Electronic Letters on Computer Vision and Image Analysis 6(3):30-41, 2007

Intelligent CCTV for Mass Transport Security: Challenges and Opportunities for Video and Face Processing

Conrad Sanderson[*], Abbas Bigdeli[*], Ting Shan[*+],
Shaokang Chen[*+], Erik Berglund[*+], and Brian C. Lovell[*+]

[*] *NICTA, 300 Adelaide St, Brisbane, QLD 4000, Australia*
[+] *ITEE, The University of Queensland, QLD 4072, Australia*

Received 5 May 2007; Accepted 9 October 2007

Abstract

CCTV surveillance systems have long been promoted as being effective in improving public safety. However due to the amount of cameras installed, many sites have abandoned expensive human monitoring and only record video for forensic purposes. One of the sought-after capabilities of an automated surveillance system is "face in the crowd" recognition, in public spaces such as mass transit centres. Apart from accuracy and robustness to nuisance factors such as pose variations, in such surveillance situations the other important factors are scalability and fast performance. We evaluate recent approaches to the recognition of faces at large pose angles from a gallery of frontal images and propose novel adaptations as well as modifications. We compare and contrast the accuracy, robustness and speed of an Active Appearance Model (AAM) based method (where realistic frontal faces are synthesized from non-frontal probe faces) against bag-of-features methods. We show a novel approach where the performance of the AAM based technique is increased by side-stepping the image synthesis step, also resulting in a considerable speedup. Additionally, we adapt a histogram-based bag-of-features technique to face classification and contrast its properties to a previously proposed direct bag-of-features method. We further show that the two bag-of-features approaches can be considerably sped up, without a loss in classification accuracy, via an approximation of the exponential function. Experiments on the FERET and PIE databases suggest that the bag-of-features techniques generally attain better performance, with significantly lower computational loads. The histogram-based bag-of-features technique is capable of achieving an average recognition accuracy of 89% for pose angles of around 25 degrees. Finally, we provide a discussion on implementation as well as legal challenges surrounding research on automated surveillance.

Key Words: surveillance, video analysis, face classification, pose, bag of words, AAM, GMM.

1 Introduction

In response to global terrorism, usage and interest in Closed-Circuit Television (CCTV) for surveillance and protection of public spaces (such as mass transit facilities) is growing at a considerable rate. A similar escalation of the installed CCTV base occurred in London late last century in response to the continual bombings linked to the conflict in Northern Ireland. Based on the number of CCTV cameras on Putney High Street, it is

Correspondence to: <lovell@itee.uq.edu.au>

Recommended for acceptance by <name>
ELCVIA ISSN:1577-5097
Published by Computer Vision Center / Universitat Autònoma de Barcelona, Barcelona, Spain

"guesstimated" [15] that there are around 500,000 CCTV cameras in the London area and 4,000,000 cameras in the UK. This suggests that in the UK there is approximately one camera for every 14 people. However, whilst it is relatively easy, albeit expensive, to install increasing numbers of cameras, it is quite another issue to adequately monitor the video feeds with security guards. Hence, the trend has been to record the CCTV feeds without monitoring and to use the video merely for a forensic, or reactive, response to crime and terrorism, often detected by other means.

In minor crimes such as assault and robbery, surveillance video is very effective in helping to find and successfully prosecute perpetrators. Thus one would expect that surveillance video would act as a deterrent to crime. Recently the immense cost of successful terrorist attacks on soft targets such as mass transport systems has indicated that forensic analysis of video after the event is simply not an adequate response. Indeed, in the case of suicide bombings there is simply no possibility of prosecution after the event and thus no deterrent effect. A pressing need is emerging to monitor all surveillance cameras in an attempt to detect events and persons-of-interest.

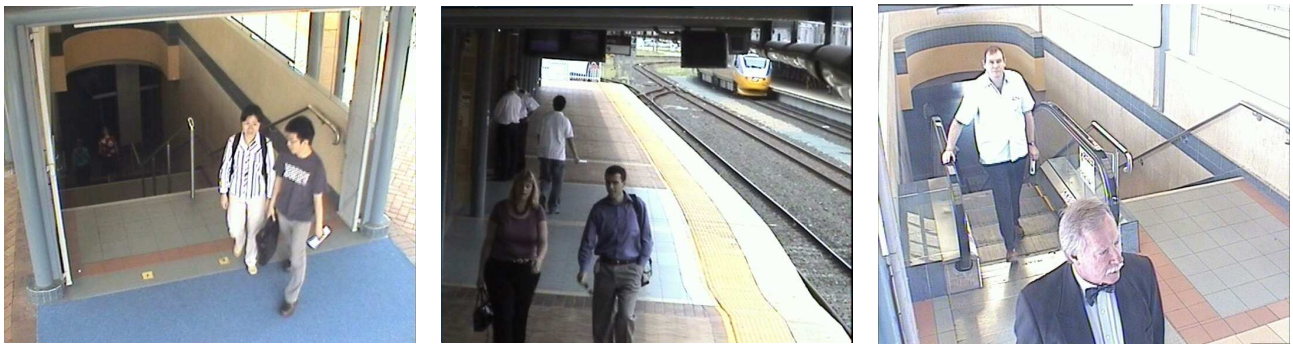
One important issue is the fact that human monitoring requires a large number of people, resulting in high ongoing costs. Furthermore, such a personnel intensive system has questionable reliability due to the attention span of humans decreasing rapidly when performing such tedious tasks. A solution may be found in advanced surveillance systems employing computer monitoring of all video feeds, delivering the alerts to human responders for triage. Indeed such systems may assist in maintaining the high level of vigilance required over many years to detect the rare events associated with terrorism — a well-designed computer system is never caught "off guard". Because of this, there has been a significant rush 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 [9, 10, 27].

