Intelligent Surveillance and Pose-Invariant 2D Face Classification

Brian C. Lovell[1,2], Conrad Sanderson[1], and Ting Shan[1]

[1] NICTA, 300 Adelaide St, Brisbane, QLD 4000, Australia
[2] SAS, ITEE, University of Queensland, Brisbane, QLD 4072, Australia

Abstract. We describe recent advances in a project being undertaken to trial and develop advanced surveillance systems for public safety. One goal of the project is to trial commercial technologies in public spaces to evaluate their performance. Another is to develop and trial enhanced capabilities that will lead to more effective surveillance systems. A key technology being developed within the group is reliable face in the crowd identification from conventional CCTV cameras. While this is acknowledged to be a challenging problem, we have made considerable progress on several fundamental issues including recognition robust to large pose angles. We also describe a reconfigurable smart camera we are developing to handle the problem of obtaining high resolution face images while simultaneously surveilling large crowds in real-time.

1 Introduction

For isolated crimes such as assault and robbery, it is well-known that video surveillance is highly effective in helping to find and successfully prosecute the perpetrators. Moreover, electronic surveillance has been shown to act as a significant deterrent to crime. Cost is mitigated by recording most of the camera feeds without any human monitoring — if an event is reported to security, the relevant video is manually extracted and reviewed. In recent times the game has changed due to the human and political cost of successful terrorist attacks on soft targets such as mass transport systems. Traditional forensic analysis of recorded video after the event is simply not an adequate response from government and large business. This seachange in the security sector is due to the fact that in the case of suicide attacks there is simply no possibility of prosecution after the event, so simply recording surveillance video provides no terrorism deterrent. Video of successful attacks may indeed add impact to the political message of the perpetrators by highlighting the failure of Western governments to protect their populace. A pressing need is emerging to detect events and persons of interest using video surveillance before such harmful actions can occur. This means that cameras must be monitored at all times. Now the problem is how do we cost-effectively monitor thousands of surveillance cameras to detect the rare events of security interest in real-time.

The problem is that human monitoring of surveillance systems requires a large number of personnel, resulting in high ongoing costs and 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."

In 2006 NICTA was awarded a research grant to conduct long term trials of advanced ICCTV technologies in important and sensitive public spaces such as major ports and railway stations [1]. One such advanced technology is a system that projects all the CCTV video feeds on to a 3D model of the environment providing rapid situational assessment facilitating a rapid response to situations arising as shown in Figure 1. The trial will highlight operational and capability deficiencies in current ICCTV systems and will focus NICTA's research on capability gaps. The project is thus a unique collaboration of researchers, vendors, and user agencies aimed at delivering advances in computer vision and pattern recognition for human activity recognition.

One of the"test-beds" we are using for our advanced surveillance field trials is a railway station in Brisbane (Australia), which provides us with

