

TOWARDS ROBUST FACE RECOGNITION FOR INTELLIGENT-CCTV BASED SURVEILLANCE USING ONE GALLERY IMAGE

Ting Shan ^{a,b}, Shaokang Chen ^{a,b}, Conrad Sanderson ^a and Brian C. Lovell ^{a,b}

^a NICTA, 300 Adelaide Street, Brisbane QLD 4000, Australia

^b ITEE, University of Queensland, Brisbane QLD 4072, Australia

Abstract

In recent years, the use of Intelligent Closed-Circuit Television (ICCTV) for crime prevention and detection has attracted significant attention. Existing face recognition systems require passport-quality photos to achieve good performance. However, use of CCTV images is much more problematic due to large variations in illumination, facial expressions and pose angle. In this paper we propose a pose variability compensation technique, which synthesizes realistic frontal face images from non-frontal views. It is based on modelling the face via Active Appearance Models and detecting the pose through a correlation model. The proposed technique is coupled with Adaptive Principal Component Analysis (APCA), which was previously shown to perform well in the presence of both lighting and expression variations. Experiments on the FERET dataset show up to 6 fold performance improvements. Finally, in addition to implementation and scalability challenges, we discuss issues related to on-going real life trials in public spaces using existing surveillance hardware.

1 Introduction

In recent years, the use of Closed-Circuit Television (CCTV) for surveillance has grown to an unprecedented level, especially after the 2005 London bombings and the 2001 terrorist attack in New York.

Hundreds of thousands of cameras have been installed in public areas all over the world, in places such as train stations, airports, car parks, Automatic Teller Machines (ATMs), vending machines and taxis. Based on the number of CCTV units on Putney High Street, it is “guesstimated” [9] that there are around 500,000 CCTV cameras in 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.

The work presented in this paper is part of a funded project to evaluate research as well as commercially available intelligent surveillance systems in the context of transit systems and public spaces. Here we focus on the automatic person recognition aspect of such systems, especially face-based methods.

Automatic face recognition under CCTV conditions is still on-going research and many problems still need to be solved before it can approach the capability of the human perception system. While recognition on passport-quality photos has achieved good results, CCTV conditions are much more challenging. This is not just due to the gross similarity of all faces but also because of the large differences between face images of the same person due to variations in lighting conditions, expression and pose (e.g. due to where the CCTV cameras are usually positioned). Examples of these variations are shown in Figures 1 through 3.

Recent research on face recognition has been focused on diminishing the impact of the abovementioned nuisance factors [7, 11, 13]. Techniques such as Adaptive Principal Component Analysis (APCA) as well as Rotated APCA were developed to compensate for illumination and expression variations [2]. In this paper we propose a pose variability compensation technique, to be used in conjunction with APCA, which synthesizes realistic frontal face images from non-frontal views.

The remainder of this paper is structured as follows. In Section 2 we present the pose compensation technique, followed by Section 3 where we overview the APCA face classification approach. In Section 4 we evaluate the performance of pose compensation coupled with APCA and contrast its performance to the standard PCA approach, with and without pose compensation. Concluding remarks and future avenues of research are given in Section 5, where we also discuss issues related to real life trials in public spaces using existing surveillance hardware, as well as the implementation and scalability of face recognition algorithms in the context of mass surveillance.



Figure 1. Examples of appearance changes due to variation in the direction of the lighting source (front, above, bottom, right, and left lighting)

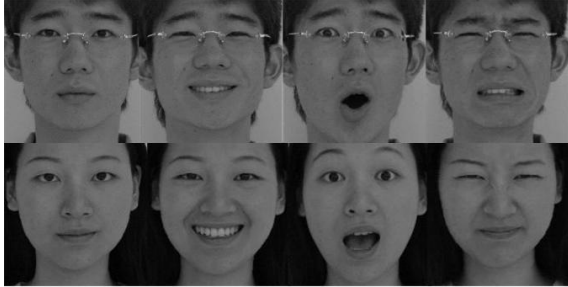


Figure 2. Examples of appearance changes due to variation in expression (neutral, smiling, surprised, and sad)



Figure 3. Examples of appearance changes due to variation in the presentation angle.

2 Pose Compensation

In this section we describe a technique for synthesizing realistic frontal face images from non-frontal views. The technique is based on deformable models popularised by Cootes et al., namely Active Shape Models (ASMs) [4] and Active Appearance Models (AAMs) [3]. We firstly overview ASMs and AAMs, then describe pose estimation via a correlation model, followed by a description of frontal view synthesis.

