Hindawi Publishing Corporation
EURASIP Journal on Image and Video Processing
Volume 2011, Article ID 790598, 14 pages
doi:10.1155/2011/790598

*Research Article*

# Face Recognition from Still Images to Video Sequences: A Local-Feature-Based Framework

**Shaokang Chen,[1,2] Sandra Mau,[1,2] Mehrtash T. Harandi,[1,2] Conrad Sanderson,[1,2] Abbas Bigdeli,[1,2] and Brian C. Lovell[1,2]**

[1] *NICTA, St Lucia, QLD 4072, Australia*
[2] *School of ITEE, The University of Queensland, St Lucia, QLD 4072, Australia*

Correspondence should be addressed to Shaokang Chen, shaokang.chen@nicta.com.au

Received 30 April 2010; Revised 30 August 2010; Accepted 9 December 2010

Academic Editor: Luigi Di Stefano



Although automatic faces recognition has shown success for high-quality images under controlled conditions, for video-based recognition it is hard to attain similar levels of performance. We describe in this paper recent advances in a project being undertaken to trial and develop advanced surveillance systems for public safety. In this paper, we propose a local facial feature based framework for both still image and video-based face recognition. The evaluation is performed on a still image dataset LFW and a video sequence dataset MOBIO to compare 4 methods for operation on feature: feature averaging (Avg-Feature), Mutual Subspace Method (MSM), Manifold to Manifold Distance (MMS), and Affine Hull Method (AHM), and 4 methods for operation on distance on 3 different features. The experimental results show that Multi-region Histogram (MRH) feature is more discriminative for face recognition compared to Local Binary Patterns (LBP) and raw pixel intensity. Under the limitation on a small number of images available per person, feature averaging is more reliable than MSM, MMD, and AHM and is much faster. Thus, our proposed framework—averaging MRH feature is more suitable for CCTV surveillance systems with constraints on the number of images and the speed of processing.

## 1. Introduction

After the bombing attack in 2005, special attentions have been paid to the use of CCTV for surveillance to prevent such attacks in the future. Based on the number of CCTV cameras on Putney High Street, it is "guesstimated" [1] that there are around 500,000 CCTV cameras in the London area and 4,000,000 cameras in the UK. This implies that there is approximately one camera for every 14 people in the UK. Given the huge number of cameras, it is impossible to hire enough security guards to constantly monitor all camera feeds. Hence, generally the CCTV feeds are recorded without monitoring, and the videos are mainly used for a forensic or reactive response to crime and terrorism after it has happened. However, the immense cost of successful terrorist attacks in public spaces shows that forensic analysis of videos after the event is simply not an adequate response. In the case of suicide attacks, there is no possibility of prosecution after the event, so only recording surveillance video provides no terrorism deterrent. There is an emerging need to detect events and persons of interest from CCTV videos before any serious attack happens. This means that cameras must be monitored at all times.

However, two main constraints restrict human monitoring of the CCTV videos. One important issue is the limitation of the number of videos that a person can monitor simultaneously. For large amount of cameras, it requires a lot of people resulting in high ongoing costs. Another issue is that such a personnel intensive system may not be reliable due to the attention span of humans decreasing rapidly when performing such tedious tasks for long time. One possible solution is advanced surveillance systems that employ computers to monitor all video feeds and deliver the alerts to human operators for response. Because of this, there has been an urgent need in both the industry and the research community to develop advanced surveillance

systems, sometimes dubbed as Intelligent CCTV (ICCTV). In particular, developing total solutions for protecting critical infrastructure has been on the forefront of R&D activities in this field [2–4].

In 2009, NICTA was awarded a research grant to conduct long-term trials of Intelligent CCTV (ICCTV) technologies in important and sensitive public spaces such as major ports and railway stations. One of our research focuses is the person identification in crowded environment. Under the context of CCTV surveillance, recognition via faces appears to be the most useful among various biometric techniques. Our starting point is developing robust face recognition technique for public railway stations using existing cameras, which arises from problems encountered in our initial real-world trials. Though automatic face recognition has achieved satisfactory results under controlled conditions, video-based face recognition is considerably more challenging. Nuisance factors such as varying illumination, expression, and pose can greatly affect recognition performance. Figure 1 shows an example of real-life CCTV images captured at a railway station. In addition to robustness and accuracy, scalability and processing time are also important for surveillance systems. A face recognition system should be able to handle large volumes of people (e.g., peak hour at a railway station). Though this can be improved by parallel processing, there are always cost considerations limiting the number of CPUs available when dealing with large amount of video streams. In this context, a face recognition algorithm should be able to run in real time or better, which necessarily limits its complexity.

The outline of this paper is as follows: we review the state-of-the-art techniques for still image and video-based face recognition in Section 2, followed by discussions of still images and video sequences for surveillance in Section 3; we then proposed the Multiregion Histogram for still image face recognition in Section 4; the extension of MRH for video-based face recognition is presented in Section 5; Section 6 comes to the conclusion and future work.

## 2. Previous Approaches

*2.1. Still Image Face Recognition.* Research on still image face recognition has been done for nearly half a century. Two main approaches have been proposed for illumination invariant recognition. One is to represent images with features that are less sensitive to illumination changes [5, 6] such as the edge maps of the image. This approach suffers from the fact that features generated from shadows are related to illumination changes and may have an impact on recognition. Experiments done by Adinj et al. in [7] show that even with the best image representations, the misclassification rate is more than 20%. Another approach is to construct a low-dimensional linear subspace for images of faces taken under different lighting conditions [8, 9]. This approach is based on an assumption that images of a convex Lambertian object under variable illuminations form a convex cone in the space of all possible images [10]. Around 3 to 9 images are required to construct the convex cone. Nevertheless, the surface of human faces is not completely Lambertian reflected and convex. Therefore, it is hard for these methods to deal with cast shadows. Furthermore, these systems need several images of the same face taken under different controlled lighting source directions to construct a model of a given face.

As for expression invariant recognition, it is still unsolved for machine recognition and is even a difficult task for humans. In [11, 12], images are morphed to be the same shape as the one used for training. But it is not guaranteed that all images can be morphed correctly; for example, an image with closed eyes cannot be morphed to a neutral image because of the lack of texture inside the eyes. Liu et al. [13] propose to use optical flow for face recognition with facial expression variations. However, it is hard to learn the local motions within the feature space to determine the expression changes of each face, since the way one person expresses a certain emotion is normally somewhat different from others. Martinez proposed a weighing method to deal with facial expressions in [14]. An image is divided into several local areas, and those that are less sensitive to expressional changes are chosen and weighed independently. But features that are insensitive to expression changes may be sensitive to illumination variations [7].

Pose variability is usually considered to be the most challenging problem. There are three main approaches developed for 2D-based pose invariant face recognition. Wiskott et al. proposed Elastic Bunch Graph Matching, which applied Gabor filter to extract pose invariant features [15]. In [16–18] multiple-view templates are used to represent faces with different poses. Multiple-view approaches require several gallery images per person under controlled view conditions to identify a face, which restricts its application when only one image is available per person. Face synthesis methods have emerged in an attempt to overcome this issue. In [19], Gao et al. constructed a Face-Specific Subspace by synthesising novel views from a single image. In [20] a method for direct synthesis of face model parameters is proposed. In [21], an Active Appearance Model- (AAM-) based face synthesis method is applied for face recognition subject to relatively small pose variations. A recurring problem with AAM-based synthesis and multiview methods is the need to reliably locate facial features to determine the pose angle for pose compensation—this turns out to a be difficult task in its own right.

The above methods can handle certain kinds of face image variation successfully, but drawbacks still restrict their application. It may be risky to rely heavily on choosing invariant features [5, 6, 14, 15], such as using edge maps of the image or choosing expression insensitive regions. This is because features insensitive to one variation may be highly sensitive to other variations, and it is very difficult to abstract features that are completely immune to all kinds of variation [7]. Some approaches attempt to construct face-specific models to describe possible variations under changes in lighting or pose [8, 9, 19, 22]. Such methods require multiple images per person taken under controlled conditions to construct a specific subspace for each person for the face representation. This leads to expensive image

(a) (b)

Figure 1: Sample images captured by CCTV cameras installed at a railway station.

capture processes, poor scalability of the face model, and does not permit applications, where only one gallery image is available per person. Other approaches divide the range of variation into several subranges (e.g., low, medium, and high pose angles) and construct multiple face spaces to describe face variations lying in the corresponding subrange [16–18]. These approaches require us to register several images representing different variations per person into the corresponding variation models so that matching can be done in each interval individually. Once again, acquiring multiple images per person under specific conditions is often very difficult, if not impossible, in practice.