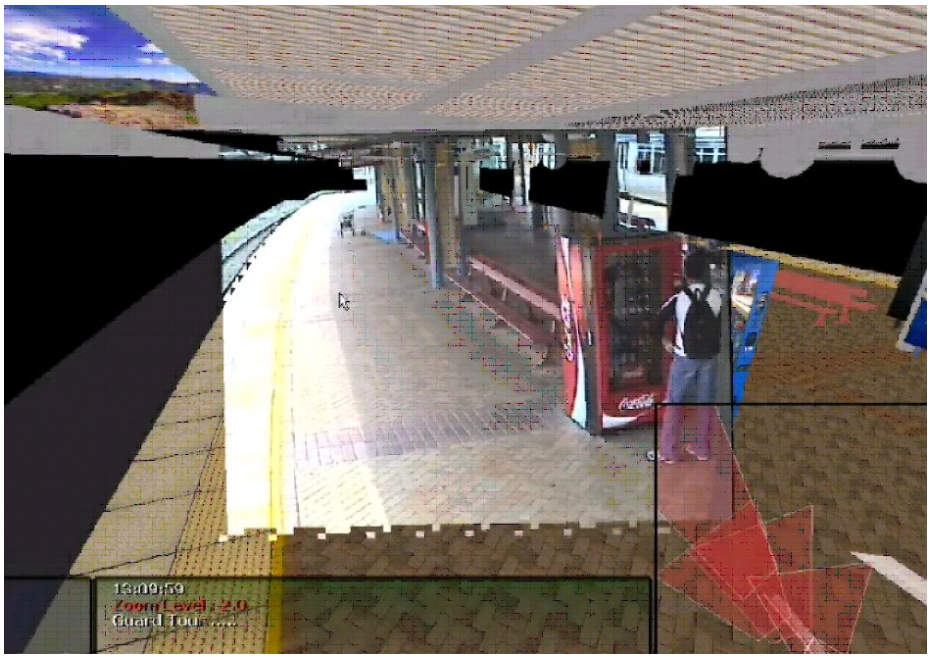

Fig. 1. Immersive 3-D Visual Presentation of Camera View and 3-D model of the railway platform.

implementation and installation issues that can be expected to arise in similar mass-transport facilities. Capturing the camera feeds in a real-world situation can be problematic, as there must be no disruption in operational capability of existing security systems. The optimal approach would be to simply use IP camera feeds. However, in many existing surveillance systems the cameras are analog and often their streams are fed to relatively old analog or digital recording equipment. Limitations of such systems may 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 suitable for an intelligent surveillance system, it is useful to tap directly into the analog video feeds and process them via dedicated analog-to-digital video matrix switches.

Apart from the technical challenges, issues in many other domains may also arise. Privacy laws or policies at the national, state, municipal or organizational 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 the needs of 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.

A key technology being developed within our group for prevention of crime and terrorism is the reliable detection of "persons of interest" through face recognition. While automatic face recognition of cooperative subjects has achieved good results in controlled applications such as passport control, CCTV conditions are considerably more challenging. Examples of real life CCTV conditions captured at the railway station are shown in Figure 2.

Nuisance factors such as varying pose, illumination, and expression (PIE) can greatly affect recognition performance. According to Phillips *et al.* head pose is believed to be the hardest factor to model [2]. 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 recognition system must have highly effective pose compensation.

A further complication is that in many practical situations there is 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 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 head pose variation include the synthesis of new images at previously unseen views [3, 4], direct synthesis of face model parameters [5] and local feature based representations [6–8]. We note that while true 3D based approaches in theory allow face matching at various poses, current 3D sensing hardware has too many limitations [9] including cost and range. Moreover unlike 2D recognition, 3D technology cannot be retrofitted to existing surveillance systems. Certainly 2D recognition presents much greater technical challenges due to difficulties presented by illumination and shadow effects as was famously noted by the great Leonardo da Vinci (1452-1519):

> After painting comes Sculpture, a very noble art, but one that does not in the execution require the same supreme ingenuity as the art of painting, since in two most important and difficult particulars, in foreshortening and in light and shade, for which the painter has to invent a process, sculpture is helped by nature.

The paper is structured as follows. In section 2 we overview the AAM-based face synthesis technique and present the modified form. Then in section 3 we overview the local feature approach. section 4 evaluates the performance of several proposed pose robust face recognition techniques on the FERET and PIE databases. In section 5 we describe our NICTA smart camera for wide-area surveillance. Finally we draw our conclusions in section 6 and then describe future directions for the project in section 7.

2 Methods Based on ASMs and AAMs

In this section we describe face modelling based on deformable models popularised by Cootes et al., namely Active Shape Models (ASMs) [10] and

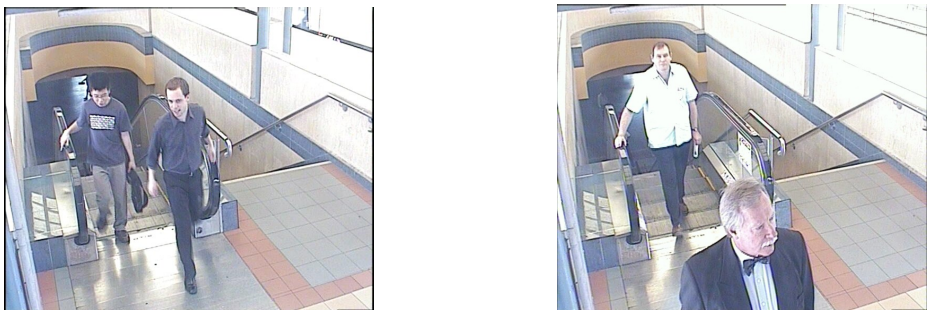

Fig. 2. Examples of typical face pose under surveillance conditions.

Active Appearance Models (AAMs) [11]. 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.