2.1 AAM Models

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:

$$\mathbf{f} = [x_1, x_2, \dots, x_N, y_1, y_2, \dots, y_N]^T \quad (1)$$

In ASM, a face shape is represented by:

$$\mathbf{f} = \bar{\mathbf{f}} + \mathbf{P}_s \mathbf{b}_s \quad (2)$$

where $\bar{\mathbf{f}}$ 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:

$$\mathbf{g} = \bar{\mathbf{g}} + \mathbf{P}_g \mathbf{b}_g \quad (3)$$

where $\bar{\mathbf{g}}$ 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:

$$\mathbf{b} = \begin{bmatrix} \mathbf{W}_s \mathbf{b}_s \\ \mathbf{b}_g \end{bmatrix} = \begin{bmatrix} \mathbf{W}_s \mathbf{P}_s^T (\mathbf{f} - \bar{\mathbf{f}}) \\ \mathbf{P}_g^T (\mathbf{g} - \bar{\mathbf{g}}) \end{bmatrix} \quad (4)$$

where \mathbf{W}_s is a diagonal matrix which represents the change between shape and texture. Via PCA we can represent \mathbf{b} as:

$$\mathbf{b} = \mathbf{P}_c \mathbf{c} \quad (5)$$

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:

$$\mathbf{f} = \bar{\mathbf{f}} + \mathbf{Q}_s \mathbf{c} \quad (6)$$

$$\mathbf{g} = \bar{\mathbf{g}} + \mathbf{Q}_g \mathbf{c} \quad (7)$$

where

$$\mathbf{Q}_s = \mathbf{P}_s \mathbf{W}_s^{-1} \mathbf{P}_{cs} \quad (8)$$

$$\mathbf{Q}_g = \mathbf{P}_g \mathbf{P}_{cg} \quad (9)$$

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.:

$$\mathbf{P}_c = \begin{bmatrix} \mathbf{P}_{cs} \\ \mathbf{P}_{cg} \end{bmatrix} \quad (10)$$

2.2 AAM Search

Finding the best parameters to represent an image can be treated as an optimisation problem where the difference between a given image and the image constructed by the model is minimised. Let us define the difference $\delta \mathbf{I} = \mathbf{I}_i - \mathbf{I}_m$, where \mathbf{I}_i is the vector of pixel values covered in the shape region, and \mathbf{I}_m is the vector of pixel values generated by the current model parameters \mathbf{c} . We've elected to minimize the magnitude of the difference, $\Delta = |\delta \mathbf{I}|^2$, by updating \mathbf{c} via a first-order Taylor expansion based method [3].

2.3 Pose Estimation using Correlation Models

Following [5], let us assume that the model parameter \mathbf{c} is approximately related to the viewing angle, θ , by a correlation model:

$$\mathbf{c} \approx \mathbf{c}_0 + \mathbf{c}_c \cos(\theta) + \mathbf{c}_s \sin(\theta) \quad (11)$$

where \mathbf{c}_0 , \mathbf{c}_c and \mathbf{c}_s are coefficient vectors for correlation model 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 $(\mathbf{c}^{[i]})_{i \in 1, \dots, |\Omega|}$ and $([1, \cos(\theta^{[i]}), \sin(\theta^{[i]})])_{i \in 1, \dots, |\Omega|}$, 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:

$$\mathbf{c}^{[new]} - \mathbf{c}_0 = [\mathbf{c}_c \ \mathbf{c}_s] \begin{bmatrix} \cos(\theta^{[new]}) & \sin(\theta^{[new]}) \end{bmatrix}^T. \quad (12)$$

Let \mathbf{R}_c^{-1} be the left pseudo-inverse of the matrix $[\mathbf{c}_c \ \mathbf{c}_s]$. Eqn. (12) can then be rewritten as:

$$\mathbf{R}_c^{-1} (\mathbf{c}^{[new]} - \mathbf{c}_0) = \begin{bmatrix} \cos(\theta^{[new]}) & \sin(\theta^{[new]}) \end{bmatrix}^T. \quad (13)$$

Let $[x_\alpha \ y_\alpha] = \mathbf{R}_c^{-1} (\mathbf{c}^{[new]} - \mathbf{c}_0)$, 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 entirely accurate due to land mark annotation errors or regression learning errors.