*2.2. Video-Based Face Recognition.* In recent years, increasing attention has been paid to the video-based face recognition. Many approaches were proposed to use temporal information to enhance face recognition for videos. One direct approach is temporal voting. A still image-matching mechanism is proposed by Satoh for matching two video sequences [23]. The distance between two videos is the minimum distance between two frames across two videos. Zhou and Chellappa presented a sequential importance sampling (SIS) method to incorporate temporal information in a video sequence for face recognition [24]. A state space model with tracking state vector and recognizing identity variable was used to characterize the identity by integrating motion and identity information over time. However, this approach only considers identity consistency in temporal domain, and thus it may not work well when the face is partially occluded. Zhang and Martinez applied a weighted probabilistic approach on appearance face models to solve the occlusion problem [25]. Their experiment shows that this approach can improve the performance for PCA, LDA, and ICA. The approach proposed in [26] uses the condensation algorithm to model the temporal structures.

Some approaches utilize spatial information by considering frames from videos as still image sets without considering their temporal information. Person-specific models are trained from video sequences to form many individual eigenspaces in [27]. Angles between subspaces are considered as the similarity between videos. In [28], each person is represented by a low-dimensional appearance manifold learned from training exemplars sampled from videos. The probabilistic likelihood of the linear models is propagating through the transition matrix between different pose manifolds. An exemplar-based probabilistic approach is proposed in [29], in which representative face images are selected as exemplars from training videos by radial basis functions. This approach can model small 2D motion effectively, but it cannot handle large pose variation or occlusion. Topkaya and Bayazit applied dimensional analysis on the representative frames selected based on facial features and the corresponding positions [30].

Most of the recent approaches utilize spatiotemporal information for face recognition in video. A sparse representation of face is learned from video for online face recognition under unconstrained conditions [31]. Principal component null space analysis (PCNSA) is proposed in [32], which is helpful for nonwhite noise covariance matrices. The Autoregressive and Moving Average (ARMA) model method is proposed in [33] to model a moving face as a linear dynamical object. Liu and Chen proposed an adaptive Hidden Markov Model (HMM) on dynamic textures for video-based face recognition. Kim et al. applied HMM to solve the visual constraints problem for face tracking and recognition [34].

The above approaches for face recognition in video have several main drawbacks. Firstly, personal specific facial dynamics are useful to discriminate different persons, but the intrapersonal temporal information that related to facial expression and emotions is also encoded and used; secondly, normally consistent weights are assigned to spatiotemporal features from the observation that some features are more helpful for recognition, but the weights are not adaptively assigned which may be harmful when face appearance changes dramatically, especially in the case of occlusion, where some features may disappear; thirdly, most of the methods require well-aligned faces, which limits their usage in practice; last but not least, most of the above approaches utilize holistic facial features, but the local facial features are

Figure 2: Normalised still face images captured by normal cameras.

not well investigated, which is shown to be useful for image analysis and face recognition on still images.

## 3. Still Image versus Video Sequence

For face recognition in surveillance scenarios, identifying a person captured on image or video is one of the key tasks. This implies matching faces on both still images and video sequences. It can be further classified into three categories: still image to still image matching, video sequence to video sequence matching, and still image to video sequence matching.

Automatic face recognition for still images with high quality can achieve satisfactory performance, but for video-based face recognition it is hard to attain similar levels of performance. Compared to still images face recognition, there are several disadvantages of video sequences. First, images captured by CCTV cameras are generally of poor quality. The noise level is higher, and images may be blurred due to movement or the subject being out of focus. Second, image resolution is normally lower for video sequences. If the subject is very far from the camera, the actual face image resolution can be as low as 64 by 64 pixels. Last, face image variations, such as illumination, expression, pose, occlusion, and motion, are more serious in video sequences. These effects are illustrated in Figure 3. 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. To comparison with the still image shown in Figure 2, it can be seen that the image quality of CCTV cameras (even high-end ones) is much worse than still images. In addition, the poor quality, low resolution, and large variation will result in uncertainty of the face detector, which is the first important step of any automatic face recognition system. Faces extracted from poor-quality videos can have higher false detection rate and larger alignment errors, which may have great influence on the performance [35].

However, there are some major advantages of video sequences. First, we can employ spatial and temporal information of faces in the video sequence to improve still images recognition performance. Second, psychophysical and neural studies have shown that dynamic information is very crucial in the human face recognition process [36]. Third, with redundant information, we can reconstruct more complex representations of faces such as a 3D face model [37] or super-resolution images [38] and apply them to improve recognition performance. Fourth, some online learning techniques can be applied for video-based face recognition to update the model over time [39].

CCTV images with relatively better quality

(a)

CCTV images with degraded quality

(b)

Figure 3: Normalised video face images captured by CCTV cameras.

Since we need to do both still image and video-based face recognition under surveillance conditions, the above approaches are not suitable. Most still image face recognition techniques are not appropriate for surveillance images due to the following concurrent and uncontrolled factors. The pose, illumination, and expression variations are shown to have great impact on face recognition [40]. Image resolution change due to variable distances to cameras is another factor that influences the recognition performance [41]. The face localization error induced by automatic face detector will definitely affect the recognition results as there are no guarantees that the localization is perfect (e.g., misalignment or wrong scale) [42]. In addition to image properties, a surveillance system may have further constraints: limitation in number of images, for example,

only one gallery image per person, as well as real-time operation requirements in order to handle large volumes of people. As many still image face recognition techniques are restricted to medium to high resolution face images and require expensive computation or multiple gallery images, which are not applicable for surveillance. Most of the video-based face recognition approaches are designed for video to video match, which cannot be used for still image recognition. Moreover, the above approaches rely heavily on the good face detection and feature localization, which is impractical under surveillance conditions, where images are of low resolution and processing should be in real-time. We thus develop a framework for both still image and video based face recognition under surveillance scenarios using local facial features. This approach can handle low resolution face image recognition with pose, illumination, and expression variations to a certain degree and is not sensitive to localization errors. Moreover, the computation for this approach is fast enough for real-time processing.

## 4. Multiregion Histogram for Still Image Face Recognition

In this section, we describe a Multiregion Histogram-(MRH-) [43] based approach with the aim of concurrently addressing the above-mentioned problems.

*4.1. Multiregion Histograms of Visual Words.* Each face is divided into $R$ fixed and adjacent regions. Each region is further divided into overlapped small blocks. Each block has a size of $8 \times 8$ pixels and overlaps neighbouring blocks by 75%. 2D DCT [44] decomposition is applied on each block to extract descriptive features. To compensate for varying contrast, each block is normalized to have zero mean and unit variance. Based on preliminary experiments we elected to retain 15 low-frequency elements out of the 64 DCT coefficients, by taking the top-left $4 \times 4$ submatrix of the $8 \times 8$ coefficient matrix and disregarding the first coefficient (as it is the mean of the block and is normalized to zero). Therefore, for region $r$ a set of feature vectors is obtained, $X_r = \{\mathbf{x}_{r,1}, \mathbf{x}_{r,2}, \ldots, \mathbf{x}_{r,N}\}$, where $N$ is the number of blocks. For each vector $\mathbf{x}_{r,i}$ obtained from region $r$, a probabilistic histogram is computed:

$$\mathbf{h}_{r,i} = \left[ \frac{w_1 p_1(\mathbf{x}_{r,i})}{\sum_{g=1}^{G} w_g p_g(\mathbf{x}_{r,i})}, \frac{w_2 p_2(\mathbf{x}_{r,i})}{\sum_{g=1}^{G} w_g p_g(\mathbf{x}_{r,i})}, \ldots, \frac{w_G p_G(\mathbf{x}_{r,i})}{\sum_{g=1}^{G} w_g p_g(\mathbf{x}_{r,i})} \right]^T, \tag{1}$$

where the $g$th element in $\mathbf{h}_{r,i}$ is the posterior probability of $\mathbf{x}_{r,i}$ according to the $g$th component of a visual dictionary model. The visual dictionary model is built from a convex mixture of Gaussians [45], parameterised by $\lambda = \{w_g, \boldsymbol{\mu}_g, \mathbf{C}_g\}_{g=1}^{G}$, where $G$ is the number of Gaussians, while $w_g$, $\boldsymbol{\mu}_g$, and $\mathbf{C}_g$ are the weight, mean vector, and covariance matrix for Gaussian $g$, respectively. The mean of each Gaussian can be regarded as a particular "visual word."