Amongst the various biometric techniques for person identification, recognition via gait and faces appears to be the most useful in the context of CCTV. Our starting point is the robust identification of persons of interest, which is motivated by problems encountered in our initial real-world trials of face recognition technologies in public railway stations using existing cameras.

While automatic face recognition of cooperative subjects has achieved good results in controlled applications such as passport control, CCTV conditions are considerably more challenging. Nuisance factors such as varying illumination, expression, and pose can greatly affect recognition performance. According to Phillips *et al*. head pose is believed to be the hardest factor to model [17]. In mass transport systems, surveillance cameras are often mounted in the ceiling in places such as railway platforms and passenger trains. Since the subjects are generally not posing for the camera, it is rare to obtain a true frontal face image. As it is infeasible to consider remounting all the cameras (in our case more than 6000) to improve face recognition performance, any practical system must have effective pose compensation or be specifically designed to handle pose variations. Examples of real life CCTV conditions are shown in Figure 1.

A further complication is that we generally only have one frontal gallery image of each person-of-interest (e.g. a passport photograph or a mugshot). In addition to robustness and accuracy, scalability and fast performance are also of prime importance for surveillance. A face recognition system should be able to handle large volumes of people (e.g. peak hour at a railway station), possibly processing hundreds of video streams. While it is possible to setup elaborate parallel computation machines, there are always cost considerations limiting the number of CPUs available for processing. In this context, a face recognition algorithm should be able to run in real-time or better, which necessarily limits complexity.

Previous approaches to addressing pose variation include the synthesis of new images at previously unseen views [1, 22], direct synthesis of face model parameters [20] and local feature based representations [3, 14, 26]. We note in passing that while true 3D based approaches in theory allow face matching at various poses, current 3D sensing hardware has too many limitations [2], including cost and range. Moreover unlike 2D recognition, 3D technology cannot be retrofitted to existing surveillance systems.



(a) (b) (c)

Figure 1: Several frames from CCTV cameras located at a railway station in Brisbane (Australia), demonstrating some of the variabilities present in real-life conditions: (a) varying face pose, (b) illumination from one side, (c) varying size and pose.

In [22], Active Appearance Models (AAMs) were used to model each face, detecting the pose through a correlation model. A frontal image could then be synthesized directly from a single non-frontal image without the need to explicitly generate a 3D head model. While the AAM-based face synthesis allowed considerable improvements in recognition accuracy, the synthesized faces have residual artefacts which may affect recognition performance.

In [20], a "bag of features" approach was shown to perform well in the presence of pose variations. It is based on dividing the face into overlapping uniform-sized blocks, analysing each block with the Discrete Cosine Transform (DCT) and modelling the resultant set of features via a Gaussian Mixture Model (GMM). The robustness to pose change was attributed to an effective insensitivity to the topology of the face. We shall refer to this method as the *direct bag-of-features*.

Inspired by text classification techniques from the fields of natural language processing and information retrieval, alternative forms of the "bag of features" approach are used for image categorisation in [7, 24, 16]. Rather than directly calculating the likelihood as in [20], histograms of occurrences of "visual words" (also known as "keypoints") are first built, followed by histogram comparison. We shall refer to this approach as the *histogram-based bag-of-features*.

The research reported in this paper has four main aims: **(i)** To evaluate the effectiveness of a novel modification of the AAM-based method, where we explicitly remove the effect of pose from the face model creating pose-robust features. The modification allows the use of the model's parameters directly for classification, thereby skipping the computationally intensive and artefact producing image synthesis step. **(ii)** To adapt the histogram-based bag-of-features approach to face classification and contrast its properties to the direct bag-of-features method. **(iii)** To evaluate the extent of speedup possible in the both bag-of-features approaches via an approximation of the exponential function, and whether such approximation affects recognition accuracy. **(iv)** To compare the performance, robustness and speed of AAM based and bag-of-features based methods in the context of face classification under pose variations.

The balance of this paper is structured as follows. In Section 2 we overview the two bag-of-features methods. In Section 3 we overview the AAM-based synthesis technique and present the modified form. Section 4 is devoted to an evaluation of the techniques on the FERET and PIE datasets. A discussion of the results, as well as implementation and legal issues surrounding research on automated surveillance, is given in Section 5.