2.4 Frontal View Synthesis

After the estimation of $\theta^{[new]}$, we can use the model to synthesize new views. Here we will synthesize a frontal view face image, which will be used for face recognition.

Let \mathbf{c}_{res} be the residual vector which is not explained by the correlation model:

$$\mathbf{c}_{res} = \mathbf{c}^{[new]} - (\mathbf{c}_0 + \mathbf{c}_c \cos(\theta^{[new]}) + \mathbf{c}_s \sin(\theta^{[new]})) \quad (14)$$

Note that bracketed term can be interpreted as the mean face for angle $\theta^{[new]}$. To reconstruct at an alternate angle, $\theta^{[alt]}$, we can add the residual vector to the mean face for that angle:

$$\mathbf{c}^{[alt]} = \mathbf{c}_{res} + (\mathbf{c}_0 + \mathbf{c}_c \cos(\theta^{[alt]}) + \mathbf{c}_s \sin(\theta^{[alt]})) \quad (15)$$

As our aim is to synthesize the frontal view face, $\theta^{[alt]}$ is set to zero. Eqn. (15) hence simplifies to:

$$\mathbf{c}^{[alt]} = \mathbf{c}_{res} + \mathbf{c}_0 + \mathbf{c}_c \quad (16)$$

Based on Eqns. (6) and (7), the shape and texture for the frontal view can then be calculated by:

$$\mathbf{f}^{[alt]} = \bar{\mathbf{f}} + \mathbf{Q}_s \mathbf{c}^{[alt]} \quad (17)$$

$$\mathbf{g}^{[alt]} = \bar{\mathbf{g}} + \mathbf{Q}_g \mathbf{c}^{[alt]} \quad (18)$$

Examples of the synthesized images resulting from this procedure can be seen in Figure 4.



Figure 4. (a): frontal view and its AAM-based synthesized representation. (b): non-frontal view, synthesized representation at the original angle, and synthesized representation at $\theta^{alt} = 0$ (i.e. synthesized frontal view).

3 Adaptive Principal Component Analysis

Adaptive Principal Component Analysis (APCA) [2] inherits characteristics from both PCA and Fisher Linear Discriminant by warping the face subspace according to the within- and between-class covariance.

In the text below we shall use the following notation: $\mathbf{s}_{j,k}$ indicates a feature vector for the j -th class and k -th sample from that class (examples are given in Figure 5), $\mathbf{s}_{j,0}$ represents the vector for the reference image in the j -th class, $s_{i,j,k}$ indicates the i -th element of the vector $\mathbf{s}_{j,k}$, K_j indicates the number of samples in class j and N is the number of classes.

APCA is comprised of three steps, which are briefly overviewed below.

- Subspace Projection, where standard PCA is used to project every face image into the face subspace to generate m -dimensional feature vectors $\mathbf{s}_{j,k}$.
- Whitening Transformation, where the subspace is whitened according to its eigenvalues, with

a whitening power p . This compensates for the overweighing of leading eigenvectors. The corresponding transformation matrix is:

$$\mathbf{C}_{cov} = \text{diag} [\lambda_1^{-2p}, \lambda_2^{-2p}, \dots, \lambda_m^{-2p}] \quad (19)$$

- Filtering, where the feature vector elements are weighted according to the *Identification to Variation* value ψ with a filtering power q . The corresponding transformation matrix is:

$$\mathbf{R} = \text{diag} [\psi_1^q, \psi_2^q, \dots, \psi_m^q] \quad (20)$$

Here ψ_i^q is the ratio of the between class covariance and the within class covariance for dimension i , and is found as follows:

$$\psi_i = \frac{\frac{1}{N} \sum_{j=1}^N \frac{1}{K_j} \sum_{k=1}^{K_j} (s_{i,j,k} - \beta_{i,k})^2}{\frac{1}{N} \sum_{j=1}^N \frac{1}{K_j} \sum_{k=1}^{K_j} (s_{i,j,k} - \mu_{i,j})^2} \quad (21)$$

where $\beta_{i,k} = \frac{1}{N} \sum_{j=1}^N s_{i,j,k}$ and $\mu_{i,j} = \frac{1}{K_j} \sum_{k=1}^{K_j} s_{i,j,k}$.