2.1 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:

$$\mathbf{f} = [\ x_1, x_2, \cdots, x_N,\ y_1, y_2, \cdots, y_N\ ]^T . \tag{1}$$

In ASM, a face shape is represented by:

$$\mathbf{f} = \overline{\mathbf{f}} + \mathbf{P}_s \mathbf{b}_s \tag{2}$$

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:

$$\mathbf{g} = \overline{\mathbf{g}} + \mathbf{P}_g \mathbf{b}_g \tag{3}$$

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:

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

where $\mathbf{W}_s$ is a diagonal matrix which represents the change between shape and texture. Through Principal Component Analysis (PCA) [12] we can represent $\mathbf{b}$ as:

$$\mathbf{b} = \mathbf{P}_c \mathbf{c} \tag{5}$$

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:

$$\mathbf{f} = \overline{\mathbf{f}} + \mathbf{Q}_s \mathbf{c} \tag{6}$$

$$\mathbf{g} = \overline{\mathbf{g}} + \mathbf{Q}_g \mathbf{c} \tag{7}$$

where

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

$$\mathbf{Q}_g = \mathbf{P}_g \mathbf{P}_{cg} \tag{9}$$

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.:

$$\mathbf{P}_c = \begin{bmatrix} \mathbf{P}_{cs} \\ \mathbf{P}_{cg} \end{bmatrix}. \tag{10}$$

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

2.2 Pose Estimation

Following [13], let us assume that the model parameter $\mathbf{c}$ is approximately related to the viewing angle, θ, by a correlation model:

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

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:

$$\mathbf{c}^{[new]} - \mathbf{c}_0 = [\ \mathbf{c}_c\ \mathbf{c}_s\ ] \left[\ \cos(\theta^{[new]})\ \sin(\theta^{[new]})\ \right]^T. \tag{12}$$

Let $\mathbf{R}_c^{-1}$ be the left pseudo-inverse of the matrix $[\ \mathbf{c}_c\ \mathbf{c}_s\ ]$. Eqn. (20) can then be rewritten as:

$$\mathbf{R}_c^{-1}\left(\mathbf{c}^{[new]} - \mathbf{c}_0\right) = \left[\ \cos(\theta^{[new]})\ \sin(\theta^{[new]})\ \right]^T. \tag{13}$$

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.

2.3 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:

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

Fig. 3. 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).

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

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

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

$$\mathbf{c}^{[alt]} = \mathbf{c}_{res} + \mathbf{c}_0 + \mathbf{c}_c. \tag{16}$$

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

$$\mathbf{f}^{[alt]} = \overline{\mathbf{f}} + \mathbf{Q}_s \mathbf{c}^{[alt]} \tag{17}$$

$$\mathbf{g}^{[alt]} = \overline{\mathbf{g}} + \mathbf{Q}_g \mathbf{c}^{[alt]}. \tag{18}$$

Examples of synthesized faces are shown in Fig. 3. Each synthesized face can then be processed via the standard Principal Component Analysis (PCA) technique to produce features which are used for classification [4].

2.4 Direct Pose-Robust Features

The bracketed term in Eqn. (14) 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.

Fig. 4. *Top row*: frontal mean-face and faces generated by adding person specific pose-independent features. *Bottom row*: mean-face at $+20^o$ and faces generated by adding person specific pose-independent features.

2.5 Remove Pose Effect using Correlation Model

Correlation Model and Pose Estimation Following [13], let us assume that the model parameter $\mathbf{c}$ is approximately related to the viewing angle, θ, by a correlation model:

$$\mathbf{c} \approx \mathbf{c}_0 + \mathbf{c}_c \cos(\theta) + \mathbf{c}_s \sin(\theta) \tag{19}$$

where $\mathbf{c}_0$, $\mathbf{c}_c$ and $\mathbf{c}_s$ are vectors which are learned from the training data. .

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:

$$\mathbf{c}^{[new]} - \mathbf{c}_0 = [\ \mathbf{c}_c\ \mathbf{c}_s\ ] \left[\ \cos(\theta^{[new]})\ \sin(\theta^{[new]})\ \right]^T \tag{20}$$

Let $\mathbf{R}_c^{-1}$ be the left pseudo-inverse of the matrix $[\ \mathbf{c}_c\ \mathbf{c}_s\ ]$. Eqn. (20) can then be rewritten as:

$$\mathbf{R}_c^{-1}\left(\mathbf{c}^{[new]} - \mathbf{c}_0\right) = \left[\ \cos(\theta^{[new]})\ \sin(\theta^{[new]})\ \right]^T \tag{21}$$

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 $\theta^{[new]} = \tan^{-1}(y_\alpha / x_\alpha)$.