2 Bag-of-Features Approaches

In this section we describe two local feature based approaches, with both approaches sharing a block based feature extraction method summarised in Section 2.1. Both methods use Gaussian Mixture Models (GMMs) to model distributions of features, but they differ in how the GMMs are applied. In the first approach (*direct bag-of-features*, Section 2.2) the likelihood of a given face belonging to a specific person is calculated directly using that person's model. In the second approach (*histogram-based bag-of-features*, Section 2.3), a generic model (not specific to any person), representing "face words", is used to build histograms which are then compared for recognition purposes. In Section 2.4 we describe how both techniques can be sped up.

2.1 Feature Extraction and Illumination Normalisation

The face is described as a set of feature vectors, $X = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\}$, which are obtained by dividing the face into small, uniformly sized, overlapping blocks and decomposing each block* via the 2D DCT [11]. Typically the first 15 to 21 DCT coefficients are retained (as they contain the vast majority of discriminatory information), except for the 0-th coefficient which is the most affected by illumination changes [3].

To achieve enhanced robustness to illumination variations, we have incorporated additional processing prior to 2D DCT decomposition. Assuming the illumination model for each pixel to be $\widehat{p}_{(x,y)} = b + c \cdot p_{(x,y)}$, where $p_{(x,y)}$ is the "uncorrupted" pixel at location (x, y), b is a bias and c a multiplier (indicating the contrast), removing the 0-th DCT coefficient only corrects for the bias. To achieve robustness to contrast variations, the set of pixels within each block is normalised to have zero mean and unit variance.

2.2 Bag-of-Features with Direct Likelihood Evaluation

By assuming the vectors are independent and identically distributed (i.i.d.), the likelihood of X belonging to person i is found with:

$$P(X|\lambda^{[i]}) = \prod_{n=1}^{N} P(\mathbf{x}_n|\lambda^{[i]}) = \prod_{n=1}^{N} \sum_{g=1}^{G} w_g^{[i]} \mathcal{N}\left(\mathbf{x}_n|\mu_g^{[i]}, \mathbf{\Sigma}_g^{[i]}\right) \tag{1}$$

where $\mathcal{N}(\mathbf{x}|\mu, \mathbf{\Sigma}) = (2\pi)^{-\frac{d}{2}} |\mathbf{\Sigma}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\mu)^T \mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)\right\}$ is a multi-variate Gaussian function [8], while $\lambda^{[i]} = \{w_g^{[i]}, \mu_g^{[i]}, \mathbf{\Sigma}_g^{[i]}\}_{g=1}^{G}$ is the set of parameters for person i. The convex combination of Gaussians, with mixing coefficients w_g, is typically referred to as a Gaussian Mixture Model (GMM). Its parameters are optimised via the Expectation Maximisation algorithm [8].

Due to the vectors being treated as i.i.d., information about the topology of the face is in effect lost. While at first this may seem counter-productive, the loss of topology in conjunction with overlapping blocks provides a useful characteristic: the precise location of face parts is no longer required. Previous research has suggested that the method is effective for face classification while being robust to imperfect face detection as well as a certain amount of in-plane and out-of-plane rotations [3, 19, 20].

The robustness to pose variations can be attributed to the explicit allowance for movement of face areas, when comparing face images of a particular person at various poses. Furthermore, significant changes of a particular face component (e.g. the nose) due to pose variations affect only the subset of face areas that cover this particular component.

*While in this work we used the 2D DCT for describing each block (or patch), it is possible to use other descriptors, for example Gabor wavelets [13].

2.3 Bag-of-Features with Histogram Matching

The technique presented in this section is an adaption of the "visual words" method used in image categorisation [7, 24, 16]. First, a training set of faces is used to build a generic model (not specific to any person). This generic model represents a dictionary of "face words" — the mean of each Gaussian can be thought of as a particular "face word". Once a set of feature vectors for a given face is obtained, a probabilistic histogram of the occurrences of the "face words" is built:

$$\vec{h}_X = \frac{1}{N}\left[\sum_{i=1}^{N}\frac{w_1 p_1\left(\vec{x}_i\right)}{\sum_{g=1}^{G} w_g p_g\left(\vec{x}_i\right)}, \sum_{i=1}^{N}\frac{w_2 p_2\left(\vec{x}_i\right)}{\sum_{g=1}^{G} w_g p_g\left(\vec{x}_i\right)}, \cdots, \sum_{i=1}^{N}\frac{w_G p_G\left(\vec{x}_i\right)}{\sum_{g=1}^{G} w_g p_g\left(\vec{x}_i\right)}\right]$$

where w_g is the weight for Gaussian g and $p_g\left(\vec{x}\right)$ is the probability of vector $\vec{x}$ according to Gaussian g.