Once the histograms are computed for each feature vector from region $r$, an average histogram for the region is built:

$$\mathbf{h}_{r,\mathtt{avg}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{h}_{r,i}. \tag{2}$$

The DCT decomposition acts like a low-pass filter, which retained features robust to small alterations due to in-plane rotations, expression changes, or smoothing due to upsampling from low-resolution images. The overlapping of blocks during feature extraction, as well as the loss of spatial relations within each region (due to averaging), results in robustness to translations of the face which are caused by imperfect face localization. We note that in the $1 \times 1$ region configuration (used in [46]) the overall topology of the face is effectively lost, while in configurations such as $3 \times 3$ it is largely retained (while still allowing for deformations in each region).

The visual dictionary is obtained by pooling a large number of feature vectors from training faces, followed by employing the Expectation Maximisation algorithm [45] to optimise the dictionary's parameters (i.e., $\lambda$).

*4.2. Normalised Distance.* Comparison of two faces is accomplished by comparing their corresponding average histograms. Based on [47] we define an $L_1$-norm-based distance measure between faces $A$ and $B$:

$$d_{\mathtt{raw}}(A, B) = \frac{1}{R} \sum_{r=1}^{R} \left\| h_{r,\mathtt{avg}}^{[A]} - h_{r,\mathtt{avg}}^{[B]} \right\|_1. \tag{3}$$

$d_{\mathtt{raw}}(A, B)$ is compared to a threshold to determine whether faces $A$ and $B$ come from the same person or from two different people. However, the threshold might be dependent on the image conditions of face $A$ and/or $B$, which are not known a priori. We propose a normalised distance in order to reduce the sensitivity of threshold selection:

$$d_{\mathtt{normalised}}(A, B) = \frac{d_{\mathtt{raw}}(A, B)}{(1/2)\left( (1/M) \sum_{i=1}^{M} d_{\mathtt{raw}}(A, C_i) + (1/M) \sum_{i=1}^{M} d_{\mathtt{raw}}(B, C_i) \right)}, \tag{4}$$

where $C_i$ is the $i$th cohort face and $M$ is the number of cohorts. In the above equation cohort faces are assumed to be reference faces that are known not to be of persons depicted in $A$ or $B$. As such, the terms $(1/M) \sum_{i=1}^{M} d_{\mathtt{raw}}(A, C_i)$ and $(1/M) \sum_{i=1}^{M} d_{\mathtt{raw}}(B, C_i)$ estimate how far away, on average, faces $A$ and $B$ are from the face of an impostor. This typically results in (4) being approximately 1 when $A$ and $B$ represent faces from two different people and less than 1 when $A$ and $B$ represent two instances of the same person. If the conditions of given images cause their raw distance to increase, the average raw distances to the cohorts will also increase. As such, the division in (4) attempts to cancel out the effect of varying image conditions.

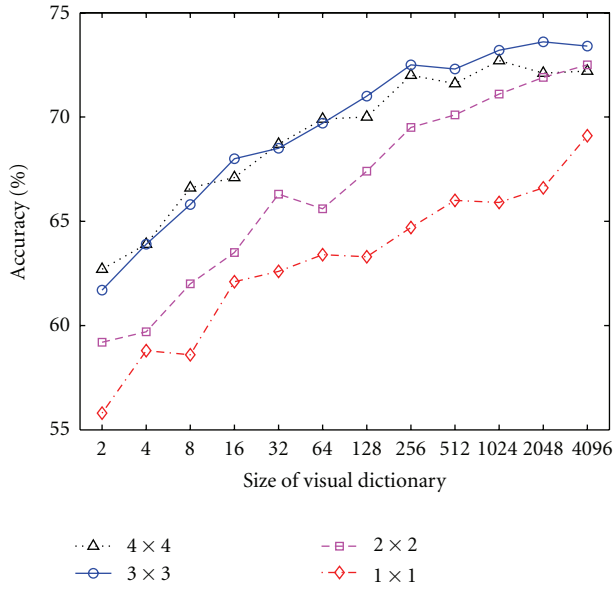



FIGURE 4: Accuracy rate for different number of regions with various size of the visual dictionary on view 1 of LFW.

*4.3. Empirical Evaluation.* This approach is evaluated on LFW dataset [48] which contains 13,233 face images with variations in pose, illumination, expression, in-plane rotation, resolution, and localization (resulting in scale or translation error). The images of LFW were obtained from the Internet, and faces were centered, scaled, and cropped based on bounding boxes provided by an automatic face locator. We normalize the extracted faces to $64 \times 64$ pixels, with an average distance between eyes of 32 pixels.

The test protocol of LFW is verification based, which is to classify whether a pair of previously unseen faces is of the same person (matched pair) or two different persons (mismatched pair). The protocol specifies two views of the dataset: *view 1*, aimed at algorithm development and model selection, and *view 2*, aimed at final performance reporting. There are 1100 matched and 1100 mismatched pairs in training set and 500 unseen matched and 500 unseen mismatched pairs in the test set in view 1. We use training set to construct the visual dictionary as well as optimizing the threshold. In view 2 the images are split into 10 sets, with each set 300 matched and 300 mismatched pairs. A 10-fold cross-validation is done by using 9 for training and 1 for testing for each of the subset, respectively. Performance is evaluated by the mean and standard error of the average accuracies for all 10 subsets. The standard error is useful for assessing the significance of performance differences across algorithms [48].

We first studied the effect of increasing the size of the visual dictionary (from 2 to 4096 components) and number of regions (from $1 \times 1$ to $4 \times 4$) on the LFW view 1. Based on preliminary experiments, we randomly selected 32 cohort faces from the training set for distance normalization. The results shown in Figure 4 suggest that performance increases continuously up to about 1024 components, then performance becomes steady with only minor change. Significant improvements are observed when the number of regions rises from $1 \times 1$ to $3 \times 3$. Whilst using more regions (i.e., $4 \times 4$) shows no further performance gains.

We then fixed the number of regions to be $3 \times 3$ and tested it on view 2 of LFW. Several configurations of MRH were evaluated in comparison with PCA (as a baseline) and RBT. Based on preliminary experiments, the baseline PCA-based system used the Euclidean distance as its raw distance. Following the suggestion in [49], 61 eigenfaces (eigenfaces 4 to 64 of the training images) were used, ignoring the first three eigenfaces. Results for RBT are obtained from http://vis-www.cs.umass.edu/lfw, using the method published in [50]. By searching the literature, we also compare with other recent methods using the following four kinds of features: Local Binary Patterns (LBP) [51], Gabor Jets Descriptors (GJD) [52], Scale Invariant Feature Transform (SIFT) [53], and the Learning-based descriptor (LE) [54]. The results for LBP, GJD, and SIFT (denoted as LBP-MSE-80, GJD-BC-100 and SIFT-SIMPLE individually in Table 1) are from [55], and the result for LE is obtained from [54]. The results shown in Table 1 indicate that the performance of MRH based systems is consistent with the previous experiment. Furthermore, the probabilistic $3 \times 3$ MRH method is much better than those methods using LBP, GJD, and SIFT features and is comparable with the more complex RBT method. The performance of PCA considerably lags behind all other approaches. The latest approach using LE features performs the best by learning comprehensive descriptors and encoder from training samples.

## 5. Enhancement of MRH for Video-based Face Recognition

For intelligent surveillance systems, automatic face recognition should be performed for both still images and video sequences. Thus, normal video-based face recognition techniques are not suitable for this task since they are designed only for video-to-video matching. In an attempt to retain the ability for still image face recognition and to be capable for still-to-video and video-to-video matching, we propose the following approaches to enhance MRH for face recognition on videos. In this section, we explore four methods that operate on features to build up a more representative model for classification as well as four methods that operate on distance between vectors to improve the performance. By investigating these approaches, we attempt to choose a best suitable method that takes advantage of multiframe information in a computationally inexpensive manner for image-set and video-set matching. As part of the investigation into this problem, a subset of LFW database is used for image set matching, test and a large-scale audiovisual database called "Mobio Biometry" (MOBIO) [56] is used for video-set matching, respectively.

