Learning-based Face Synthesis for Pose-Robust Recognition from Single Image

Akshay Asthana[1]
aasthana@rsise.anu.edu.au
Conrad Sanderson[2]
conradsand@ieee.org
Tom Gedeon[1]
tom.gedeon@anu.edu.au
Roland Goecke[1,3]
roland.goecke@ieee.org

[1] RSISE and SoCS, CECS, Australian National University, Australia
[2] NICTA and University of Queensland, Australia
[3] Faculty of Information Sciences and Engineering, University of Canberra, Australia

Abstract

Face recognition in real-world conditions requires the ability to deal with a number of conditions, such as variations in pose, illumination and expression. In this paper, we focus on variations in head pose and use a computationally efficient regression-based approach for synthesising face images in different poses, which are used to extend the face recognition training set. In this data-driven approach, the correspondences between facial landmark points in frontal and non-frontal views are learnt offline from manually annotated training data via Gaussian Process Regression. We then use this learner to synthesise non-frontal face images from any unseen frontal image. To demonstrate the utility of this approach, two frontal face recognition systems (the commonly used PCA and the recent Multi-Region Histograms) are augmented with synthesised non-frontal views for each person. This synthesis and augmentation approach is experimentally validated on the FERET dataset, showing a considerable improvement in recognition rates for $\pm 40^\circ$ and $\pm 60^\circ$ views, while maintaining high recognition rates for $\pm 15^\circ$ and $\pm 25^\circ$ views.

1 Introduction

A major challenge for automatic face based identity inference is the shear magnitude of uncontrolled factors than can result in a considerable change of shape and appearance, such as expression, illumination and pose. The problem is complicated further when only one image per person is available for training. Previous approaches for addressing the specific problem of pose variations include 3D deformable models (3DMM) [4, 5], parts-based probabilistic approaches [7, 15] and tied factor analysis [12].

The system presented in [4] first fits a 3D model to a given probe image followed by computing a set of model coefficients which are then used for finding the best match in a database. In [5], another 3DMM based approach was presented, which extracted 3D shape, texture and a set of 3D scene parameters. A synthetic frontal view of the probe image was then generated, followed by employing a view-based recognition algorithm to compare the

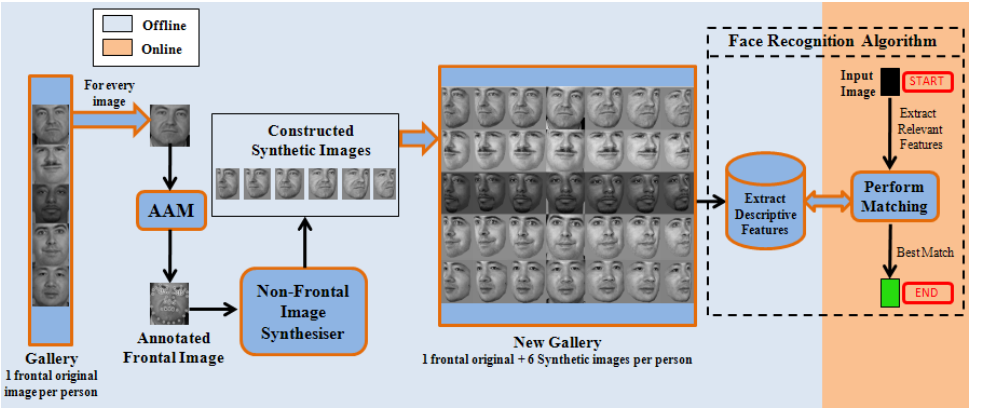


Figure 1: Conceptual example of the overall synthesis and augmentation approach for pose-robust recognition.

synthetic frontal image with a database of frontal images. There are several drawbacks to the above 3DMM approaches, which may limit their applicability: (i) automatically achieving accurate fitting, across different poses, is a very difficult task [7], (ii) heavy reliance on precise initialisation of the fitting procedure, (iii) the fitting procedure is computationally expensive.

In [15], the problem was approached by extending the training dataset with artificially synthesised statistical models of face patches (rather than images) for non-frontal views. In [7], a patch-based approach was proposed that models the joint probability of gallery and probe images across different poses. In [12], a purely machine learning based approach for pose-robust recognition was presented, which creates a generative model that best represents the changes in the face image induced by pose variation.

In this paper, we present an approach that uses a data driven and computationally efficient 2D technique for the synthesis of high quality non-frontal faces from a single frontal input face. The technique is based on Active Appearance Models (AAM) [6] and Gaussian Process Regression (GPR) [13]. The synthesised faces can then be used for extending the training set for each person, enabling existing 2D face recognition algorithms to improve their performance when dealing with pose variations. A conceptual overview of the proposed synthesis and augmentation approach is given in Figure 1.

We continue the paper as follows. The regression based method for construction of synthetic non-frontal images is presented in Section 2. Section 3 details the experiments and discusses the results. Section 4 provides the conclusions and an outlook on future work.

