Regression Based Non-Frontal Face Synthesis for Improved Identity Verification

Yongkang Wong[1,2], Conrad Sanderson[1,2], and Brian C. Lovell[1,2]

[1] NICTA, PO Box 6020, St Lucia, QLD 4067, Australia *
[2] The University of Queensland, School of ITEE, QLD 4072, Australia

Abstract. We propose a low-complexity face synthesis technique which transforms a 2D frontal view image into views at specific poses, without recourse to computationally expensive 3D analysis or iterative fitting techniques that may fail to converge. The method first divides a given image into multiple overlapping blocks, followed by synthesising a non-frontal representation through applying a multivariate linear regression model on a low-dimensional representation of each block. To demonstrate one application of the proposed technique, we augment a frontal face verification system by incorporating multi-view reference (gallery) images synthesised from the frontal view. Experiments on the pose subset of the FERET database show considerable reductions in error rates, especially for large deviations from the frontal view.

1 Introduction

Face based identity inference subject to pose variations is a challenging problem, as previously studied and documented in FRVT test reports [1]. In certain applications the only reference (gallery) face images available are in one pose — e.g. frontal passport photos. Under typical surveillance conditions, CCTV cameras are unable to provide good quality frontal face images, largely due to the positioning of the cameras. In such situations an identity inference system based on frontal reference views will tend to have poor accuracy, unless extra processing is used to reduce the pose mismatch between the reference and acquired surveillance images.

The mismatch reduction can be accomplished through transforming the acquired surveillance images to be of the same pose as the reference image, or vice-versa. Recent face transformation methods include techniques based on Active Appearance Models (AAMs) [2, 3] and fitting a 2D image into a 3D morphable model [4, 5]. The AAM based synthesis approach requires an initialisation stage to label the important facial features (e.g. ~ 60 points). The initialisation can be done manually or automatically, where it may fail to converge. The morphable model based approach estimates the 3D shape and texture from a single image, with the fitting process taking about 4.5 minutes on a 2 GHz Pentium 4 workstation.

* NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy, as well as the Australian Research Council through the ICT Centre of Excellence program.

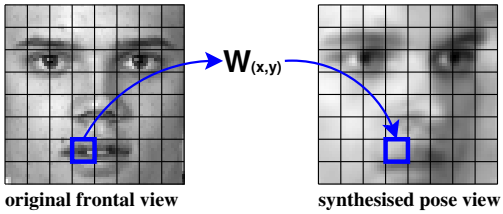

Fig. 1. Conceptual example of block-by-block image synthesis. The transformation matrix $\boldsymbol{W}_{(x,y)}$ is specific to location (x, y).

In this paper we propose a low-complexity, non-iterative face synthesis technique which transforms a given frontal view image into views at specific poses, without recourse to computationally expensive 3D analysis or fitting techniques that may fail to converge. The method first divides a given image into multiple overlapping blocks, followed by applying a multivariate linear regression model on a low-dimensional representation of each block. A conceptual example of this process is shown in Fig. 1. We demonstrate an application of the technique by augmenting a frontal face verification system with artificial multi-view reference (gallery) images synthesised from the frontal view.

We continue as follows. The details of the face synthesis method are described in Section 2. A preliminary comparison with AAM based image synthesis is given in Section 3. Identity verification experiments on the pose subset of the FERET database are given in Section 4. Conclusions and an outlook are presented in Section 5.

2 Face Synthesis with Multivariate Linear Regression

The proposed face synthesis method is a non-iterative process comprised of five steps: **(1)** block-based image analysis, **(2)** low-dimensional representation of each block, **(3)** transformation with multivariate linear regression, **(4)** block reconstruction, **(5)** block-based image synthesis. The steps are elaborated below, followed by a subsection which explores the effects of several parameters.