*5.1. Operation on Feature.* In this approach, several methods are inspected, which utilize multiple feature vectors of

Table 1: Results on *view 2* of LFW. MRH approaches used a 1024-component visual dictionary.

| Method | Mean accuracy | Standard error |
|---|---|---|
| $3 \times 3$ MRH (probabilistic, normalised distance) | 72.95 | 0.55 |
| $3 \times 3$ MRH (probabilistic, raw distance) | 70.38 | 0.48 |
| $1 \times 1$ MRH (probabilistic, normalised distance) | 67.85 | 0.42 |
| PCA (normalised distance) | 59.82 | 0.68 |
| PCA (raw distance) | 57.23 | 0.68 |
| Randomised Binary Trees (RBT) | 72.45 | 0.40 |
| LBP-MSE-80 | 65.27 | 0.47 |
| GJD-BC-100 | 67.98 | 0.65 |
| SIFT-SIMPLE | 62.95 | 0.71 |
| Single LE + holistic | 81.22 | 0.53 |

the sample images in a set to build up a more representative model of faces. In other words, they attempt to extract more meaningful new features from the existing features. In the following sections, we will discuss them in more detail.

*5.1.1. Feature Averaging.* To extend still image face recognition for video sequences, a direct approach is applying still image recognition for each frame in the video set. But this approach is computationally expensive and does not fully utilize spatial and temporal information of the video. Given an example, to identify a face from a probe video with $f$ frames in a video database with $V$ video sequences, the thorough search needs to perform the still image matching by $V \times m \times v$ times, where $v$ is the average frames per sequence. Generally, for only a 10-second video, it would contain about 300 frames (with a normal frame rate at 30 fps). This means that the calculation for video is about 90000 times of that for still image.

Inspired by a recent paper published in Science [57], we propose the following approach by averaging MRH facial features. Different from [57], where a simple image averaging is applied, we average the features due to the observation that image averaging is only helpful for holistic facial features and impairs the local facial features [58]. Assume that MRH is applied on frame $k$ of video $p$ to extract the histogram $h_{p,k}$. Then the final description of the face in this video is by averaging as follows:

$$h_p = \frac{1}{v_p} \sum_{k=1}^{v_p} h_{p,k}, \tag{5}$$

where $v_p$ is the number of selected frames for video $p$. By the above averaging, we statistically average both spatial and temporal information of faces. The average over frames straightly integrates temporal information, and the region averaging of MRH accomplishes spatial merge.

The similarity measure between two videos is the normalized distance between the average histograms as defined in (4). As can be easily seen, with this averaging approach, the recognition is done by only $V$ times distance calculation, comparable to the still image recognition.

*5.1.2. Manifold Distance.* Manifold distance is an emerging area in image-set matching [59]. Mutual Subspace Method (MSM) was one of the earliest (and still competitive) approaches within this school of thought [60]. In MSM the principal angle between two subspaces is considered as the similarity measure. The basic idea of principal angles has been extended into kernel principal angles [61] or discriminative canonical correlation [62] with promising results.

Assume that $W \in i^{d \times n_1}$ and $X \in i^{d \times n_2}$ are two linear subspaces with minimum rank $r = \min(n_1, n_2)$. Then there are exactly $r$ uniquely defined canonical correlations between $W$ and $X$ as follows:

$$\cos \theta_i = \max_{w_i \in W} \max_{x_i \in X} w_i^T x_i, \tag{6}$$

where $w_i^T w_i = x_i^T x_i = 1$ and $w_i^T w_j = x_i^T x_j = 0, i \neq j$. $\theta_i$ is the principal angle between the two subspaces and $0 \leq \theta_i \leq \pi/2, i \in [1 \cdot r]$. One straight and numerical robust way to compute the canonical correlations is based on Singular Value Decomposition (SVD). Considering that $O_1$ and $O_2$ are orthogonal bases for subspaces $W$ and $X$, respectively, the canonical correlations are the singular values of $O_1^T O_2$. In MSM the largest eigenvalue is used as the distance between two manifolds.

A more comprehensive extension of MSM is the Manifold to Manifold Distance (MMD) proposed in [59]. A Maximal Linear Patch (MLP) method is used to cluster the sample images in the data set to form several local linear patches (linear subspaces). The MMD of two image sets is the minimal distance between the MLPs of these two sets; that is,

$$d(M_1, M_2) = \min_{C_i \in M_1} \min_{C_j' \in M_2} d\left(C_i, C_j'\right), \tag{7}$$

where $C_i$ and $C_j'$ are the MLPs of image set $M_1$ and $M_2$, respectively.

*5.1.3. Affine Hull Method.* The Affine Hull Method proposed in [63] represents images as points in a linear feature space and each image set is approximated by a convex geometric region, named as affine hull, spanned by its feature points.

Table 2: Verification results for image-set matching of LFW.

| Features | Number of images | Operation on feature | | | | Operation on distance | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | MSM | MMD | AHM | Avg-feature | Min-min | Max-min | Avg-min | Min-avg |
| MRH | 3 | 86.45 | 86.45 | 81.62 | **88.06** | 86.77 | 77.10 | 84.84 | 84.19 |
| | 4 | 90.28 | 88.89 | 90.74 | **92.59** | 89.35 | 78.70 | 88.43 | 87.04 |
| LBP | 3 | 75.81 | 75.81 | 73.23 | 77.74 | 77.42 | 67.74 | **78.39** | 75.16 |
| | 4 | 78.70 | 80.09 | 81.48 | **83.80** | 80.09 | 67.13 | 79.17 | 76.39 |
| Pixel intensity | 3 | 65.48 | 57.87 | **66.13** | 61.29 | 59.03 | 58.39 | 60.0 | 56.13 |
| | 4 | **72.69** | 67.13 | 68.52 | 67.13 | 64.35 | 58.80 | 63.89 | 63.89 |

The dissimilarity of two sets is measured by the geometric distances between two convex models. Given a face image vector $x_{ci}$, where $c = [1, \ldots, C]$ is the index of the $C$ image sets and $i = [1, \ldots, n_c]$ is the index of the $n_c$ samples of image set $c$; the affine hull that represents the images set is modelled as

$$H_c = \{x = \mu_c + U_c v_c\}, \quad \mu_c = \frac{1}{n_c}\sum_{k=1}^{n_c} x_{ck}, \tag{8}$$

where $U_c$ is the set of eigenvectors spanned by the affine subspace and $v_c$ is a vector of free parameters. The distance of two convex sets $H_i$ and $H_j$ is the infimum of the distances between any point in $H_i$ and any point in $H_j$; that is,

$$D\left(H_i, H_j\right) = \min_{x \in H_i, y \in H_j} ||x - y|| = \left|\left|(I - P)\left(\mu_i - \mu_j\right)\right|\right|, \tag{9}$$

where $P = U(U^T U)^{-1} U^T, U = (U_i - U_j)$.

*5.2. Operation on Distance.* In contrast to operation on features, the operation on distance applies some simple statistical analysis on the distance between feature vectors without generating new features. The following four methods are investigated:

$$\begin{aligned} \text{Min-Min}: &\ \min_i \min_j \left(d\left(w_i, x_j\right)\right), \\ \text{Max-Min}: &\ \max_i \min_j \left(d\left(w_i, x_j\right)\right), \\ \text{Avg-Min}: &\ \frac{1}{n_1}\sum_{i=1}^{n_1} \min_j \left(d\left(w_i, x_j\right)\right), \\ \text{Min-Avg}: &\ \min_i \frac{1}{n_2}\sum_{j=1}^{n_2} \left(d\left(w_i, x_j\right)\right), \end{aligned} \tag{10}$$

where $d(w_i, x_j)$ is the distance between $w_i$ and $x_j$.

*5.3. Empirical Evaluation.* The above approaches for video-based face recognition are evaluated on LFW and MOBIO datasets. For fair comparison, the four methods with operation on features are actually applied on three different facial features: MRH, Local Binary Patterns (LBP), and raw pixel intensity. The four methods for operation on distance are actually applied on the defined distance between these features. In the following section, we will describe the experiments on the above two datasets individually.