Putting all the steps together, each face is represented by:

$$\hat{\mathbf{s}}_{j,k} = \mathbf{C}_{cov} \mathbf{R} \mathbf{s}_{j,k}. \quad (22)$$

The nearest neighbour rule is used for classification of each face vector. Automatic determination of p and q is achieved by optimisation of a cost function, which is a combination of the error rate and the ratio of between-class distance and within-class distance:

$$f_{cost}(p, q) = \sum_{j=1}^N \sum_{k=1}^{K_j} \sum_{n=1}^N e_{j,k,n} \quad (23)$$

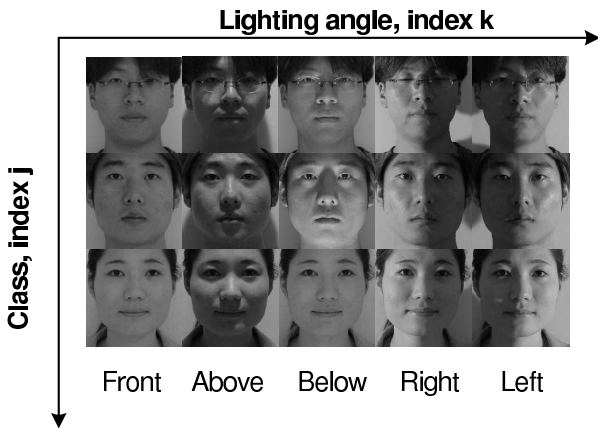


Figure 5. Examples of images taken under different illuminations in the Asian Face Image Database [8].

where

$$e_{j,k,n} = \begin{cases} \frac{d_{jj,k0}}{d_{jn,k0}} & \text{if } d_{jn,k0} < d_{jj,k0} \\ 0 & \text{otherwise} \end{cases} \quad (24)$$

and

$$d_{jj',kk'} = |\hat{\mathbf{s}}_{j,k} - \hat{\mathbf{s}}_{j',k'}|^2. \quad (25)$$

In the above, $d_{jj,k0}$ is the within-class distance between the illumination varied sample $\mathbf{s}_{j,k}$ and the standard reference image $\mathbf{s}_{j,0}$ while $d_{jn,k0}$ is the between-class distance between sample $\mathbf{s}_{j,k}$ and the normal lighting image $\mathbf{s}_{n,0}$ for class n . The cost function is optimized by a search in the space for p and q in the range $[-100, 100]$ with a pyramid scheme [12, 6].

APCA has been previously shown to work considerably better than standard PCA in the presence of illumination changes. To achieve robustness to expression changes, a further extension known as Rotated APCA can be used [2].

4 Experiments

The APCA classifier was first trained on the Asian Face Database [8]. Pose compensation experiments were performed on a subset of the FERET dataset [10], using face images from 46 persons with good AAM search results. For each person we used images at the following head poses: left 25° , 15° , 0° to right 15° and 25° . The original images were processed, via the method described in Section 2, to obtain synthesized frontal views.

For obtaining baseline performance of the APCA and standard PCA based classifiers, the original 0° view faces (46 images, one image per person) were used as gallery images and the remainder of the original faces (184 images) were used for testing. To evaluate the proposed pose compensation approach, the training set was comprised of synthesized frontal view faces which originated from the original 0° view images. The remaining synthesized faces (which originated from left/right 25° and 15° views) were used for testing. Recognition results are presented in Figure 6.

The results firstly indicate that the proposed pose compensation approach considerably increases the performance for non-frontal faces, with most of the gains occurring at 25° . The best overall performance is obtained when coupling the approach with the APCA based classifier, though large error reductions are also observed when pose compensation is coupled with standard PCA. Looking across combinations, the standard PCA approach with no pose compensation achieves a recognition rate of 9% for faces at left 25° , while APCA coupled with pose-compensation achieves a recognition rate of 57%. The results also show that the APCA approach is able to achieve relatively small error reductions by itself, even though it wasn't explicitly designed for handling varying pose.

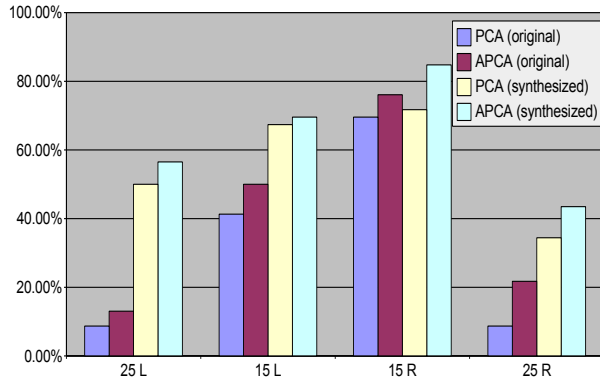


Figure 6. Recognition rates for PCA and APCA on original and synthesized faces.