1. A given frontal image, $\boldsymbol{X}^{[\texttt{frontal}]}$, is analysed on an overlapping block-by-block basis, resulting in a set of 2D blocks, $\left\{\boldsymbol{b}^{[\texttt{frontal}]}_{(p,q)}, \boldsymbol{b}^{[\texttt{frontal}]}_{(r,s)}, \cdots\right\}$, where the subscript indicates the position of the block within the image. Based on preliminary experiments (using face images with a size of 64×64 pixels), each block has a size of 8×8 pixels.
2. To ameliorate dimensionality problems described later, each block $\boldsymbol{b}^{[\texttt{frontal}]}_{(x,y)}$ is represented by a vector of coefficients, $\boldsymbol{v}^{[\texttt{frontal}]}_{(x,y)}$, resulting from the 2D Discrete Cosine Transform (DCT) [6]. To achieve dimensionality reduction, only the lower order coefficients are retained (specifically, to reduce the dimensionality from 64 to 16, coefficients are taken from the top-left 4×4 sub-matrix of the 8×8 coefficient matrix).
3. Each vector obtained from the frontal view is then transformed to represent a non-frontal view Θ using:

$$\boldsymbol{v}^{[\Theta]}_{(x,y)} = \left[1 \ \left(\boldsymbol{v}^{[\texttt{frontal}]}_{(x,y)}\right)^T\right] \boldsymbol{W}^{[\Theta]}_{(x,y)} \tag{1}$$

where $\boldsymbol{W}^{[\Theta]}_{(x,y)}$ is a transformation matrix specific to view Θ and location (x, y).

Two sets of training vectors, obtained from frontal and non-frontal faces, are required to obtain $\boldsymbol{W}^{[\Theta]}_{(x,y)}$. In each set the vectors are also specific to location (x, y). Let us place the training frontal vectors into an extended matrix $\boldsymbol{A}$ and the training non-frontal vectors into matrix $\boldsymbol{B}$. If both matrices are constrained to have N number of vectors, we can define a linear regression model as follows:

$$\boldsymbol{B}_{(x,y)} = \boldsymbol{A}_{(x,y)} \quad \boldsymbol{W}^{[\Theta]}_{(x,y)} \tag{2}$$

$$\begin{bmatrix} \boldsymbol{b}_1^T \\ \boldsymbol{b}_2^T \\ \vdots \\ \boldsymbol{b}_N^T \end{bmatrix} = \begin{bmatrix} 1 \ \boldsymbol{a}_1^T \\ 1 \ \boldsymbol{a}_2^T \\ \vdots \\ 1 \ \boldsymbol{a}_N^T \end{bmatrix} \begin{bmatrix} w_{1,1} & \cdots & w_{1,D} \\ w_{2,1} & \cdots & w_{2,D} \\ \vdots & \vdots & \vdots \\ w_{D+1,1} & \cdots & w_{D+1,D} \end{bmatrix} \tag{3}$$

where D is the dimensionality of the vectors and $N > D + 1$. Under the sum-of-least-squares regression criterion, $\boldsymbol{W}^{[\Theta]}_{(x,y)}$ can be found using [7]:

$$\boldsymbol{W}^{[\Theta]}_{(x,y)} = \left(\boldsymbol{A}^T_{(x,y)}\boldsymbol{A}_{(x,y)}\right)^{-1}\boldsymbol{A}^T_{(x,y)}\boldsymbol{B}_{(x,y)} \tag{4}$$

Due to the constraint on N, the higher the dimensionality of the vectors, the more training faces are required. Given that there might be a limited number of such faces, or there might be memory constraints for solving Eqn. (4), it is preferable to keep the dimensionality low.

4. Each synthesised non-frontal vector $\boldsymbol{v}^{[\Theta]}_{(x,y)}$ is converted to a non-frontal block $\boldsymbol{b}^{[\Theta]}_{(x,y)}$ through an inverse 2D DCT. The omitted DCT coefficients are set to zero.

5. A synthesised image $\boldsymbol{X}^{[\Theta]}$ for non-frontal view Θ is constructed from blocks $\left\{\boldsymbol{b}^{[\Theta]}_{(p,q)}, \boldsymbol{b}^{[\Theta]}_{(r,s)}, \cdots\right\}$ through an averaging operation. An auxiliary matrix, $\boldsymbol{C}$, is used for keeping the count of pixels placed at each location. Elements of $\boldsymbol{X}^{[\Theta]}$ and $\boldsymbol{C}$ are first set to zero. A block $\boldsymbol{b}^{[\Theta]}_{(x,y)}$ is placed into $\boldsymbol{X}^{[\Theta]}$ at location (x, y) by adding to the elements already present in $\boldsymbol{X}^{[\Theta]}$. The corresponding elements of $\boldsymbol{C}$ are increased by one. This process is repeated until all the blocks have been placed. Finally, each element of $\boldsymbol{X}^{[\Theta]}$ is divided by the corresponding element in $\boldsymbol{C}$.