Comparison of two faces is then accomplished by comparing their corresponding histograms. This can be done by the so-called χ^2 distance metric [25], or the simpler approach of summation of absolute differences [12]:

$$d\left(\vec{h}_A, \vec{h}_B\right) = \sum\nolimits_{g=1}^{G}\left|\vec{h}_A^{[g]} - \vec{h}_B^{[g]}\right| \tag{2}$$

where $\vec{h}_A^{[g]}$ is the g-th element of $\vec{h}_A$. As preliminary experiments suggested that there was little difference in performance between the two metrics, we've elected to use the latter one.

Note that like in the direct method presented in the previous section, information about the topology of the face is lost. However, the direct method requires that the set of features from a given probe face is processed using all models of the persons in the gallery. As such, the amount of processing can quickly become prohibitive as the gallery grows[†]. In contrast, the histogram-based approach requires the set of features to be processed using only one model, potentially providing savings in terms of storage and computational effort.

Another advantage of the histogram-based approach is that the face similarity measurement, via Eqn. (2), is symmetric. This is not the case for the direct approach, as the representation of probe and gallery faces differs — a probe face is represented by a set of features, while a gallery face is represented by a model of features (the model, in this case, can be thought of as a compact approximation of the set of features from the gallery face).

2.4 Speedup via Approximation

In practice the time taken by the 2D DCT feature extraction stage is negligible and hence the bulk of processing in the above two approaches is heavily concentrated in the evaluation of the exponential function. As such, a considerable speedup can be achieved through the use of a fast approximation of this function [21]. A brief overview follows: rather than using a lookup table, the approximation is accomplished by exploiting the structure and encoding of a standard (IEEE-754) floating-point representation. The given argument is transformed and injected as an integer into the first 32 bits of the 64 bit representation. Reading the resulting floating point number provides the approximation. Experiments in Section 4 indicate that the approximation does not affect recognition accuracy.

[†]For example, assuming each model has 32 Gaussians, going through a gallery of 1000 people would require evaluating 32000 Gaussians. Assuming 784 vectors are extracted from each face, the number of exp() evaluations is around 25 million.

3 Active Appearance Models

In this section we describe face modelling based on deformable models popularised by Cootes et al., namely Active Shape Models (ASMs) [5] and Active Appearance Models (AAMs) [4]. We first provide a brief description of the two models, followed by pose estimation via a correlation model and finally frontal view synthesis. We also show that the synthesis step can be omitted by directly removing the effect of the pose from the model of the face, resulting in (theoretically) pose independent features.

3.1 Face Modelling

Let us describe a face by a set of N landmark points, where the location of each point is tuple (x, y). A face can hence be represented by a $2N$ dimensional vector:

$$\mathbf{f} = [\, x_1, x_2, \cdots, x_N,\ y_1, y_2, \cdots, y_N \,]^T . \tag{3}$$

In ASM, a face shape is represented by:

$$\mathbf{f} = \overline{\mathbf{f}} + \mathbf{P}_s \mathbf{b}_s \tag{4}$$

where $\overline{\mathbf{f}}$ is the mean face vector, $\mathbf{P}_s$ is a matrix containing the k eigenvectors with largest eigenvalues (of a training dataset), and $\mathbf{b}_s$ is a weight vector. In a similar manner, the texture variations can be represented by:

$$\mathbf{g} = \overline{\mathbf{g}} + \mathbf{P}_g \mathbf{b}_g \tag{5}$$

where $\overline{\mathbf{g}}$ is the mean appearance vector, $\mathbf{P}_g$ is a matrix describing the texture variations learned from training sets, and $\mathbf{b_g}$ is the texture weighting vector.

The shape and appearance parameters $\mathbf{b}_s$ and $\mathbf{b}_g$ can be used to describe the shape and appearance of any face. As there are correlations between the shape and appearance of the same person, let us first represent both aspects as:

$$\mathbf{b} = \begin{bmatrix} \mathbf{W}_s \mathbf{b}_s \\ \mathbf{b}_g \end{bmatrix} = \begin{bmatrix} \mathbf{W}_s \mathbf{P}_s^T (\mathbf{f} - \overline{\mathbf{f}}) \\ \mathbf{P}_g^T (\mathbf{g} - \overline{\mathbf{g}}) \end{bmatrix} \tag{6}$$

where $\mathbf{W}_s$ is a diagonal matrix which represents the change between shape and texture. Through Principal Component Analysis (PCA) [8] we can represent $\mathbf{b}$ as:

$$\mathbf{b} = \mathbf{P}_c \mathbf{c} \tag{7}$$

where $\mathbf{P}_c$ are eigenvectors, $\mathbf{c}$ is a vector of appearance parameters controlling both shape and texture of the model, and $\mathbf{b}$ can be shown to have zero mean. Shape $\mathbf{f}$ and texture $\mathbf{g}$ can then be represented by:

$$\mathbf{f} = \overline{\mathbf{f}} + \mathbf{Q}_s \mathbf{c} \tag{8}$$

$$\mathbf{g} = \overline{\mathbf{g}} + \mathbf{Q}_g \mathbf{c} \tag{9}$$

where

$$\mathbf{Q}_s = \mathbf{P}_s \mathbf{W}_s^{-1} \mathbf{P}_{cs} \tag{10}$$

$$\mathbf{Q}_g = \mathbf{P}_g \mathbf{P}_{cg} \tag{11}$$

In the above, $\mathbf{Q}_s$ and $\mathbf{Q}_g$ are matrices describing the shape and texture variations, while $\mathbf{P}_{cs}$ and $\mathbf{P}_{cg}$ are shape and texture components of $\mathbf{P}_c$ respectively, i.e.:

$$\mathbf{P}_c = \begin{bmatrix} \mathbf{P}_{cs} \\ \mathbf{P}_{cg} \end{bmatrix} \tag{12}$$