*5.3.1. LFW Multiple Images Set Match.* The image-set matching is evaluated in subsets of LFW. We follow the similar image pair verification protocol to LFW. We first evaluate the image-set matching with 3 images per set. 620 image-set pairs are generated from the LFW dataset, with 310 pairs for training and 310 pairs for testing. Each pair contains two image sets with 3 images in each set. Images in the testing are never included in the training. In order to remove bias, the pairs generated in our experiments are balanced so that the number of matched pairs and mismatched pairs is the same. Similarly, 432 pairs (216 training pairs and 216 testing pairs) are generated for image-set matching with 4 images in a set. We test the following four methods that operates on features: feature averaging (Avg-Feature), the manifold distance by applying MSM and MMD on facial features, and the affine hull method (AHM). To comprehensively investigate the influence of operation on features, we test the following three features: Multiregion Histogram (MRH), Local Binary Patterns (LBP) [51, 64], and raw pixel intensity. For comparison, we also apply four methods that operates on the distance between vectors (features) for thorough image-set matching.

Table 2 presents the verification accuracy for the above 8 methods on different features following LFW protocol. It can be seen from the result that those four methods of operation on features generally perform slightly better than the four methods of operation on distance. The accuracy of 4 images per set is a little better than using 3 images for all of the methods across all different features, which is predictable as more information is utilized. When scanning the table vertically, we notice that the performance of the above 8 methods on MRH feature (top two rows) is about 10 percent higher than LBP and 20 percent higher than pixel intensity, respectively. This indicates that MRH feature is more discriminative for face recognition. It is interesting that Avg-Feature performs the best among the 4 methods of operation on feature with MRH and LBP features. That might be because the face images extracted and normalized from video sequence have great variations, large alignment errors, and even false detections. MSM, MMD, and AHM

may model the faces biasedly with a small number of samples. The highest accuracy is obtained by averaging features (Avg-Feature) for both 3 and 4 images per set with MRH feature. The second best with 3 images per set is MSM, and the one for 4 images per set is AHM, which are 2 percent worse than Avg-Feature.

MMD is slightly worse than MSM because of the limitation on the number of images (only 3 or 4 images) per set. With only a few images per set, MSM would construct a more representative subspace than the subspaces modelled by MMD, because MSM uses all available features to construct a linear subspace, whilst MMD only uses a subset of the feature to construct several linear subspaces. Thus, the performance of MSM is slightly better than MMD in LFW. In some extreme cases, if only two images available per set, MMD cannot be applied to generate linear subspaces. MMD performs better than MSM only when there are much more images in an image set. Results reported in [59] shows improvement of MMD over MSM with 300 to 500 frames per video. Under the constraint that only few images (less than 10) are acquired for each set, MMD generally can not perform as good as MSM.

*5.3.2. MOBIO Videos Set Match.* The MOBIO dataset was collected as part of a European collaborative project to facilitate research and development on robust-to-illumination face and speaker video verification systems in uncontrolled environments. The quality of mobile images is generally poor with blurred images from motion and smudged lenses and changes in illumination between scenes, which is similar to those experienced in CCTV videos with out of focus, motion blur, and cameras with dirty lenses. The experiments used in this paper focused on only the development subset of the MOBIO database. In this subset, 1,500 probe videos are each compared to every person in the gallery (20 females and 27 males) whom each have 5 videos.

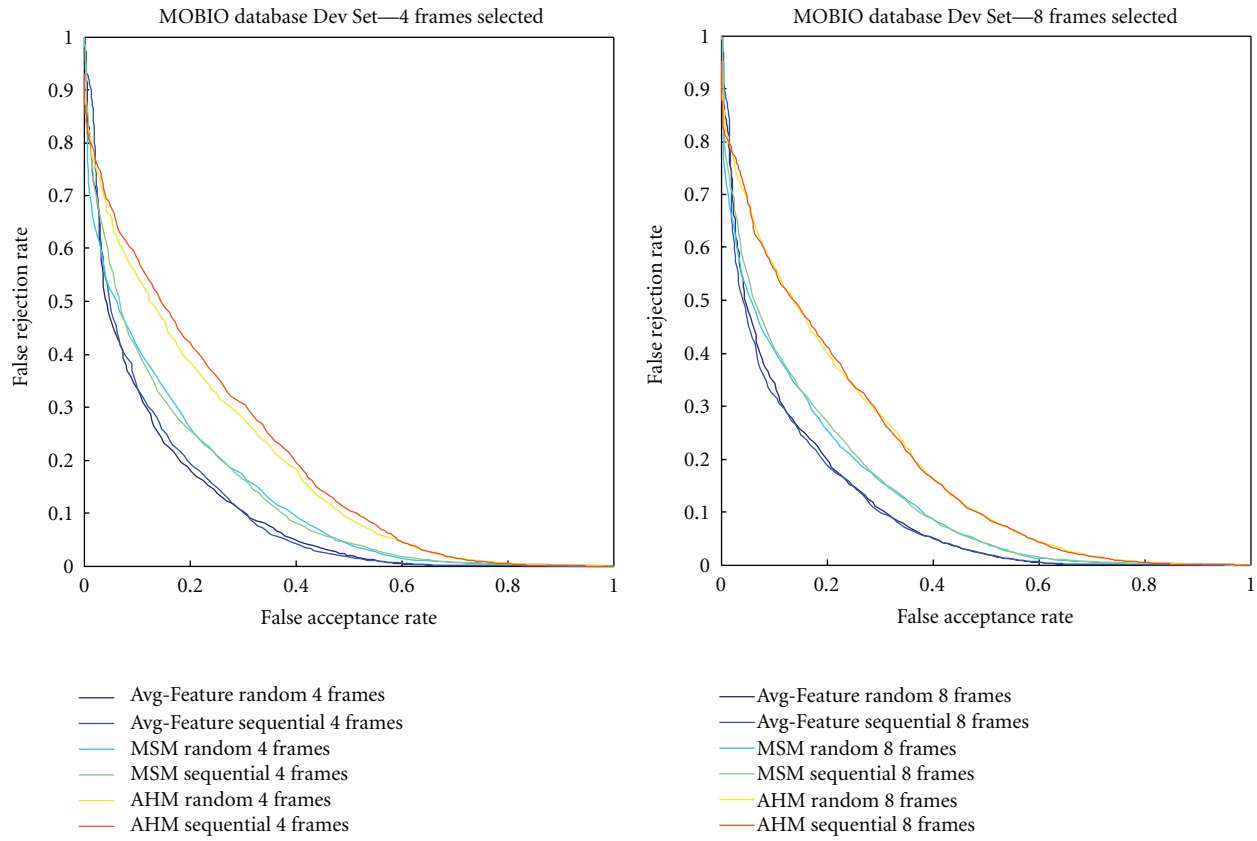
Because MOBIO database does not provide face locations, the OpenCV's Haar Feature-based Cascade Classifier [35] is used to detect faces in each frame. The faces are then tracked over multiple frames using Continuously Adaptive Mean-SHIFT Tracker (CAMSHIFT) [65] with colour histograms. Once the faces are detected, eyes are further located within the face using a Haar-based classifier. If no eyes are located, they are approximated from the size of the face detected. The faces are then resized and cropped such that the eyes are centered with a 32-pixel intereye distance. For these experiments, a closely cropped face of $64 \times 64$ pixels was used which excludes outer features surrounding the face such as hair and beard. In the surveillance context, such peripheral features can be easily used as disguises. Due to the low quality of the videos in MOBIO database and the robustness of the face detector, there are 7% of all the videos with less than or equal to 2 face images extracted.

Based on the observation in LFW test that methods of operation on features perform better than that of operation on distance and operations on MRH features outperform other features, in this video set match, we only evaluate operation on feature approaches for MRH. Due to the limitation of the face detector, less than 2 images with faces can be extracted from some videos, and MMD methods are not applicable to those videos. Thus, we only test the following three methods: MSM, AHM, and Avg-Feature.