Removing Pose Effect in Appearance After the estimation of $\theta^{[new]}$, we can use the correlation model to remove the effect of pose. Now

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

represents the standard parameter vector at pose θ, note that it's fixed at specific angle θ and changes when pose changes. Let $\mathbf{c}_{feature}$ be the feature vector which is generated by removing the pose effect from the correlation model: Note that the bracketed term in (14) can be interpreted as the mean face for angle $\theta^{[new]}$. Given any face image, we can use Active Appearance Models (AAMs) to estimate face model parameters $\mathbf{c}$ and use the correlation model as described above to remove pose effect. Each face image then can be characterized by $\mathbf{c}_{feature}$, which is pose-independent. Figure 5 shows different face images generated from mean face at certain angle by adding $\mathbf{c}_{feature}$. Note for recognition, there is no need to construct the face image itself.

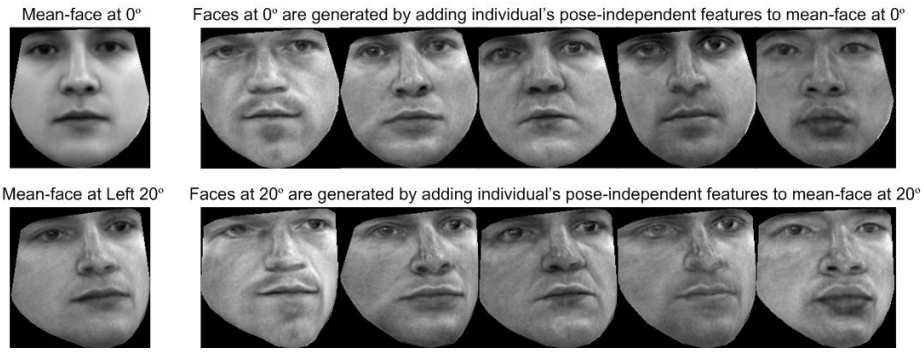

Fig. 5. Faces generated by adding individual's pose-independent features to mean-face

2.6 Face Recognition using Pose-Independent Features

Both the gallery face images and the given unknown face image can be represented by parameter vector $\mathbf{c}_{feature}$. To recognize a given face image becomes a problem of measuring the similarity between the parameter vector of the given face image and the vectors of the gallery images stored in the database. We applied two well known pattern recognition techniques: Mahalanobis distance and cosine measure for classification.

3 Methods Based on Bag-of-Features Approach

In this section we describe two local feature based approaches, with both approaches sharing a block based feature extraction method summarised in

section 3.1. 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 3.2) 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 3.3), a generic model (not specific to any person), representing "face words", is used to build histograms which are then compared for recognition purposes.

3.1 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[3] via the 2D DCT [16]. 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 [6].

3.2 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:

$$P(X|\lambda^{[i]}) = \prod_{n=1}^{N} P(\mathbf{x}_n|\lambda^{[i]}) = \prod_{n=1}^{N} \sum_{g=1}^{G} w_g^{[i]} \mathcal{N}\left(\mathbf{x}_n|\mu_g^{[i]}, \mathbf{\Sigma}_g^{[i]}\right) \tag{22}$$

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 [12], 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 [12].

3.3 Bag-of-Features with Histogram Matching

The technique presented in this section is an adaption of the "visual words" method used in image categorisation [17–19]. 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:

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

[3] While in this work we used the 2D DCT for describing each block (or patch), it is possible to use other descriptors, for example SIFT [14] or Gabor wavelets [15].

where w_g is the weight for Gaussian g and $p_g(\boldsymbol{x})$ is the probability of vector $\boldsymbol{x}$ 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 [20], or the simpler approach of summation of absolute differences [21]:

$$d(\boldsymbol{h}_A, \boldsymbol{h}_B) = \sum_{g=1}^{G} \left| \boldsymbol{h}_A^{[g]} - \boldsymbol{h}_B^{[g]} \right| \tag{23}$$

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

4 Face Recognition Robust to Pose

These pose robust face recognition algorithms were first proposed and compared in Sanderson, Shan, and Lovell (2007) [22]. The methods we compare are 1) baseline PCA or "eigenfaces" 2) Synthesis + PCA, where we synthesise frontal views from high pose angle views using deformable models popularised by Cootes *et al*. namely Active Shape Models (ASMs) [10] and Active Appearance Models (AAMs) [11], 3) Pose-robust features based on a modification of the previous method which avoids the synthesis step, 4) Direct bag of features based on GMMs, and 5) Histogram bag of features which is a faster and more scalable version of the previous method.