5 Discussion

In this paper we proposed a pose variability compensation technique which synthesizes realistic frontal face images from non-frontal views. The technique is based on modelling faces via Active Appearance Models and Active Shape Models. Coupled with Adaptive Principal Component Analysis (APCA) based face classification approach, the resulting system shows considerable reductions in error rates.

In the current pose compensation approach we are explicitly synthesizing face images, which are then processed by APCA prior to classification. Recently it has been observed that it is possible to accomplish synthesis directly in the feature domain, indicating that the image synthesis step may be bypassed [11], reducing the complexity of the resulting system. Within the ASM/AAM based pose compensation framework, a similar approach might be based on using \mathbf{c}_{res} from Eqn. (14) directly for classification. The reasoning is as follows. In Eqn. (14) the term $(\mathbf{c}_0 + \mathbf{c}_c \cos(\theta) + \mathbf{c}_s \sin(\theta))$ can be interpreted as the mean face at angle θ . The difference between \mathbf{c} (which represents the given face at angle θ) and the above term can hence be interpreted as removing the effect of the angle, resulting in a pose-independent representation. We are currently looking into this aspect in more detail.

As a part of the project mentioned in Section 1, we are working towards real-life trials of face recognition technologies. One of our test-beds 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 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 an IP feed. 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 into the analogue video feeds and process them via dedicated analogue-to-digital video matrix switches.

Apart from the technical challenges, issues in other domains can arise. Laws or policies at the national, state, municipal or organisational level may limit surveillance footage from being used for purposes other than security. In other words, while CCTV can be expressly used for surveillance, it does not automatically mean that CCTV recordings can be used for surveillance research. When surveillance footage is not directly used for security, 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 (e.g. as part of publicly available datasets). In these cases, explicit protocols for handling and accessibility of surveillance footage need to be agreed on. Furthermore, plaques and warning signs indicating when surveillance recordings are being gathered for research purposes allow people to consciously avoid surveilled areas.

Our first trial of a relatively straightforward PCA-based face recognition system highlighted the need to address the challenges presented in Section 1 (see Figure 7 for real-life examples). In addition, there are several other important issues before a face recognition system can be employed as a component in an intelligent surveillance system. We discuss these below.

Scalability and real-time performance. A face recognition technique should be able to handle large volumes of people (e.g. peak hour at a railway station). Furthermore, a surveillance system is often comprised of a multitude of cameras, meaning there is a multitude of video streams to be processed. Obviously this is not a trivial amount of raw information. While it is possible to setup elaborate parallel computation environments, there are cost considerations limiting the number of CPUs available for processing. As such a face recognition algorithm for ICCTV environments should be able to run in at least real-time, which necessarily limits its complexity. This limitation may exclude promising but computationally expensive face recognition techniques, such as those based on pseudo-2D Hidden Markov Models [1].

Conversion of research code into production code. Research code is often written by scientists/engineers (not necessarily professional programmers) for the explicit purpose of evaluating new methods. While this is sufficient to obtain experimental results which can be published, there can be little incentive to keep the code in a maintainable state or to guarantee that the underlying

algorithm implementation is actually stable. Furthermore, research code is often written in Matlab which requires conversion into a language such as C++, to allow faster processing and integration with other software (e.g. via an SDK). The conversion may not be trivial if, for example, the experimental implementation relies on elaborate functions and toolkits included with Matlab.