2.1 Effects of Vector Dimensionality and Degree of Block Overlap

In this section we show the effects of the degree of dimensionality reduction as well as the amount of block overlap. For evaluation we use frontal and non-frontal faces from subset b of the FERET dataset [8]. The subset has 200 persons in 9 views (frontal, $\pm 60^o$, $\pm 40^o$, $\pm 25^o$ and $\pm 15^o$). Each image was size normalised and cropped so that the eyes were at the same positions in all images. The resulting image size was 64×64 pixels. 100 randomly selected persons were used to train the transformation matrices for each pose angle. Frontal images from the remaining 100 persons were then fed to the proposed synthesis technique.

Examples of typical effects of the amount of block overlap are shown in Fig. 2, where a frontal face is transformed to a synthetic $+40^o$ view. 25% of the DCT

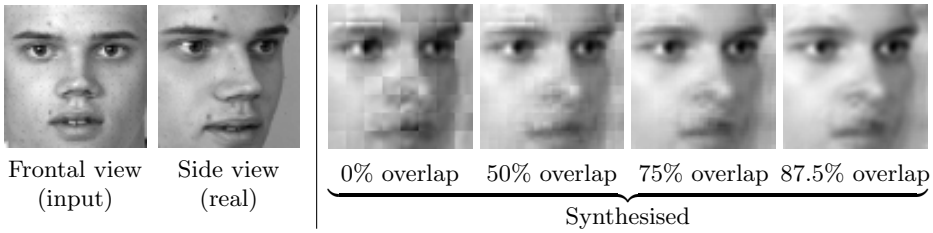

Fig. 2. Examples of non-frontal (side view) face synthesis with various degrees of block overlap. The images are synthesised from the frontal input image. The real side view, for the person depicted in the frontal image, is never used by the system. 25% of the DCT coefficients were retained.

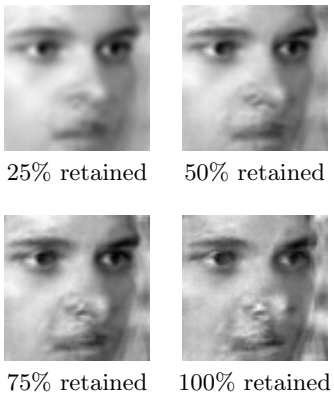

Fig. 3. Examples of face synthesis with various amounts of retained DCT coefficients. There is a maximum of 64 coefficients per block. The block overlap was set to 87.5%.

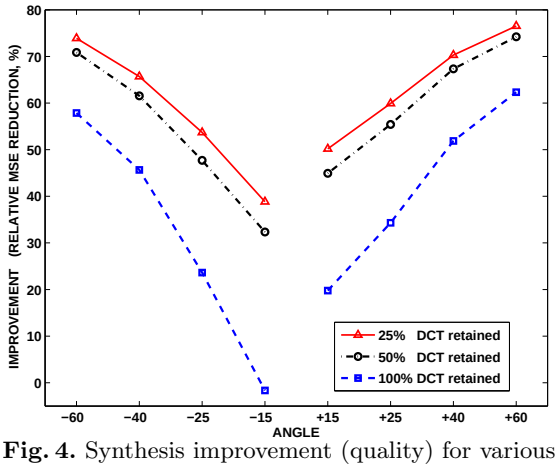

Fig. 4. Synthesis improvement (quality) for various amounts of retained DCT coefficients. A baseline mean square error (MSE) was obtained by comparing each frontal image with the corresponding real side image. A secondary MSE was then obtained by comparing real side images with corresponding synthesised side images. The improvement is how much smaller the secondary MSE is compared to the baseline (i.e. relative reduction).

coefficients were retained (i.e. 16 out of 64) for each block. The block overlap varied from 0% to 87.5%. A 50% overlap indicates that each 8×8 pixel block overlapped its neighbours by 4 pixels.