Table 3: Average time for processing one video for different number of selected frames.

| Method | Number of frames | Average time |
|---|---|---|
| Avg-feature | 4 | 0.51 |
| | 8 | 0.51 |
| | 16 | 0.52 |
| MSM | 4 | 3.0 |
| | 8 | 5.48 |
| | 16 | 10.44 |
| AHM | 4 | 23.95 |
| | 8 | 44.95 |
| | 16 | 88.73 |

Figure 5 shows the Receiver Operating Characteristic (ROC) curve for the False Acceptance Rate (FAR) versus False Rejection Rate (FRR) of MSM, AHM and Avg-Feature on MOBIO dataset with 4 and 8 frames selected randomly and sequentially, respectively. Random selection is to choose frames in a video randomly with equal probability whilst, sequential selection is to choose frames with equal time interval. As can be seen from Figures 5(a) and 5(b), Avg-Feature performs the best among the three methods followed by MSM. The random selection and sequential selection have very limited influence on the performance of all of the three methods. In addition to the performance, we also evaluated the time cost for these three methods. Table 3 presents the average time of processing one video for MOBIO dataset with different number of selected frames. The Avg-Feature approach is much faster as all calculations are linear (euclidean averaging and L1-norm distances), whilst the time cost for MSM and AHM is much higher due to the complex Singular Value Decomposition, which exponentially increases with the number of frames. This fact can be observed from Table 3, where Avg-Feature is 6 times faster than MSM and 40 times faster than AHM with 4 frames and even much faster with 16 frames. Due to the real-time requirement for surveillance systems, MSM and AHM are not applicable especially when hundreds of frames are used. The second experiment done for MOBIO database is investigating the effect of the number of frames on performance. The performance of the three methods with different number of frames selected is evaluated. Because MSM and AHM are very slow especially with more frames selected, we only test them with 4 and 8 frames. With 16 frames or more, it will take at least a month for such a big dataset as MOBIO. For Avg-Feature method, we run the test for up to all frames selected from each video. Figures 7 and 6 illustrate the effect of using multiple frames as a biometric ROC of FAR versus FRR for the three methods. As can be seen from Figures 7(a) and 7(b), the performance for MSM and AHM with 8 frames is slightly better than that of 4 frames. The same trend can be observed from Figure 6 for




(a) ROC for 4 frames selected

(b) ROC for 8 frames selected

Figure 5: ROC of FAR versus FRR for MOBIO Dev Set with 4 and 8 frames selected.

Table 4: Half total error rate results on MOBIO dataset obtained from [66].

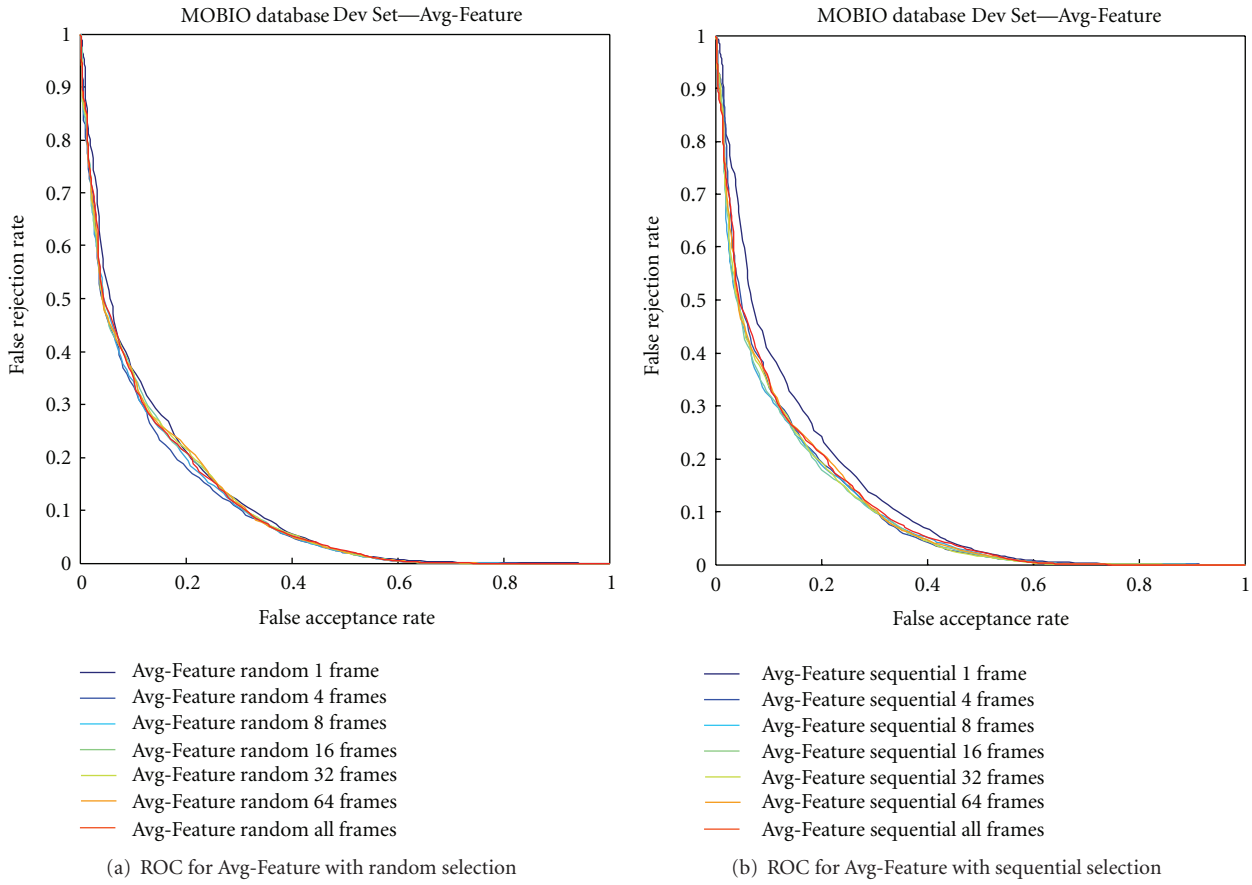
| Method | Male | Female | Average |
|---|---|---|---|
| IDIAP | 25.45 | 24.39 | 24.92 |
| ITI | 16.92 | 17.85 | 17.38 |
| Averaging MRH | 25.43 | 20.83 | 23.13 |
| TEC | 31.36 | 29.08 | 30.22 |
| UNIS | 9.75 | 12.07 | 10.91 |
| VISIDON | 10.30 | 14.95 | 12.62 |
| UON | 29.80 | 23.89 | 26.85 |
| NTU | 20.50 | 27.26 | 23.88 |
| UPV | 21.86 | 23.84 | 22.85 |

Avg-Feature, in which a small improvement in recognition is shown with the use of multiple frames compared to just 1 frame. But it does not imply that the more frames used the better the performance. The best performance for Avg-Feature is achieved with 8 frames selected for random selection and 16 frames selected for sequential selection.

To compare with other state-of-the-art techniques on MOBIO dataset, we also report the results on MOBIO test set in Table 4, which are obtained from [66]. Our proposed averaging MRH method performs the fourth and fifth for female and male test set separately. However, those methods that perform better than the proposed method use more reliable proprietary software for face detection. Compared with those methods using OpenCV face detector, MRH Averaging method performs better than them.