The process of "interpretation" of faces is hence comprised of finding a set of model parameters which contain information about the shape, orientation, scale, position, and texture.

3.2 Pose Estimation

Following [6], let us assume that the model parameter $\mathbf{c}$ is approximately related to the viewing angle, θ, by a correlation model:

$$\mathbf{c} \approx \mathbf{c}_0 + \mathbf{c}_c \cos(\theta) + \mathbf{c}_s \sin(\theta) \tag{13}$$

where $\mathbf{c}_0$, $\mathbf{c}_c$ and $\mathbf{c}_s$ are vectors which are learned from the training data. (Here we consider only head turning. Head nodding can be dealt with in a similar way).

For each face from a training set Ω, indicated by superscript $[i]$ with associated pose $\theta^{[i]}$, we perform an AAM search to find the best fitting model parameters $\mathbf{c}^{[i]}$. The parameters $\mathbf{c}_0$, $\mathbf{c}_c$ and $\mathbf{c}_s$ can be learned via regression from $\left(\mathbf{c}^{[i]}\right)_{i\in 1,\cdots,|\Omega|}$ and $\left(\left[1, \cos(\theta^{[i]}), \sin(\theta^{[i]})\right]\right)_{i\in 1,\cdots,|\Omega|}$, where $|\Omega|$ indicates the cardinality of Ω.

Given a new face image with parameters $\mathbf{c}^{[new]}$, we can estimate its orientation as follows. We first rearrange $\mathbf{c}^{[new]} = \mathbf{c}_0 + \mathbf{c}_c \cos(\theta^{[new]}) + \mathbf{c}_s \sin(\theta^{[new]})$ to:

$$\mathbf{c}^{[new]} - \mathbf{c}_0 = [\ \mathbf{c}_c\ \mathbf{c}_s\ ] \left[\ \cos(\theta^{[new]})\ \sin(\theta^{[new]})\ \right]^T. \tag{14}$$

Let $\mathbf{R}_c^{-1}$ be the left pseudo-inverse of the matrix $[\ \mathbf{c}_c\ \mathbf{c}_s\ ]$. Eqn. (14) can then be rewritten as:

$$\mathbf{R}_c^{-1} \left(\mathbf{c}^{[new]} - \mathbf{c}_0\right) = \left[\ \cos(\theta^{[new]})\ \sin(\theta^{[new]})\ \right]^T. \tag{15}$$

Let $[\ x_\alpha\ y_\alpha\ ] = \mathbf{R}_c^{-1} \left(\mathbf{c}^{[new]} - \mathbf{c}_0\right)$. Then the best estimate of the orientation is $\theta^{[new]} = \tan^{-1}(y_\alpha / x_\alpha)$. Note that the estimation of $\theta^{[new]}$ may not be accurate due to land mark annotation errors or regression learning errors.

3.3 Frontal View Synthesis

After the estimation of $\theta^{[new]}$, we can use the model to synthesize frontal face views. Let $\mathbf{c}_{res}$ be the residual vector which is not explained by the correlation model:

$$\mathbf{c}_{res} = \mathbf{c}^{[new]} - \left(\mathbf{c}_0 + \mathbf{c}_c \cos(\theta^{[new]}) + \mathbf{c}_s \sin(\theta^{[new]})\right) \tag{16}$$

To reconstruct at an alternate angle, $\theta^{[alt]}$, we can add the residual vector to the mean face for that angle:

$$\mathbf{c}^{[alt]} = \mathbf{c}_{res} + \left(\mathbf{c}_0 + \mathbf{c}_c \cos(\theta^{[alt]}) + \mathbf{c}_s \sin(\theta^{[alt]})\right) \tag{17}$$

To synthesize the frontal view face, $\theta^{[alt]}$ is set to zero. Eqn. (17) hence simplifies to:

$$\mathbf{c}^{[alt]} = \mathbf{c}_{res} + \mathbf{c}_0 + \mathbf{c}_c \tag{18}$$

Based on Eqns. (8) and (9), the shape and texture for the frontal view can then be calculated by:

$$\mathbf{f}^{[alt]} = \overline{\mathbf{f}} + \mathbf{Q}_s \mathbf{c}^{[alt]} \tag{19}$$

$$\mathbf{g}^{[alt]} = \overline{\mathbf{g}} + \mathbf{Q}_g \mathbf{c}^{[alt]} \tag{20}$$

Examples of synthesized faces are shown in Fig. 2. Each synthesized face can then be processed via the standard Principal Component Analysis (PCA) technique to produce features which are used for classification [22].