The quality of the synthesised images improves remarkably as the overlap increases. This can be attributed to the considerable increase in the number of transformation matrices (from 64 in the 0% overlap case to 3249 in the 87.5% case), leading to the overall image transformation being much more detailed. Furthermore, mistakes in the synthesis of individual blocks tend to be reduced through the averaging process described in step 5 of the algorithm.

The effect of the degree of dimensionality reduction is shown qualitatively in Fig. 3 and quantitatively in Fig. 4. The optimal amount of retained coefficients

appears to be around 25%, which has the effect of removing high frequencies. Using more than 25% of the coefficients results in poorer quality images — this can be attributed to the dimensionality being too high in relation to the available number of training examples.

Fig. 4 shows the relative reduction in pixel based mean square error (MSE) for faces not used during training, which can be used to quantitatively gauge the improvement in image quality. Here a baseline MSE was obtained by comparing the frontal image with the real side image for each person. A secondary MSE was then obtained by comparing real side images with corresponding synthesised side images. We define the relative reduction in MSE as how much smaller the secondary MSE is compared to the baseline MSE. The quantitative results presented in Fig. 4 confirm the qualitative results shown in Fig. 3, where using 25% of the DCT coefficients is better than using 100%. It can also be observed that the best improvement occurs for high pose angles ($\pm 60^o$).

3 Comparison with AAM Based Image Synthesis

The Active Appearance Model (AAM) based image synthesis approach [2, 3] requires a initialisation stage to label important facial features (e.g. ~ 60 points). The initialisation can be done manually or automatically, where it can fail to converge. This is in contrast the proposed technique, which only requires the positions of the eyes — this can be obtained automatically and relatively robustly by modern face localisation (detection) algorithms [9].

Fig. 5 shows a preliminary and qualitative comparison of image synthesis using the proposed regression based technique and the AAM based method described in [3]. We follow the setup described in [3], which is similar to that of Section 2.1. The main difference is that side view images are transformed to a frontal view rather than vice-versa.

On first sight, the synthesised image produced by the regression method appears to be less defined than the AAM generated image. However, facial features such as the eyes and nose are actually better approximated when compared directly with the real frontal view image. Specifically, the eyes synthesised by the regression method retain their overall shape (outer edges are pointing downwards) and the nose remains thin. This is not the case for the AAM-based technique, where the eyes lose much of their original shape and the nose is considerably wider.

Non-frontal input image | Real frontal view | AAM based synthesis of frontal view | Regression based synthesis of frontal view

Fig. 5. Comparison of frontal view synthesis from non-frontal input view. Results for the AAM based approach taken from [3].

4 Face Verification with Synthesised Non-Frontal Images

This section shows an application of the proposed face synthesis technique. Synthesised faces are used to build a multi-angle model to address the pose mismatch problem described in Section 1. As the baseline we use the PCA/GMM face verification system described in [10], which is easily extendable to multi-angle models while remaining in a probabilistic framework.

The first step is Principal Component Analysis (PCA) based feature extraction. Briefly, a given face image is represented by a matrix containing grey pixel values. The matrix is converted to a face vector, $\boldsymbol{f}$, by column concatenation. A D-dimensional feature vector, $\boldsymbol{x}$, is obtained by:

$$\boldsymbol{x} = \boldsymbol{U}^T \left(\boldsymbol{f} - \boldsymbol{f}_\mu \right) \tag{5}$$

where $\boldsymbol{U}$ contains D eigenvectors (corresponding to the D largest eigenvalues) of the training data covariance matrix and $\boldsymbol{f}_\mu$ is the mean of training face vectors [11].

In the verification task we wish to find out whether a given biometric sample belongs to the true claimant or an impostor. A claim for the identity of person C is classified as coming from the that person (i.e. the true claimant) when

$$\frac{p(\boldsymbol{x}|\lambda_C)}{p(\boldsymbol{x}|\lambda_{\texttt{impostor}})} > t \tag{6}$$

and as coming from an impostor otherwise. Here t is a decision threshold, λ_C is the model for person C and $\lambda_{\texttt{impostor}}$ is the approximate impostor model. The distribution of feature vectors is described by a Gaussian Mixture Model (GMM):

$$p\left(\boldsymbol{x}|\lambda\right) = \sum\nolimits_{g=1}^{N_G} w_g \mathcal{N}\left(\boldsymbol{x}|\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g\right) \tag{7}$$

where $\lambda = \left\{w_g, \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g\right\}_{g=1}^{N_G}$ and $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is a D-dimensional Gaussian function with mean $\boldsymbol{\mu}$ and diagonal covariance matrix $\boldsymbol{\Sigma}$. N_G is the number of Gaussians and w_g is the weight for Gaussian g (with constraints $\sum_{g=1}^{N_G} w_g = 1$) [11, 12]. Due to the relatively small amount of training data for each person (i.e. one frontal image), a common covariance matrix is used for all Gaussians and all models.