We are currently in the process of creating a suitable dataset for face classification in CCTV conditions. 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 [24] (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 [4]. 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 [12].

For the bag-of-features approaches, in a similar manner to [5], 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

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.

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.

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.

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.

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 (22), while for the histogram-based approach this involves just the sum of absolute differences between two histograms (Eqn. (23)). 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).

When using a fast approximation of the exp() 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 NICTA smart camera

One of the challenges of face recognition in a surveillance environment is to obtain faces of sufficient resolution to allow accurate recognition. To facilitate

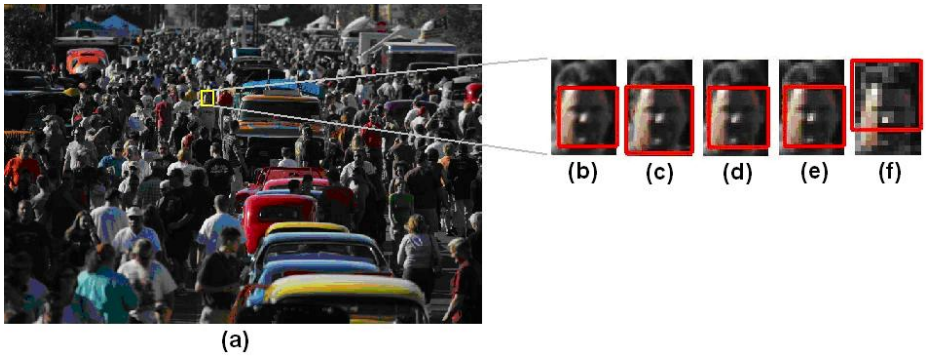

Fig. 6. Overall scene (a) ROI extracted from scene with resolution of 7Mp(b), 5Mp(c), 3Mp(d), 1Mp(e) and VGA(f).

this goal, we have been developing a smart camera which can surveil crowds while simultaneously extracting high resolution face images [25, 26].

5.1 Proposed smart camera architecture

Most existing smart camera designs use sensors ranging from 192x124 pixels to 640x480 pixels (VGA standard). In our design, we have decided to tailor the smart camera design to the task of face detection for crowd surveillance. Crowd surveillance usually surveils a wide area and often has multiple objects of interest in view. To classify these objects reliably we need high resolution images. However most of the scene is of little interest for automated analysis and can thus be acquired at much lower resolution. Our camera is designed to extract the objects of interest, in this case faces, at full sensor resolution while simultaneously obtaining a much lower resolution video of the entire scene. Figure 6 shows an example of how obtaining such high resolution images affects the accuracy of face detection performance.

In the example, the region of interest (ROI) is extracted from an image of a crowd of people (a). The face (b) extracted from a 7 MP (MegaPixel) high resolution image is much more recognizable than (f) extracted from the lower resolution (VGA) image. The extracted faces were also tested for suitability for automatic detection using a Viola-Jones face detection module. The images (c), (d) and (e) taken with 5, 3, and 1MP sensors were suitable for face detection. However, face cannot be correctly detected in the VGA image (f) because the image does not have enough details for the face detection module to work correctly.