3.4 Direct Pose-Robust Features

The bracketed term in Eqn. (16) can be interpreted as the mean face for angle $\theta^{[new]}$. The difference between $\mathbf{c}^{[new]}$ (which represents the given face at the estimated angle $\theta^{[new]}$) and the bracketed term can hence be interpreted as removing the effect of the angle, resulting in a (theoretically) pose independent representation. As such, $\mathbf{c}_{res}$ can be used directly for classification, providing considerable computational savings — the process of face synthesis and PCA feature extraction is omitted. Because of this, we're avoiding the introduction of imaging artefacts (due to synthesis) and information loss caused by PCA-based feature extraction. As such, the pose-robust features should represent the faces more accurately, leading to better discrimination performance. We shall refer to this approach as the *pose-robust features* method.

4 Evaluation

We are currently in the process of creating a suitable dataset for face classification in CCTV conditions (part of a separately funded project). As such, in these experiments we instead used subsets of the PIE dataset [23] (using faces at -22.5^o, 0^o and $+22.5^o$) as well as the FERET dataset [18] (using faces at -25^o, -15^o, 0^o, $+15^o$ and $+25^o$).

To train the AAM based approach, we first pooled face images from 40 FERET individuals at -15^o, 0^o, $+15^o$. Each face image was labelled with 58 points around the salient features (the eyes, mouth, nose, eyebrows and chin). The resulting model was used to automatically find the facial features (via an AAM search) for the remainder of the FERET subset. A new dataset was formed, consisting of 305 images from 61 persons with successful AAM search results. This dataset was used to train the correlation model and evaluate the performances of all presented algorithms. In a similar manner, a new dataset was formed from the PIE subset, consisting of images for 53 persons.

For the synthesis based approach, the last stage (PCA based feature extraction from synthesized images) produced 36 dimensional vectors. The PCA subsystem was trained as per [22]. The pose-robust features approach produced 43 dimensional vectors for each face. For both of the AAM-based techniques, Mahalanobis distance was used for classification [8].

For the bag-of-features approaches, in a similar manner to [20], we used face images with a size of 64×64 pixels, blocks with a size of 8×8 pixels and an overlap of 6 pixels. This resulted in 784 feature vectors per face. The number of retained DCT coefficients was set to 15 (resulting in 14 dimensional feature vectors, as the 0-th coefficient was discarded). The faces were normalised in size so that the distance between the eyes was 32 pixels and the eyes were in approximately the same positions in all images.

Figure 2: Top row: frontal view and its AAM-based synthesized representation. Bottom row: non-frontal view as well as its AAM-based synthesized representation at its original angle and $\theta^{[alt]} = 0$ (i.e. synthesized frontal view).

Method	Pose			
	-25^o	-15^o	$+15^o$	$+25^o$
PCA	23.0	54.0	49.0	36.0
Synthesis + PCA	50.0	71.0	67.4	42.0
pose-robust features	**85.6**	88.2	88.1	66.8
Direct bag-of-features	83.6	93.4	**100.0**	72.1
Histogram bag-of-features	83.6	**100.0**	96.7	**73.7**

Table 1: Recognition performance on the FERET pose subset.

Method	Pose	
	-22.5^o	$+22.5^o$
PCA	13.0	8.0
Synthesis + PCA	60.0	56.0
pose-robust features	83.3	80.6
Direct bag-of-features	**100.0**	90.6
Histogram bag-of-features	**100.0**	**100.0**

Table 2: Recognition performance on PIE.

For the direct bag-of-features approach, the number of Gaussians per model was set to 32. Preliminary experiments indicated that accuracy for faces at around 25^o peaked at 32 Gaussians, while using more than 32 Gaussians provided little gain in accuracy at the expense of longer processing times.

For the histogram-based bag-of-features method, the number of Gaussians for the generic model was set to 1024, following the same reasoning as above. The generic model (representing "face words") was trained on FERET *ba* data (frontal faces), excluding the 61 persons described earlier.

Tables 1 and 2 show the recognition rates on the FERET and PIE datasets, respectively. The AAM-derived pose-robust features approach obtains performance which is considerably better than the circuitous approach based on image synthesis. However, the two bag-of-features methods generally obtain better performance on both FERET and PIE, with the histogram-based approach obtaining the best overall performance. Averaging across the high pose angles ($\pm 25^o$ on FERET and $\pm 22.5^o$ on PIE), the histogram-based method achieves an average accuracy of 89%.

Table 3 shows the time taken to classify one probe face by the presented techniques (except for PCA). The experiments were performed on a Pentium-M machine running at 1.5 GHz. All methods were implemented in C++. The time taken is divided into two components: (1) one-off cost per probe face, and (2) comparison of one probe face with one gallery face.