## 6. Conclusion and Future Work

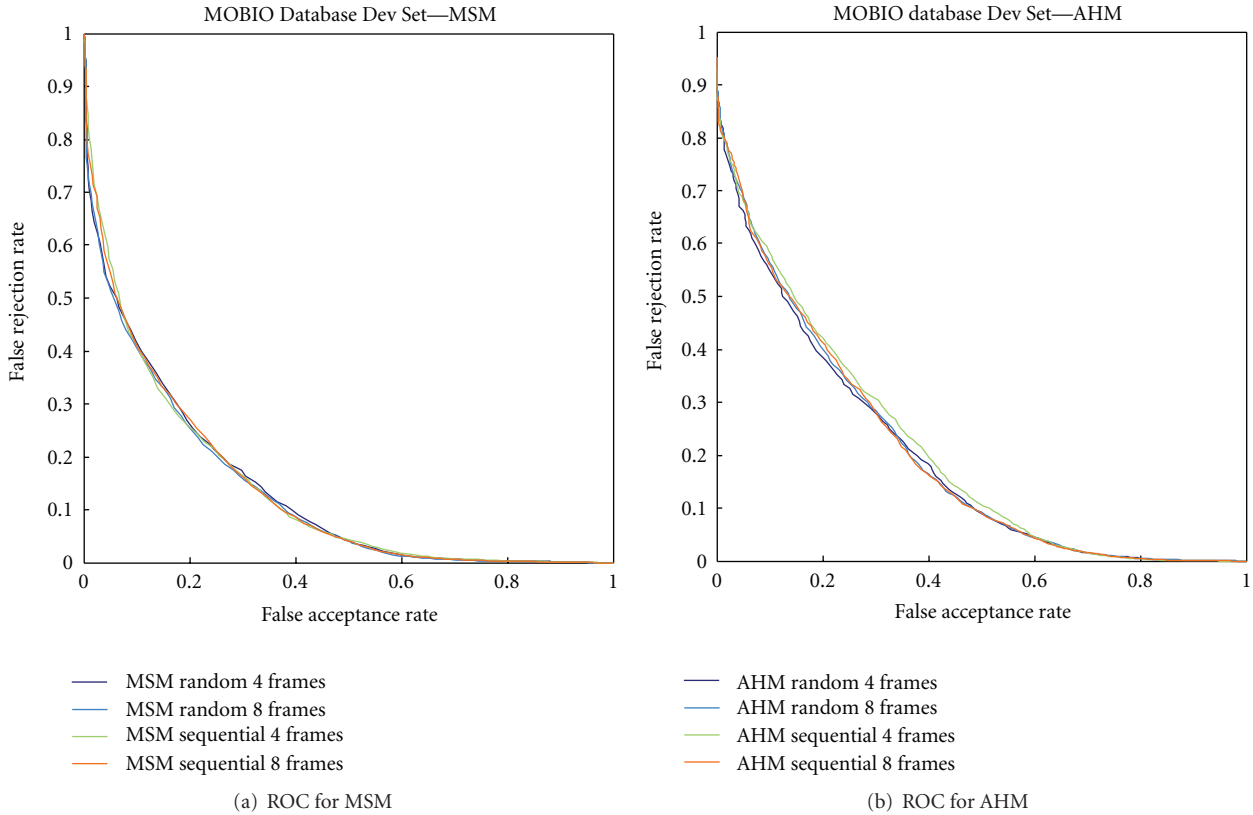
In this paper, we reviewed state-of-the-art face recognition techniques for still images and video sequences. Most of these existing approaches need well-aligned face images and only perform either still image face recognition or video-to-video match. They are not suitable for face recognition under surveillance scenarios because of the following reasons: limitation in the number (around ten) of face images extracted from each video due to the large variation in pose and lighting change; no guarantee of the face image alignment resulted from the poor video quality, constraints in the resource for calculation influenced by the real time processing. We then proposed a local facial feature-based framework for still image and video-based face recognition under surveillance conditions. This framework is generic to be capable of still-to-still, still-to-video and video-to-video matching in real-time. Evaluation of this approach is done for still image and video based face recognition on LFW image dataset and MOBIO video dataset. Experimental




(a) ROC for Avg-Feature with random selection

(b) ROC for Avg-Feature with sequential selection

Figure 6: ROC of FAR versus FRR for MOBIO Dev set for Avg-Feature.




(a) ROC for MSM

(b) ROC for AHM

Figure 7: ROC of FAR versus FRR for MOBIO Dev Set for MSM and AHM.

results show that MRH feature is more discriminative for face recognition with illumination, pose, and expression variations and is less sensitive to alignment errors. Empirical evaluation on video-based recognition with 8 methods for operation on feature and operation on distance shows that operation on features generally performs better. The best performance achieved is by Avg-Feature compared to other recent advanced methods such as MSM, MMD, and AHM, when the number of images per set is small (less than 10). MSM, MMD and AHM attempt to overfit to small number of samples, though they might outperform Avg-Feature with hundreds of images available per set. But the speed of the former is much slower than the latter. Thus, for face recognition under surveillance scenario, Avg-Feature is more suitable, subjected to the constraints in the number of images and real-time processing. Though experiments show that MRH feature is more reliable then other local features, such as LBP, GJD, and SIFT, recent research discovers some more robust features, for example, Learning-based Descriptors (LE) [54]. It is worth investigating the averaging effect on these features.

Besides technical challenges, data collection is one of the main issues for research on surveillance systems. Privacy laws or policies may prevent surveillance footage being used for research even if the video is already being used for security monitoring. Careful consultation and negotiation should be carried out before any real-life trials of intelligent surveillance systems.

## Acknowledgments

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] M. McCahill and C. Norris, *Urbaneye: CCTV in London*, Centre for Criminology and Criminal Justice, University of Hull, UK, 2002.

[2] G. Francisco, S. Roberts, K. Hanna, and J. Heubusch, "Critical infrastructure security confidence through automated thermal imaging," in *Infrared Technology and Applications XXXII*, vol. 6206 of *Proceedings of SPIE*, Kissimmee, Fla, USA, April 2006.

[3] L. M. Fuentes and S. A. Velastin, "From tracking to advanced surveillance," in *Proceedings of the International Conference on Image Processing (ICIP '03)*, pp. 121–124, September 2003.

[4] F. Ziliani, S. Velastin, F. Porikli et al., "Performance evaluation of event detection solutions: the CREDS experience," in *Proceedings of IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS '05)*, pp. 201–206, September 2005.

[5] Y. Gao and M. K. H. Leung, "Face recognition using line edge map," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 6, pp. 764–779, 2002.

[6] A. Yilmaz and M. Gökmen, "Eigenhill vs. eigenface and eigenedge," in *Proceedings of the International Conference on Pattern Recognition*, pp. 827–830, 2000.

[7] Y. Adinj, Y. Moses, and S. Ullman, "Face recognition: the problem of compensation for changes in illumination direction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 721–732, 1997.

[8] R. Basri and D. W. Jacobs, "Lambertian reflectance and linear subspaces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 2, pp. 218–233, 2003.

[9] A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman, "From few to many: illumination cone models for face recognition under variable lighting and pose," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 643–660, 2001.

[10] P. N. Belhumeur and D. J. Kriegman, "What is the set of images of an object under all possible illumination conditions?" *International Journal of Computer Vision*, vol. 28, no. 3, pp. 245–260, 1998.

[11] D. Beymer and T. Poggio, "Face recognition from one example view," in *Proceedings of the 5th International Conference on Computer Vision*, pp. 500–507, June 1995.

[12] M. J. Black, D. J. Fleet, and Y. Yacoob, "Robustly estimating changes in image appearance," *Computer Vision and Image Understanding*, vol. 78, no. 1, pp. 8–31, 2000.

[13] X. Liu, T. Chen, and B. V. K. V. Kumar, "Face authentication for multiple subjects using eigenflow," *Pattern Recognition*, vol. 36, no. 2, pp. 313–328, 2003.

[14] A. M. Martínez, "Recognizing imprecisely localized, partially occluded, and expression variant faces from a single sample per class," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 6, pp. 748–763, 2002.

[15] L. Wiskott, J. M. Fellous, N. Krüger, and C. D. Von Malsburg, "Face recognition by elastic bunch graph matching," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 775–779, 1997.

[16] D. Beymer, "Feature correspondence by interleaving shape and texture computations," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 921–928, June 1996.

[17] A. Pentland, B. Moghaddam, and T. Starner, "View-based and modular eigenspaces for face recognition," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 84–91, June 1994.

[18] P. Sankaran and V. Asari, "A multi-view approach on modular PCA for illumination and pose invariant face recognition," in *Proceedings of the 33rd Applied Imagery Pattern Recognition Workshop*, pp. 165–170, October 2004.

[19] W. Gao, S. Shan, X. Chai, and X. Fu, "Virtual face image generation for illumination and pose insensitive face recognition," in *Proceedings of IEEE International Conference on Accoustics, Speech, and Signal Processing*, pp. 776–779, April 2003.

[20] C. Sanderson, S. Bengio, and Y. Gao, "On transforming statistical models for non-frontal face verification," *Pattern Recognition*, vol. 39, no. 2, pp. 288–302, 2006.

[21] T. Shan, B. C. Lovell, and S. Chen, "Face recognition robust to head pose from one sample image," in *Proceedings of the 18th International Conference on Pattern Recognition (ICPR '06)*, pp. 515–518, August 2006.