5.2 System design constraints

There are several constraints that we have taken into consideration in designing our smart camera the main ones being:

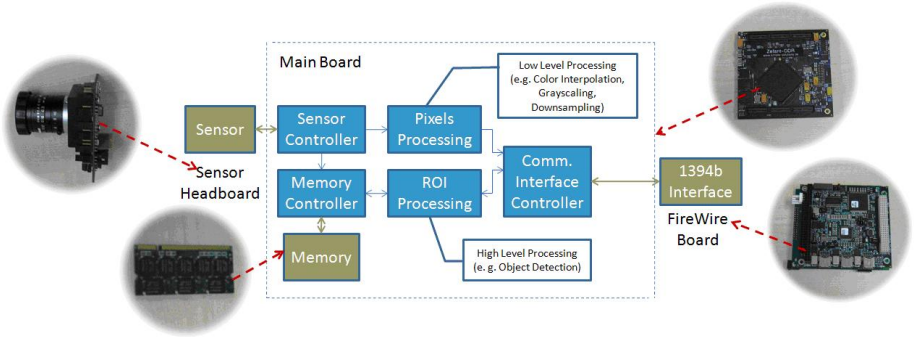

Fig. 7. Smart camera system architecture.

1. Real-time constraint: Meeting real-time constraint is an important issue since our smart camera targeted application in the surveillance area with real-time response requirement.
2. Hardware resources: In our smart camera design, we have chosen the Spartan FPGA-based series. A Spartan FPGA is a low cost version of the high performance Virtex family. Hence, the Spartan series has reduced hardware resources.
3. Bandwidth constraint: Higher resolution image would require higher data transfer rate. In our current design, our smart camera has a limited communication bandwidth to a host PC of 800Mbps.
4. Memory capacity: The memory storage is highly dependable on the image sensor. For example, a 5Mp image sensor has a total raw pixels value of 5Mp times Bit depth.

5.3 Hardware specification of NICTA smart camera prototype

Figure 7 shows an overview of our proposed smart camera architecture design. To meet our design requirement, we have chosen a 5 Megapixel (2592x1944 pixels) CMOS image sensor headboard manufactured by Micron. This sensor headboard could operate up to 14fps (frame-per-second) at full resolution. The main reasons to choose the CMOS image sensor are because unlike CCD image sensor, CMOS image sensor has parallel data access for faster data manipulation, low power consumption and the on-chip functionality. The main board of our smart camera has a Spartan-3 series FPGA chip (XS3C5000). For this project, Spartan FPGA is chosen as the processing target device primarily because of its low cost and low power consumption. The main board also has a DDR SDRAM slot. To ensure that our smart camera system could handle the high data rates (from the high resolution image sensor), we have installed a 1GB DDR SDRAM as the main frame buffer of the camera. As for the camera communication interface to the host PC, we have decided to use a FireWire 800 (1394b) communication protocol. A FireWire board that consists of Texas Instrument's 1394b Link Layer and

Physical Layer controller chips and 3 FireWire 800 ports is used in our design. All three boards were interfaced together and powered using a custom-designed PCB board (interface board).

Figure 8 shows a picture of our smart camera prototype and Table 4 summarises the basic specification of our prototype.

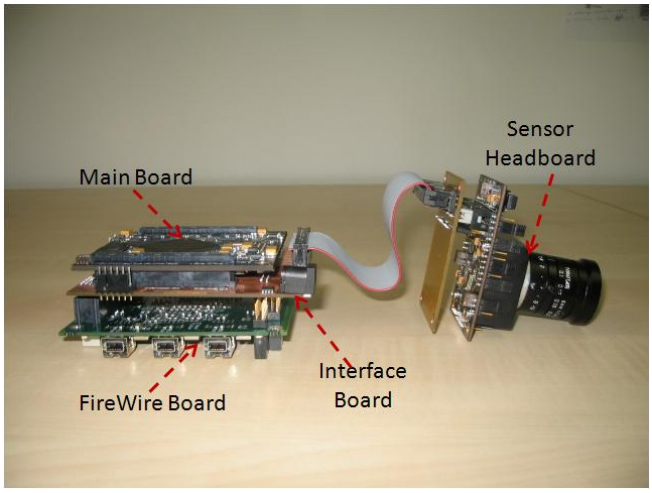

Fig. 8. NICTA smart camera prototype

Table 4. NICTA smart camera specification

Parameter	**Value)**
Sensor Type	CMOS
Resolution	2592 x 1944
Processing Element	Spartan-3 FPGA
Comm. Interface	FireWire800
Physical Dimension	90 x 90 x 150 mm^3

6 Conclusions

In this paper we have described our advanced surveillance project and the need for computer based monitoring of CCTV video feeds. Next we described recent advances in robust face recognition primarily addressing the issue of pose compensation. We acknowledge that apart from pose variations, imperfect face localisation [27] 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. Finally we described our reconfigurable smart camera project designed to deliver high quality face images from crowd scenes.