The one-off cost is the time required to convert a given face into a format which will be used for matching. For the synthesis approach this involves an AAM search, image synthesis and PCA based feature extraction. For the pose-robust features method, in contrast, this effectively involves only an AAM search. For the bag-of-features approaches, the one-off cost is the 2D DCT feature extraction, with the histogram-based approach additionally requiring the generation of the "face words" histogram.

The second component, for the case of the direct bag-of-features method, involves calculating the likelihood using Eqn. (1), while for the histogram-based approach this involves just the sum of absolute differences between two histograms (Eqn. (2)). For the two AAM-based methods, the second component is the time taken to evaluate the Mahalanobis distance.

As expected, the pose-robust features approach has a speed advantage over the synthesis based approach, being about 50% faster. However, both of the bag-of-features methods are many times faster, in terms of the first component — the histogram-based approach is about 7 times faster than the pose-robust features method. While the one-off cost for the direct bag-of-features approach is much lower than for the histogram-based method, the time required for the second component (comparison of faces after conversion) is considerably higher, and might be a limiting factor when dealing with a large set of gallery faces (i.e. a scalability issue).

Method	**Approximate time taken (sec)**	
	One-off cost per probe face	Comparison of one probe face with one gallery face
Synthesis + PCA	1.493	< 0.001
pose-robust features	0.978	< 0.001
Direct bag-of-features	0.006	0.006
Histogram bag-of-features	0.141	< 0.001

Table 3: Average time taken for two stages of processing: (1) conversion of a probe face from image to format used for matching (one-off cost per probe face), (2) comparison of one probe face with one gallery face, after conversion.

When using the fast approximation of the exponential function, the time required by the histogram-based method (in the first component) is reduced by approximately 30% to 0.096, with no loss in recognition accuracy. This makes it over 10 times faster than the pose-robust features method and over 15 times faster than the synthesis based technique. In a similar vein, the time taken by the second component of the direct bag-of-features approach is also reduced by approximately 30%, with no loss in recognition accuracy.

5 Discussion

With an aim towards improving intelligent surveillance systems, in this paper we have made several contributions. We proposed a novel approach to Active Appearance Model based face classification, where pose-robust features are obtained without the computationally expensive image synthesis step. Furthermore, we've adapted a *histogram-based bag-of-features* technique (previously employed in image categorisation) to face classification, and contrasted its properties to a previously proposed *direct bag-of-features* method. We have also shown that the two bag-of-features approaches, both based on Gaussian Mixture Models, can be considerably sped up without a loss in classification accuracy via an approximation of the exponential function.

In the context of pose mismatches between probe and gallery faces, experiments on the FERET and PIE databases suggest that while there is merit in the AAM based methods, the bag-of-features techniques generally attain better performance, with the histogram-based method achieving an average recognition rate of 89% for pose angles of around 25 degrees. Furthermore, the bag-of-features approaches are considerably faster, with the histogram-based method (using the fast exponential function) being over 10 times quicker than the pose-robust features method.

We note that apart from pose variations, imperfect face localisation [19] is also an important issue in a real life surveillance system. Imperfect localisations result in translations as well as scale changes, which adversely affect recognition performance. To that end, we are currently extending the histogram-based bag-of-features approach to also deal with scale variations.

As mentioned in the introduction, the research presented here is motivated by application to real-life conditions. One of our "test-beds" intended for field trials is a railway station in Brisbane (Australia), which provides us with implementation and installation issues that can be expected to arise in similar mass-transport facilities. Capturing the video feed in a real-world situation can be problematic, as there should be no disruption in operational capability of existing security systems. The optimal approach would be to simply use Internet Protocol (IP) camera feeds, however, in many existing surveillance systems the cameras are analogue and often their streams are fed to relatively old digital recording equipment. Limitations of such systems can include low resolution, recording only a few frames per second, non-uniform time delay between frames, and proprietary codecs. To avoid disruption while at the same time obtaining video streams which are more appropriate for an intelligent surveillance system, it is useful to tap directly into the analogue video feeds and process them via dedicated analogue-to-digital video matrix switches.

The face recognition techniques were implemented with an aim to be fast as well as integrable into larger commercial intelligent surveillance systems. This necessitated the conversion of Matlab code into C++, which was non-trivial. Certain parts of the original code relied on elaborate functions and toolkits included with Matlab, which we had to re-implement. Furthermore, our experience also shows that while research code written by scientists/engineers (who are not necessarily professional programmers) might be sufficient to obtain experimental results which can be published, more effort is required to ensure the code is in a maintainable state as well as to guarantee that the underlying algorithm implementation is stable.

Apart from the technical challenges, issues in many other domains may also arise. Privacy laws or policies at the national, state, municipal or organisational level may prevent surveillance footage being used for research even if the video is already being used for security monitoring — the primary purpose of the data collection is the main issue here. Moreover, without careful consultation and/or explanation, privacy groups as well as the general public can become uncomfortable with security research. Some people may simply wish not to be recorded as they have no desire in having photos or videos of themselves being viewable by other people. Plaques and warning signs indicating that surveillance recordings are being gathered for research purposes may allow people to consciously avoid monitored areas, possibly invalidating results. Nevertheless, it is our experience that it is possible to negotiate a satisfying legal framework within which real-life trials of intelligent surveillance systems can take place.

