subset 6) where there are strong shadows (see Fig. 2).

	PIE illumination subset					
	1	2	3	4	5	6
Asym_shrp [26]	<b>97.1</b>	2.9	0	0	0	0
Gabor_asym [29]	51.5	5.9	2.9	39.7	4.4	0
DFFS [5]	0	0	4.4	88.2	7.4	0
Proposed	94.1	5.9	0	0	0	0

### 4.3. Quality Assessment on Cast Shadow Variations

Here we evaluated the accuracy of selecting frontal face images with the least amount of cast shadow within a set of images subject to varying illumination direction. The direction ranged from frontal (subset 1) to side (subset 6), where severe cast shadows exist (as shown in Fig. 2).

The results, presented in Table 4, show that Asym\_shrp achieved the best performance (correctly labelling frontally illuminated faces as having high quality), with the proposed method a close second. In contrast, Gabor\_asym was confused between subsets 1 and 4, while DFFS erroneously labelled most faces in subset 4 (containing significant shadows) as having the highest quality.

## 5. Experiments on Video: Subset Selection

In this section, we study the effectiveness of using quality measurements to select a subset of images for video-based face verification. To demonstrate the effectiveness of the quality assessment for a variety of face recognition systems, we used two facial feature extraction algorithms and two classification techniques, specifically designed for dealing with sets of faces (ie. image set matching).

Specifically, we separately used Multi-Region Histograms (MRH) [28] and Local Binary Patterns (LBP) [4] to extract features from each face. The comparison between two sets of faces was performed using (i) Mutual Subspace Method (MSM) [37] (for both MRH and LBP), and (ii) feature averaging [8, 20] (for MRH only).

The experiments were conducted on the ChokePoint dataset, using the video-to-video protocol (see Sec. 3.2). Each set of face images for a particular person was rank ordered according to the quality scores of the images, followed by keeping the top  $N$  images.

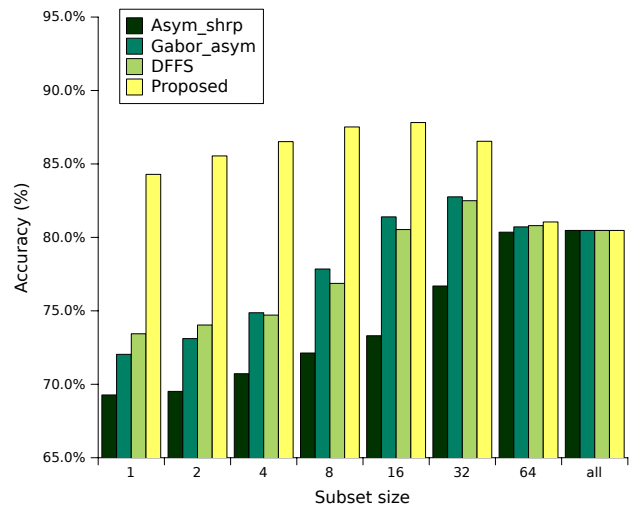
As per Section 4, the proposed face quality measurement method was compared against three other methods: Asym\_shrp, Gabor\_asym and DFFS. The ‘fa’ subset of FERET, which is totally independent from ChokePoint, was used for training DFFS and the proposed quality measurement method.

In the first experiment,  $N$  varied from 4 to 16. The results, reported in Table 5, indicate that the proposed quality measurement method consistently leads to better face verification performance than the other three methods, regardless of the facial feature extraction algorithm used. The improvement is most prevalent for  $N = 4$ , indicating that the proposed method assigns high scores to high quality images more accurately.

In the second experiment,  $N$  varied from 1 to the size of the set (labelled as “all”). Each face set was represented by an average MRH signature; LBP feature extraction was not used as it isn’t suitable for feature averaging. Face sets were then compared by using an  $L_1$ -norm based distance between their corresponding average MRH signatures [8, 20, 28].