7 Future Work

In 2008 and beyond we are additionally targeting critical infrastructure protection in the maritime environment as illustrated in Figure 9. We will be applying intelligent surveillance techniques to address terrorism concerns through identity recognition as well as the day to day operational issues such as monitoring traffic on the land and ships in the 100km long shipping channel. Note that surveillance systems are installed primarily to address commercial concerns such as theft, property damage, and liability issues. They are typically not installed to address the very rare events associated with terrorism. So a system designed for enhanced counter-terrorism capabilities must also be operationally more efficient than a conventional surveillance system to justify the enormous cost of system upgrading.

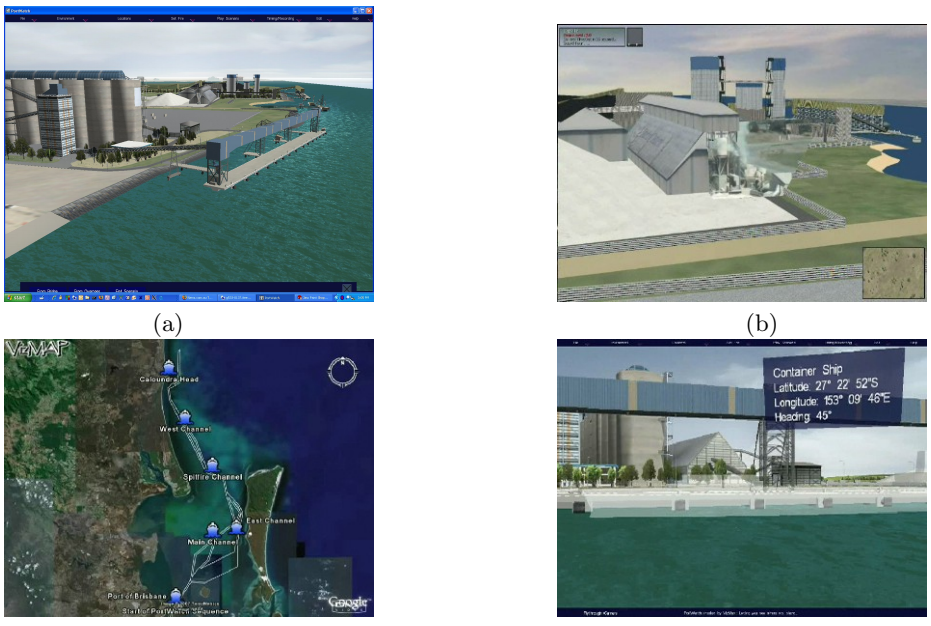

Fig. 9. Images from port project: (a) Grain handling facilities (b) surveillance video of real fire projected on model of port (c) over 100 km of channel to protect (d) shipping labelled and positioned in the port model using AIS data feeds.

Ships currently use the Automatic Identification System (AIS) which provides a means for them to electronically exchange ship data including: identification, position, course, and speed, with other nearby ships and ports. The new project will require the integration of surveillance cameras with all other relevant information such as AIS data, harbour marine radar, weather stations, and possibly even CBR (Chemical, Biological, Radiation) wireless sensor networks. This will provide an integrated information system for the port which will lead to greater operational efficiency, less lost time, and faster recovery due to better management of incidents.

8 Acknowledgements

This project is supported by a grant from the Australian Government Department of the Prime Minister and Cabinet and by the Australian Research Council through the Research Network for Securing Australia. NICTA is funded by the Australian Government's *Backing Australia's Ability* initiative, in part through the Australian Research Council. The author thanks Abbas Bigdeli, Shaokang Chen, Amelia Azman, Yasir Mustafah, and Erik Berglund for their major contributions to this work.