Frontal face models, for each person enrolled in the system, are comprised of one Gaussian. The Gaussian's mean is equal to the PCA-derived feature vector obtained from the frontal face. In a similar manner, the approximate impostor model is comprised of 32 Gaussians, where the Gaussian means are taken to be equal to the PCA-derived feature vectors of 32 randomly selected persons. The weights are all equal.

4.1 Synthesised Multi-Angle Models

In order for the system to automatically handle non-frontal views, each frontal face model is extended by concatenating it with models generated from synthesised non-frontal views. Formally, an extended (or multi-angle) model for person C is created using:

$$\lambda_C^{[\texttt{extended}]} = \lambda_C^{[\texttt{frontal}]} \sqcup \lambda_C^{[+60^\circ]} \sqcup \lambda_C^{[+40^\circ]} \cdots \sqcup \lambda_C^{[-40^\circ]} \sqcup \lambda_C^{[-60^\circ]} \tag{8}$$

$$= \sqcup_{i \in \Phi} \lambda_C^{[i]} \tag{9}$$

where $\lambda_C^{[\texttt{frontal}]}$ represents the frontal model, Φ is a set of angles, e.g. $\Phi = \{\pm 15^o, \pm 25^o, \pm 40^o, \pm 60^o\}$, and $\sqcup$ is an operator for joining GMM parameter sets, defined as follows. Let us suppose we have two GMM parameter sets, $\lambda^{[\texttt{a}]}$ and $\lambda^{[\texttt{b}]}$, comprised of parameters for $N_G^{[\texttt{a}]}$ and $N_G^{[\texttt{b}]}$ Gaussians, respectively. The $\sqcup$ operator is defined as follows:

$$\lambda^{[\texttt{joined}]} = \lambda^{[\texttt{a}]} \sqcup \lambda^{[\texttt{b}]} = \left\{\alpha w_g^{[\texttt{a}]}, \boldsymbol{\mu}_g^{[\texttt{a}]}, \boldsymbol{\Sigma}_g^{[\texttt{a}]}\right\}_{g=1}^{N_G^{[\texttt{a}]}} \cup \left\{\beta w_g^{[\texttt{b}]}, \boldsymbol{\mu}_g^{[\texttt{b}]}, \boldsymbol{\Sigma}_g^{[\texttt{b}]}\right\}_{g=1}^{N_G^{[\texttt{b}]}} \quad (10)$$

where $\alpha = N_G^{[\texttt{a}]} / \left(N_G^{[\texttt{a}]} + N_G^{[\texttt{b}]}\right)$ and $\beta = 1 - \alpha$.

4.2 Experiments and Results

The experiments were done[3] using data from two subsets of the FERET dataset. Subset *f* was used to train the PCA based feature extractor (i.e. $\boldsymbol{U}$ and $\boldsymbol{f}_\mu$) and to obtain the common covariance matrix for the GMMs. Faces for $\lambda_{\texttt{impostor}}$ were also selected from this subset. Subset *b* was randomly split into three disjoint groups: group A, group B and an impostor group. Group A had 100 persons and was used to train the transformation matrices for each pose view (i.e. from frontal to dedicated pose). Group B had 80 persons and the remaining 20 persons were placed in the impostor group. The latter two groups were used in verification tests, which were comprised of 80 true claims and $20 \times 80 = 1600$ impostor attacks per view angle.

Two systems were evaluated: (1) frontal models and (2) synthesised multi-angle models. In the latter case each person's model had 9 Gaussians, with each Gaussian representing a particular view (i.e. the original frontal and synthesised $\pm 15^o$, $\pm 25^o$, $\pm 40^o$, $\pm 60^o$ views). For each angle the results are reported in terms of the Equal Error Rate (EER) [12].