**Table 5.** Video-based face verification performance on the *ChokePoint* dataset, using MRH and LBP feature extraction algorithms coupled with the Mutual Subspace Method (MSM) for classifying face sets. Each set of face images for a particular person was rank ordered according to the quality scores of the images, followed by retaining top  $N$  quality images (ie.  $N$  is the subset size). Faces were segmented using automatic face localisation (detection). The average face verification rate is reported (see Sec. 3.2). Best performance is highlighted in bold.

Subset Selection Method	Recognition Method					
	MRH + MSM			LBP + MSM		
	N=4	N=8	N=16	N=4	N=8	N=16
Asym_shrp [26]	67.5	70.3	75.4	65.3	67.6	70.5
Gabor_asym [29]	75.4	78.6	84.0	69.3	71.4	74.5
DFFS [5]	74.7	78.1	83.4	69.4	70.3	74.6
Proposed	<b>82.5</b>	<b>84.5</b>	<b>86.7</b>	<b>73.5</b>	<b>74.7</b>	<b>75.8</b>



**Figure 5.** Video-based face verification performance on the *ChokePoint* dataset using average MRH signatures. Each set of face images for a particular person was rank ordered according to the quality scores of the images, followed by selecting a predefined number of top quality images to create a subset. Faces were segmented using automatic face localisation (detection). The average face verification rate is reported (see Sec. 3.2).

From the results shown in Fig. 5, it can be observed that using all captured faces generally does not lead to the best performance. It can also be observed that the proposed method considerably outperforms the other three methods for  $N \leq 32$ , and furthermore leads to the best verification performance (which occurs at  $N = 16$ ). We note that even when only one face is selected by the proposed method (ie.  $N = 1$ ), relatively high verification accuracy is still achieved. This suggests that the proposed method has a high chance of picking the “best” face out of a set of faces.

## 6. Main Findings

In this paper we presented a novel patch-based face image quality assessment algorithm. Unlike previous methods, the proposed approach is capable of simultaneously handling issues such as pose variations, cast shadows, blurriness as well as alignment errors caused by automatic face localisation (eg. in-plane rotations, horizontal and vertical shifts).

The proposed method was evaluated on two still face datasets (FERET and PIE), using faces subject to pose and illumination direction changes, as well as simulated geometric alignment errors and decreased sharpness. Experiments show that the proposed method has the best overall performance, identifying images which are the most frontal, well-aligned, illuminated and sharp. This is accomplished without requiring parameter tuning or retraining for each dataset tested.

The proposed method was also evaluated in a video-based face verification setting, on a new surveillance dataset termed *ChokePoint*. For each given set of face images for a person, the proposed method was used to rank the images according to their quality. By selecting a subset containing only the top quality images, verification accuracy was considerably improved when compared to using all available images. Furthermore, the proposed method consistently led to higher quality subsets (leading to higher verification accuracy) than previous image quality assessment algorithms,

The proposed method is capable of assigning low-quality scores to images with cast shadows (eg. due to self-shadowing caused by strong directed illumination), however it is currently unlikely to detect more subtle variations in illumination. This is due to its elaborate illumination normalisation steps, necessary for generalisation purposes (ie. not being tied to the level of contrast and/or illumination bias in a particular training dataset). The proposed method is also unlikely to detect minor expression variations, as only low frequency information is used. According to [3, 33], expression changes mainly lie in high frequency bands. However, many of the recent face recognition algorithms are capable of handling relatively minor variations in both illumination and expression [15, 17, 24, 28], thus these characteristics of the quality assessment method might be more of a feature than a limitation.

## Acknowledgements

NICTA is funded by the Australian Government as represented by the *Department of Broadband, Communications and the Digital Economy*, as well as the Australian Research Council through the *ICT Centre of Excellence* program. We thank Dr Mehrtash Harandi for useful discussions. We also thank all the volunteers who participated in the recording of the ChokePoint dataset.