[22] M. T. Harandi, M. Nili Ahmadabadi, and B. N. Araabi, "Optimal local basis: a reinforcement learning approach for face recognition," *International Journal of Computer Vision*, vol. 81, no. 2, pp. 191–204, 2009.

[23] S. Satoh, "Comparative evaluation of face sequence matching for content-based video access," in *Proceedings of IEEE Conference on Automatic Face and Gesture Recognition*, pp. 163–168, 2000.

[24] S. Zhou and R. Chellappa, "Probabilistic human recognition from video," in *Proceedings of the European Conference on Computer Vision*, pp. 681–697, Copenhagen, Denmark, 2002.

[25] Y. Zhang and A. M. Martinez, "A weighted probabilistic approach to face recognition form multiple images and video sequences," *Asian Security Review*, vol. 24, pp. 626–638, 2006.

[26] S. Zhou, V. Krueger, and R. Chellappa, "Face recognition from video: a condensation approach," in *Proceedings of IEEE Conference on Automatic Face and Gesture Recognition*, pp. 221–228, Washington, DC, USA, 2002.

[27] G. Shakhnarovich and B. Moghaddam, "Face recognition in subspaces," in *Handbook of Face Recognition*, Springer, New York, NY, USA, 2004.

[28] K. C. Lee, J. Ho, M. H. Yang, and D. Kriegman, "Video-based face recognition using probabilistic appearance manifolds," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 313–320, June 2003.

[29] V. Kruger and S. Zhou, "Exemplar-based face recognition from video," in *Proceedings of the European Conference on Computer Vision*, pp. 361–365, Copenhagen, Denmark, 2002.

[30] I. S. Topkaya and N. G. Bayazit, "Improving face recognition from videos with preprocessed representative faces," in *Proceedings of the 23rd International Symposium on Computer and Information Sciences (ISCIS '08)*, October 2008.

[31] J. Tangelder and B. Schouten, "Learning a sparse representation from multiple still images for online face recognition in an unconstrained environment," in *Proceedings of the International Conference on Pattern Recognition*, vol. 3, pp. 1087–1090, 2006.

[32] N. Vaswani and R. Chellappa, "Principal components null space analysis for image and video classification," *IEEE Transactions on Image Processing*, vol. 15, no. 7, pp. 1816–1830, 2006.

[33] S. Soatto, G. Doretto, and Y. Wu, "Dynamic textures," in *Proceedings of the International Conference on Computer Vision*, vol. 2, pp. 439–446, Vancouver, Canada, 2001.

[34] M. Kim, S. Kumar, V. Pavlovic, and H. Rowley, "Face tracking and recognition with visual constraints in real-world videos," in *Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition (CVPR '08)*, June 2008.

[35] P. Viola and M. J. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.

[36] A. J. O'Toole, D. A. Roark, and H. Abdi, "Recognizing moving faces: a psychological and neural synthesis," *Trends in Cognitive Sciences*, vol. 6, no. 6, pp. 261–266, 2002.

[37] A. R. Chowdhury, R. Chellappa, R. Krishnamurthy, and T. Vo, "3d face reconstruction from video using a generic model," in *Proceedings of the International Conference on Multimedia and Expo*, Lausanne, Switzerland, 2002.

[38] S. Baker and T. Kanade, "Limits on super-resolution and how to break them," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 9, pp. 1167–1183, 2002.

[39] X. Liu, T. Chen, and S. M. Thornton, "Eigenspace updating for non-stationary process and its application to face recognition," *Pattern Recognition*, vol. 36, no. 9, pp. 1945–1959, 2003.

[40] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, "Face recognition: a literature survey," *ACM Computing Surveys*, vol. 35, no. 4, pp. 399–458, 2003.

[41] J. Wang, C. Zhang, and H. Y. Shum, "Face image resolution versus face recognition performance based on two global methods," in *Proceedings of the Asian Conference on Computer Vision*, Jeju Island, Korea, 2004.

[42] Y. Rodriguez, F. Cardinaux, S. Bengio, and J. Mariéthoz, "Measuring the performance of face localization systems," *Image and Vision Computing*, vol. 24, no. 8, pp. 882–893, 2006.

[43] C. Sanderson and B. C. Lovell, "Multi-region probabilistic histograms for robust and scalable identity inference," in *Proceedings of the 3rd International Conference on Advances in Biometrics (ICB '09)*, pp. 199–208, June 2009.

[44] R. Gonzales and R. Woods, *Digital Image Processing*, Prentice Hall, Englewood Cliffs, NJ, USA, 3rd edition, 2007.

[45] C. Bishop, *Pattern Recognition and Machine Learning*, Springer, Berlin, Germany, 2006.

[46] C. Sanderson, T. Shang, and B. C. Lovell, "Towards pose-invariant 2D face classification for surveillance," in *Proceedings of the 3rd International Workshop on Analysis and Modeling of Faces and Gestures (AMFG '07)*, pp. 276–289, October 2007.

[47] T. Kadir and M. Brady, "Saliency, scale and image description," *International Journal of Computer Vision*, vol. 45, no. 2, pp. 83–105, 2001.

[48] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled Faces in the Wild: a database for studying face recognition in unconstrained environments," Tech. Rep. 07-49, 2007.

[49] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: recognition using class specific linear projection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711–720, 1997.

[50] E. Nowak and F. Jurie, "Learning visual similarity measures for comparing never seen objects," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '07)*, June 2007.

[51] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

[52] M. Lades, J. C. Vorbrueggen, J. Buhmann et al., "Distortion invariant object recognition in the dynamic link architecture," *IEEE Transactions on Computers*, vol. 42, no. 3, pp. 300–311, 1993.

[53] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[54] Z. Cao, Q. Yin, X. Tang, and J. Sun, "Face recognition with learning-based descriptor," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, San Francisco, Calif, USA, 2010.

[55] J. Ruiz-Del-Solar, R. Verschae, and M. Correa, "Recognition of faces in unconstrained environments: a comparative study," *EURASIP Journal on Advances in Signal Processing*, vol. 2009, Article ID 184617, 19 pages, 2009.

[56] S. Marcel, C. McCool, P. M. Ahonen et al., "Mobile biometry (mobio) face and speaker verification evaluation," in *Proceedings of the 20th International Conference on Pattern Recognition*, 2010.

[57] R. Jenkins and A. M. Burton, "100% Accuracy in automatic face recognition," *Science*, vol. 319, no. 5862, p. 435, 2008.

[58] S. Zhao, X. Zhang, and Y. Gao, "A comparative evaluation of average face on holistic and local face recognition approaches," in *Proceedings of the 19th International Conference on Pattern Recognition (ICPR '08)*, December 2008.

[59] R. Wang, S. Shan, X. Chen, and W. Gao, "Manifold-manifold distance with application to face recognition based on image set," in *Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition (CVPR '08)*, June 2008.

[60] O. Yamaguchi, K. Fukui, and K. Maeda, "Face recognition using temporal image sequence," in *Proceedings of the 3rd IEEE International Conference on Automatic Face and Gesture Recognition*, Nara, Japan, 1998.

[61] L. Wolf and A. Shashua, "Learning over sets using kernel principal angles," *Journal of Machine Learning Research*, vol. 4, no. 6, pp. 913–931, 2004.

[62] T. K. Kim, J. Kittler, and R. Cipolla, "Discriminative learning and recognition of image set classes using canonical correlations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 6, pp. 1005–1018, 2007.

[63] H. Cevikalp and B. Triggs, "Face recognition based on image sets," in *IEEE Conference on Computer Vision and Pattern Recognition*, San Francisco, Calif, USA, 2010.

[64] T. Ahonen, A. Hadid, and M. Pietikainen, "Face recognition with local binary patterns," in *Proceedings of the 8th European Conference on Computer Vision (ECCV '04)*, Prague, Czech Republic, 2004.

[65] G. R. Bradski, "Computer video face tracking for use in a perceptual user interface," *Intel Technology Journal Q2*, 1998.

[66] S. Marcel, C. McCool, P. Matejka et al., "On the results of the first mobile biometry (mobio) face and speaker verification evaluation," in *Proceedings of IEEE Conference on Pattern Recognition (ICPR '10)*, Istanbul, Turkey, 2010.