The results, shown in Fig. 6, demonstrate considerable error reductions across all pose views. The largest improvement in performance occurs for large deviations from the frontal view ($\pm 60^o$), where the errors are remarkably reduced by an absolute difference of about 15 percentage points, or a relative difference of about 30%.

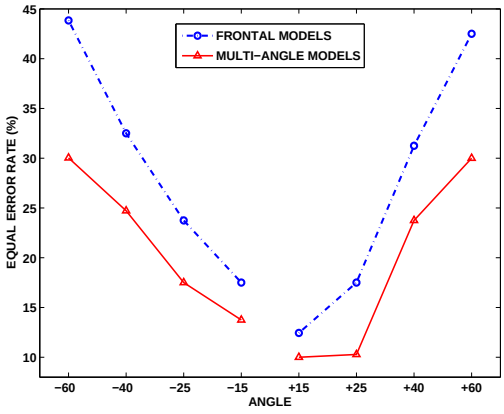

Fig. 6. Verification error rates using frontal and multi-angle models. The latter use synthesised non-frontal faces.

[3] The experiments were performed with the aid of the Armadillo C++ linear algebra library, available from http://arma.sourceforge.net

5 Conclusions and Outlook

In this paper we proposed a low-complexity face synthesis technique which transforms a 2D frontal view image into views at specific poses, without recourse to computationally expensive 3D analysis or iterative fitting techniques that may fail to converge (as used by Active Appearance Models [3]). The proposed method first divides a given image into multiple overlapping blocks, followed by synthesising a non-frontal representation through applying a multivariate linear regression model on a low-dimensional representation of each block.

The proposed synthesis method is relatively straightforward, with low computational requirements for both training and image synthesis. Using 100 persons for training, learning the regression matrices took about 3 seconds on a 2 GHz Intel Core 2 processor. Synthesis of 1 test image took less than 0.04 seconds (C++ implementation, gcc 4.1.2, Linux 2.6.26).

To demonstrate one application of the proposed technique, we augmented a frontal face verification system by incorporating multi-view reference (gallery) images synthesised from the frontal view. Experiments on the pose subset of the FERET database indicate considerable reductions in error rates, especially for large deviations from the frontal view.

Improvements in synthesis quality could be obtained through a more precise low-dimensional representation of each block. For example, rather than using the generic 2D DCT, a position dependent local PCA could be employed, which may have the additional advantage of further reducing the dimensionality.

References

1. Phillips, P., Grother, P., Micheals, R., Blackburn, D., Tabassi, E., Bone, M.: Face recognition vendor test 2002. In: Proc. Int. Workshop on Analysis and Modeling of Faces and Gestures (AMFG). (2003) 44
2. Cootes, T., Walker, K., Taylor, C.: View-based active appearance models. In: Proc. 4th IEEE Int. Conf. Automatic Face and Gesture Recognition. (2000) 227–232
3. Shan, T., Lovell, B., Chen, S.: Face recognition robust to head pose from one sample image. In: Proc. ICPR. Volume 1. (2006) 515–518
4. Blanz, V., Vetter, T.: Face recognition based on fitting a 3D morphable model. IEEE Trans. Pattern Analysis and Machine Intelligence **25**(9) (2003) 1063–1074
5. 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
6. Gonzales, R., Woods, R.: Digital Image Processing. 3 edn. Prentice Hall (2007)
7. Rice, J.: Mathematical Statistics and Data Analysis. 2 edn. Duxbury Press (1995)
8. Phillips, P., Moon, H., Rizvi, S., Rauss, P.: The FERET evaluation methodology for face-recognition algorithms. IEEE Trans. Pattern Analysis and Machine Intelligence **22**(10) (2000) 1090–1104
9. Viola, P., Jones, M.J.: Robust real-time face detection. International Journal of Computer Vision **57**(2) (2004) 137–154
10. Rodriguez, Y., Cardinaux, F., Bengio, S., Mariethoz, J.: Measuring the performance of face localization systems. Image and Vision Comput. **24** (2006) 882–893
11. Bishop, C.: Pattern Recognition and Machine Learning. Springer (2006)
12. Sanderson, C.: Biometric Person Recognition: Face, Speech and Fusion. VDM-Verlag (2008)