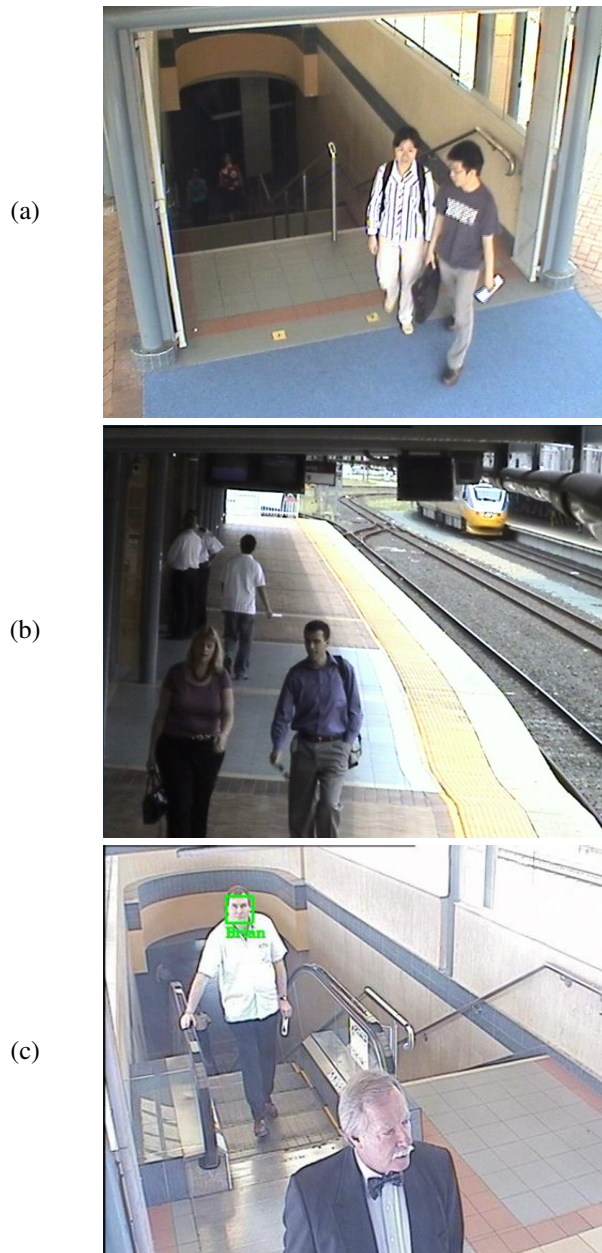


Figure 7. Several frames from CCTV cameras located at a railway station in Brisbane (Australia), demonstrating some of the variabilities present in real-life conditions: (a) varying face pose, (b) illumination from one side, (c) varying size and pose. In (c), the green box shows Brian Lovell being recognized by our face recognition algorithms in an early trial.

Acknowledgements

The authors thank Abbas Bigdeli and Erik Berglund for useful suggestions. 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] F. Cardinaux, C. Sanderson, and S. Bengio. User authentication via adapted statistical models of face images. *IEEE Transactions on Signal Processing*, 54(1):361–373, 2006.
- [2] S. Chen and B. Lovell. Illumination and expression invariant face recognition with one sample image. In *Proceedings of 17th International Conference on Pattern Recognition*, volume 1, pages 300–303, 2004.
- [3] T. Cootes, G. Edwards, and C. Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):681–685, 2001.
- [4] T. Cootes and C. Taylor. Active shape models - ‘smart snakes’. In *Proceedings of British Machine Vision Conference*, pages 267–275, 1992.
- [5] 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.
- [6] A. R. F. da Silva. Evolutionary time-frequency analysis. In *Proceedings of 2000 Congress on Evolutionary Computation*, pages 2:1102–1109, 2000.
- [7] Y. Gao and M. Leung. Face recognition using line edge map. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(6):764–779, 2002.
- [8] K.-A. Kim, S.-Y. Oh, and H.-C. Choi. Facial feature extraction using PCA and wavelet multi-resolution images. In *Proceedings of 6th International Conference on Automatic Face and Gesture Recognition*, pages 439–444, 2004.
- [9] M. McCahill and C. Norris. *Urbaneye: CCTV in London*. Centre for Criminology and Criminal Justice, University of Hull, UK, 2002.
- [10] P. Phillips, H. Moon, S. Rizvi, and P. Rauss. The FERET evaluation methodology for face-recognition algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(10):1090–1104, 2000.
- [11] C. Sanderson, S. Bengio, and Y. Gao. On transforming statistical models for non-frontal face verification. *Pattern Recognition*, 39(2):288–302, 2006.
- [12] P. Thevenaz and M. Unser. Pyramid approach to sub-pixel image fusion based on mutual information. In *Proceedings of IEEE International Conference on Image Processing*, volume 1, pages 265–268, 1996.
- [13] A. Yilmaz and M. Gökmen. Eigenhill vs. eigenface and eigenedge. *Pattern Recognition*, 34(1):181–184, 2001.