References

1. Bigdeli, A., Lovell, B., Sanderson, C.: Vision processing in intelligent cctv for mass transport security. In: Proc. of SAFE 2007: Workshop on Signal Processing Applications for Public Security and Forensics. (2007)
2. Phillips, P., Grother, P., Micheals, R., Blackburn, D., Tabassi, E., Bone, M.: Face recognition vendor test 2002. In: Proc. Analysis and Modeling of Faces and Gestures. (2003) 44
3. Blanz, V., Grother, P., Phillips, P., Vetter, T.: Face recognition based on frontal views generated from non-frontal images. In: Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition. Volume 2. (2005) 454–461
4. Shan, T., Lovell, B., Chen, S.: Face recognition robust to head pose from one sample image. In: Proc. 18th Int. Conf. Pattern Recognition (ICPR). Volume 1. (2006) 515–518
5. Sanderson, C., Bengio, S., Gao, Y.: On transforming statistical models for non-frontal face verification. Pattern Recognition **39** (2006) 288–302
6. Cardinaux, F., Sanderson, C., Bengio, S.: User authentication via adapted statistical models of face images. IEEE Trans. Signal Processing **54** (2006) 361–373
7. Lucey, S., Chen, T.: Learning patch dependencies for improved pose mismatched face verification. In: IEEE Conf. Computer Vision and Pattern Recognition. Volume 1. (2006) 909–915
8. Wiskott, L., Fellous, J., Kuiger, N., Malsburg, C.V.: Face recognition by elastic bunch graph matching. IEEE Trans. Pattern Analysis and Machine Intelligence **19** (1997) 775–779
9. Bowyer, K., Chang, K., Flynn., P.: A survey of approaches and challenges in 3D and multi-modal 3D+2D face recognition. Computer Vision and Image Understanding **101** (2006) 1–15

10. Cootes, T., Taylor, C.: Active shape models - 'smart snakes'. In: Proc. British Machine Vision Conference. (1992) 267–275
11. Cootes, T., Edwards, G., Taylor, C.: Active appearance models. IEEE Trans. Pattern Analysis and Machine Intelligence **23** (2001) 681–685
12. Duda, R., Hart, P., Stork, D.: Pattern Classification. 2nd edn. Wiley (2001)
13. Cootes, T., Walker, K., Taylor, C.: View-based active appearance models. In: Proc. 4th IEEE International Conference on Automatic Face and Gesture Recognition. (2000) 227–232
14. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision **60** (2004) 91–110
15. Lee, T.S.: Image representation using 2D Gabor wavelets. IEEE Trans. Pattern Analysis and Machine Intelligence **18** (1996) 959–971
16. Gonzales, R., Woods, R.: Digital Image Processing. Addison-Wesley (1992)
17. Csurka, G., Dance, C., Fan, L., Willamowski, J., Bray, C.: Visual cetegorization with bags of keypoints. In: Workshop on Statistical Learning in Computer Vision (in conjunction with ECCV'04). (2004)
18. Sivic, J., Zisserman, A.: Video google: A text retrieval approach to object matching in videos. In: Proc. 9th International Conference on Computer Vision (ICCV). Volume 2. (2003) 1470–1477
19. Nowak, E., Jurie, F., Triggs, B.: 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. (2006) 490–503
20. Wallraven, C., Caputo, B., Graf, A.: Recognition with local features: the kernel recipe. In: Proc. 9th International Conference on Computer Vision (ICCV). Volume 1. (2003) 257–264
21. Kadir, T., Brady, M.: Saliency, scale and image description. International Journal of Computer Vision **45** (2001) 83–105
22. Sanderson, C., Shan, T., Lovell, B.: Towards pose-invariant 2d face classification for surveillance. In: Third IEEE International Workshop on Analysis and Modelling of Faces and Gestures at ICCV2007. (2007)
23. Sim, T., Baker, S., Bsat, M.: The CMU pose, illumination, and expression database. IEEE. Trans. Pattern Analysis and Machine Intelligence **25** (2003) 1615–1618
24. Phillips, P., Moon, H., Rizvi, S., Rauss, P.: The FERET evaluation methodology for face-recognition algorithms. IEEE Trans. Pattern Analysis and Machine Intelligence **22** (2000) 1090–1104
25. Azman, A., Mustafah, Y.M., Bigdeli, A., Lovell, B.: Optimizing resources of an fpga-based smart camera architecture. In: Proc. of Digital Image Computing: Techniques and Applications. (2007) 600–606
26. Mustafah, Y.M., Shan, T., Azman, A.W., Bigdeli, A., Lovell, B.: Real-time face detection and tracking for high resolution smart camera system. In: Proc. of Digital Image Computing: Techniques and Applications. (2007) 387–393
27. Rodriguez, Y., Cardinaux, F., Bengio, S., Mariethoz, J.: Measuring the performance of face localization systems. Image and Vision Computing **24** (2006) 882–893