## References

- [1] ISO/IEC 19794-5 (published version). *Information technology - Biometric Data Interchange Formats*, June 2005.
- [2] Machine readable travel documents. *International Civil Aviation Organization*, August 2006.
- [3] L. Aguado, I. Serrano-Pedraza, S. Rodriguez, and F. J. Roman. Effects of spatial frequency content on classification of face gender and expression. *The Spanish Journal of Psychology*, 13(2):525–537, 2010.
- [4] T. Ahonen, A. Hadid, and M. Pietikäinen. Face recognition with local binary patterns. In *ECCV, Lecture Notes in Computer Science (LNCS)*, volume 3021, pages 469–481, 2004.
- [5] H. Bae and S. Kim. Real-time face detection and recognition using hybrid-information extracted from face space and facial features. *Image and Vision Computing*, 23(13):1181–1191, 2005.
- [6] S.-A. Berrani and C. Garcia. Enhancing face recognition from video sequences using robust statistics. In *IEEE International Conference on Video and Signal Based Surveillance (AVSS)*, pages 324–329, 2005.
- [7] L. Chang, I. Rodés, H. Méndez, and E. del Toro. Best-shot selection for video face recognition using FPGA. In *CIARP, Lecture Notes in Computer Science (LNCS)*, volume 5197, pages 543–550, 2008.
- [8] S. Chen, S. Mau, M. T. Harandi, C. Sanderson, A. Bigdeli, and B. C. Lovell. Face recognition from still images to video sequences: A local-feature-based framework. *EURASIP Journal on Image and Video Processing*, 2011.
- [9] W. Chen, M. J. Er, and S. Wu. Illumination compensation and normalization for robust face recognition using discrete cosine transform in logarithm domain. *IEEE Trans. Systems, Man and Cybernetics (Part B)*, 36(2):458–466, 2006.
- [10] X. Gao, S. Z. Li, R. Liu, and P. Zhang. Standardization of face image sample quality. In *ICB, Lecture Notes in Computer Science (LNCS)*, volume 4642, pages 242–251, 2007.
- [11] R. Gonzales and R. Woods. *Digital Image Processing*. Prentice Hall, 3rd edition, 2007.
- [12] R. Gottumukkal and V. K. Asari. An improved face recognition technique based on modular PCA approach. *Pattern Recognition Letters*, 25(4):429–436, 2004.
- [13] A. Hadid and M. Pietikäinen. From still image to video-based face recognition: An experimental analysis. In *Proc. Automatic Face and Gesture Recognition (AFGR)*, pages 813–818, 2004.
- [14] M. T. Harandi, M. N. Ahmadabadi, and B. N. Araabi. Optimal local basis: A reinforcement learning approach for face recognition. *International Journal of Computer Vision*, 81(2):191–204, 2009.
- [15] M. T. Harandi, C. Sanderson, S. Shirazi, and B. C. Lovell. Graph embedding discriminant analysis on Grassmannian manifolds for improved image set matching. In *IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pages 2705–2712, 2011.
- [16] R.-L. V. Hsu, J. Shah, and B. Martin. Quality assessment of facial images. In *Biometrics Symposium*, 2006.