6 Acknowledgements

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

References

[1] V. Blanz, P. Grother, P. Phillips, and T. Vetter. Face recognition based on frontal views generated from non-frontal images. In *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition*, volume 2, pages 454–461, 2005.

[2] K. Bowyer, K. Chang, and P. Flynn. A survey of approaches and challenges in 3D and multi-modal 3D+2D face recognition. *Computer Vision and Image Understanding*, 101(1):1–15, 2006.

[3] F. Cardinaux, C. Sanderson, and S. Bengio. User authentication via adapted statistical models of face images. *IEEE Trans. Signal Processing*, 54(1):361–373, 2006.

[4] T. Cootes, G. Edwards, and C. Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):681–685, 2001.

[5] T. Cootes and C. Taylor. Active shape models - 'smart snakes'. In *Proceedings of British Machine Vision Conference*, pages 267–275, 1992.

[6] T. Cootes, K. Walker, and C. Taylor. View-based active appearance models. In *Proceedings of 4th IEEE International Conference on Automatic Face and Gesture Recognition*, pages 227–232, 2000.

[7] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray. Visual cetegorization with bags of keypoints. In *Workshop on Statistical Learning in Computer Vision (in conjunction with ECCV'04)*, 2004.

[8] R. Duda, P. Hart, and D. Stork. *Pattern Classification*. Wiley, 2nd edition, 2001.

[9] G. Francisco, S. Roberts, K. Hanna, J.S., and Heubusch. Critical infrastructure security confidence through automated thermal imaging. In *Infrared Technology and Applications XXXII*, volume SPIE 6206, 2006.

[10] L. Fuentes and S. Velastin. From tracking to advanced surveillance. In *Proceedings of International Conference on Image Processing Conference (ICIP 2003)*, volume 3, pages III 121–4, 2003.

[11] R. Gonzales and R. Woods. *Digital Image Processing*. Addison-Wesley, 1992.

[12] T. Kadir and M. Brady. Saliency, scale and image description. *International Journal of Computer Vision*, 45(2):83–105, 2001.

[13] T. S. Lee. Image representation using 2D Gabor wavelets. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 18(10):959–971, 1996.

[14] S. Lucey and T. Chen. Learning patch dependencies for improved pose mismatched face verification. In *IEEE Conf. Computer Vision and Pattern Recognition*, volume 1, pages 909–915, 2006.

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

[16] E. Nowak, F. Jurie, and B. Triggs. Sampling strategies for bag-of-features image classification. In *European Conference on Computer Vision (ECCV), Part IV, Lecture Notes in Computer Science (LNCS)*, volume 3954, pages 490–503, 2006.

[17] P. Phillips, P. Grother, R. Micheals, D. Blackburn, E. Tabassi, and M. Bone. Face recognition vendor test 2002. In *Proceedings of Analysis and Modeling of Faces and Gestures*, page 44, 2003.

[18] P. Phillips, H. Moon, S. Rizvi, and P. Rauss. The FERET evaluation methodology for face-recognition algorithms. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22(10):1090–1104, 2000.

[19] Y. Rodriguez, F. Cardinaux, S. Bengio, and J. Mariethoz. Measuring the performance of face localization systems. *Image and Vision Computing*, 24:882–893, 2006.

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

[21] N. Schraudolph. A fast, compact approximation of the exponential function. *Neural Computation*, 11:853–862, 1999.

[22] T. Shan, B. Lovell, and S. Chen. Face recognition robust to head pose from one sample image. In *Proc. 18th Int. Conf. Pattern Recognition (ICPR)*, volume 1, pages 515–518, 2006.

[23] T. Sim, S. Baker, and M. Bsat. The CMU pose, illumination, and expression database. *IEEE. Trans. Pattern Analysis and Machine Intelligence*, 25(12):1615–1618, 2003.

[24] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *Proceedings of 9th International Conference on Computer Vision (ICCV)*, volume 2, pages 1470–1477, 2003.

[25] C. Wallraven, B. Caputo, and A. Graf. Recognition with local features: the kernel recipe. In *Proc. 9th International Conference on Computer Vision (ICCV)*, volume 1, pages 257–264, 2003.

[26] L. Wiskott, J. Fellous, N. Kuiger, and C. V. Malsburg. Face recognition by elastic bunch graph matching. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 19(7):775–779, 1997.

[27] F. Ziliani, S. Velastin, F. Porikli, L. Marcenaro, T. Kelliher, A. Cavallaro, and P. Bruneaut. Performance evaluation of event detection solutions: the creds experience. In *Proceedings of IEEE Conference on Advanced Video and Signal Based Surveillance*, pages 201–206, 2005.