- [17] N. Kumar, A. Berg, P. Belhumeur, and S. Nayar. Attribute and simile classifiers for face verification. In *Int. Conf. Computer Vision (ICCV)*, pages 365–372, 2009.
- [18] H. Luo. A training-based no-reference image quality assessment algorithm. In *International Conference on Image Processing (ICIP)*, pages 2973–2976, 2004.
- [19] S. Marcel, C. McCool, P. Matejka, T. Ahonen, J. Cernocky, S. Chakraborty, V. Balasubramanian, S. Panchanathan, C. Chan, J. Kittler, et al. On the results of the first mobile biometry (MOBIO) face and speaker verification evaluation. In *Recognizing Patterns in Signals, Speech, Images and Videos, Lecture Notes in Computer Science (LNCS)*, volume 6388, pages 210–225, 2010.
- [20] S. Mau, S. Chen, C. Sanderson, and B. C. Lovell. Video face matching using subset selection and clustering of probabilistic multi-region histograms. In *International Conference on Image and Vision Computing New Zealand (IVCNZ)*, 2010.
- [21] K. Nasrollahi and T. B. Moeslund. Face quality assessment system in video sequences. In *BIOID, Lecture Notes in Computer Science (LNCS)*, volume 5372, pages 10–18, 2008.
- [22] N. Ozay, Y. Tong, W. Frederick W, and X. Liu. Improving face recognition with a quality-based probabilistic framework. In *Computer Vision and Pattern Recognition (CVPR) Biometrics Workshop*, pages 134–141, 2009.
- [23] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss. The FERET evaluation methodology for face-recognition algorithms. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(10):1090–1104, 2000.
- [24] P. J. Phillips, W. T. Scruggs, A. J. O’Toole, P. J. Flynn, K. W. Bowyer, C. L. Schott, and M. Sharpe. FRVT 2006 and ICE 2006 large-scale experimental results. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(5):831–846, 2010.
- [25] Y. Rodriguez, F. Cardinaux, S. Bengio, and J. Mariethoz. Measuring the performance of face localization systems. *Image and Vision Computing*, 24(8):882–893, 2006.
- [26] E. A. Rúa, J. L. A. Castro, and C. G. Mateo. Quality-based score normalization and frame selection for video-based person authentication. In *BIOID, Lecture Notes in Computer Science (LNCS)*, pages 1–9, 2008.
- [27] C. Sanderson. Armadillo: An open source C++ linear algebra library for fast prototyping and computationally intensive experiments. Technical report, NICTA, 2010. <http://arma.sourceforge.net>
- [28] C. Sanderson and B. C. Lovell. Multi-region probabilistic histograms for robust and scalable identity inference. In *Lecture Notes in Computer Science (LNCS)*, volume 5558, pages 199–208, 2009.
- [29] J. Sang, Z. Lei, and S. Z. Li. Face image quality evaluation for ISO/IEC standards 19794-5 and 29794-5. In *ICB, Lecture Notes in Computer Science (LNCS)*, volume 5558, pages 229–238, 2009.
- [30] S. Shan, W. Gao, B. Cao, and D. Zhao. Illumination normalization for robust face recognition against varying lighting conditions. In *Workshop on Analysis and Modeling of Faces and Gestures (AMFG)*, pages 157–164, 2003.
- [31] T. Sim, S. Baker, and M. Bsat. The CMU pose, illumination, and expression database. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(1):1615 – 1618, 2003.
- [32] M. Subasic, S. Loncaric, T. Petkovic, H. Bogunovic, and V. Krivec. Face image validation system. In *International Symposium on Image and Signal Processing and Analysis (ISPA)*, pages 30–33, 2005.
- [33] J. D. Swisher, C. Brooking, and D. Somers. Spatial frequency and facial expressions of emotion. *Journal of Vision*, 4(8):905, 2004.
- [34] A. Torralba and P. Shina. Detecting faces in impoverished images. *Technical Report 028, MIT AI Lab*, 2001.
- [35] Y. Wong, C. Sanderson, S. Mau, and B. C. Lovell. Dynamic amelioration of resolution mismatches for local feature based identity inference. In *International Conference on Pattern Recognition (ICPR)*, pages 1200–1203, 2010.
- [36] X. Xie and K.-M. Lam. An efficient illumination normalization method for face recognition. *Pattern Recognition Letters*, 27:609–617, 2006.
- [37] O. Yamaguchi, K. Fukui, and K. Maeda. Face recognition using temporal image sequence. In *Proc. Automatic Face and Gesture Recognition (AFGR)*, pages 318–323, 1998.
- [38] Z. Yang, H. Ai, B. Wu, S. Lao, and L. Cai. Face pose estimation and its application in video shot selection. In *International Conference on Pattern Recognition (ICPR)*, pages 322–325, 2004.
- [39] G. Zhang and Y. Wang. Asymmetry-based quality assessment of face images. In *ISVC, Lecture Notes in Computer Science (LNCS)*, volume 5876, pages 499–508, 2009.
- [40] W. Zhao, R. Chellappa, A. Rosenfeld, and P. Phillips. Face recognition: A literature survey. *ACM Computing Surveys*, 35(4):399–458, 2003.