2 Constructing Synthetic Non-Frontal Images

Given a gallery containing one *frontal* image per person, the goal is to generate high quality synthetic images at various poses using original shape and texture information. We use a regression-based approach to generate synthetic images at various poses (e.g. covering the range of $-67.5°$ to $+67.5°$ on the x-axis). For every image in the gallery, a set of landmark points is extracted by using AAMs [6] (Section 2.1). These landmark points are used by the *Non-Frontal Image Synthesiser* (Section 2.2) to generate synthetic images at various poses.

2.1 Automatic Extraction of Landmark Points From Frontal Images

We use offline AAM fitting in order to obtain the locations of landmark points in a given frontal face image. For training the AAM that is used to perform fitting on unseen faces, we manually annotated frontal images for a total of 400 people from the CMU PIE [18], Face Pointing [9] and FERET [11] databases. Our proposed framework uses the *Simultaneous Inverse Compositional* (SIC) method [2] to perform fitting. SIC is one of the most powerful generative fitting method in which the update model is generated directly from background-free components (i.e. the mean appearance and their modes of variation) and, hence, has no specialisation to any particular background. Since the construction of synthetic images is an offline process, finding accurate landmark point locations takes priority over speed, making SIC best suited for this purpose. Moreover, if the initialisation[1] of the AAM fitting procedure is close to the optimum, the problem of performing AAM fitting on the frontal images is relatively simpler as compared to AAM fitting on images with varying viewpoints.

2.2 Non-Frontal Image Synthesiser

Once the locations of the landmark points from the gallery images have been extracted, we use a regression-based approach [1] to generate synthetic images at various poses. We first learn the correspondence between the landmark points of the set of frontal images exhibiting arbitrary facial expressions and their corresponding non-frontal images at arbitrary pose. Once this learner has been trained, a synthetic image of any other *unseen* face (at a pose for which the learner has been trained) can be generated by using a regression method such as GPR [13]. Regression is used to predict the locations of the new landmark points, followed by warping of the texture from the original frontal image. The detailed step-by-step procedure is as follows.

We start by learning the correspondence between the landmark points of the set of frontal images and their corresponding non-frontal images at an arbitrary, but known, pose (e.g. between the frontal and 45° pose). We extract the *Normalisation* (**N**), *Centroid* (**C**) and *Point* (**P**) vectors from every frontal and its corresponding non-frontal image.

Normalisation Vector (**N**) - 1D vector containing the normalisation distances. Horizontal normalisation distance (N_h) is the horizontal distance between the eye corners. Vertical normalisation distance (N_v) is the vertical distance between the eye corners and the nose tip (also the reference point in the normalised frame).

$$\mathbf{N} = [N_h; N_v]^T \tag{1}$$

Centroid Vector (**C**) - 1D vector containing the location of the centroids of six individual facial features (left and right eyebrows, left and right eyes, nose and mouth) in the normalised frame. For this, a dictionary of landmark points is created that contains the information about which of the six facial features each of the n landmark points represents. Hence, if the number of landmark points c represents a facial feature, e.g. the mouth, then the centroid $(\vec{x}, \vec{y})$ of this facial feature is computed as $\vec{x} = \left(\frac{\sum_{i=1}^{c} x_i}{c}\right)$

[1] For the experiments presented in this paper, we use faceAPI (Seeing Machines) to detect the location of the face, which is required to initialise the AAM fitting procedure. However, we wish to highlight that any suitably accurate face detector will suffice and that our approach is not dependent on a particular face detector.

Algorithm 1: Learning the Regression Model

Require: $\mathbf{N}$, $\mathbf{C}$ and $\mathbf{P}$ from m frontal and non-frontal images.

1 Extract $\mathbf{N^f}$ from frontal and $\mathbf{N^n}$ from non-frontal image

$$\mathbf{N^f} = [N_h; N_v]^T \qquad \mathbf{N^n} = [N'_h; N'_v]^T$$

2 Extract $\mathbf{C^f}$ from frontal and $\mathbf{C^n}$ from non-frontal image

$$\mathbf{C^f} = [\vec{x_1}; \vec{y_1}; \ldots; \vec{x_6}; \vec{y_6}]^T \qquad \mathbf{C^n} = [\vec{x'_1}; \vec{y'_1}; \ldots; \vec{x'_6}; \vec{y'_6}]^T$$

3 Extract $\mathbf{P^f}$ from frontal and $\mathbf{P^n}$ from non-frontal image

$$\mathbf{P^f} = [x_1; y_1; \ldots; x_n; y_n]^T \qquad \mathbf{P^n} = [x'_1; y'_1; \ldots; x'_n; y'_n]^T$$

4 There are $m \times 3$ pairs of normalisation, centroid and point vectors, respectively, with each pair representing a frontal and its corresponding non-frontal image.

5 Construct 3 different training sets

$$\mathscr{T}_{\mathscr{N}} = \left\{ (\mathbf{N_i^f}, \mathbf{N_i^n}) \mid i \in (1, 2, \ldots, m) \right\}$$
$$\mathscr{T}_{\mathscr{C}} = \left\{ (\mathbf{C_i^f}, \mathbf{C_i^n}) \mid i \in (1, 2, \ldots, m) \right\}$$
$$\mathscr{T}_{\mathscr{P}} = \left\{ (\mathbf{P_i^f}, \mathbf{P_i^n}) \mid i \in (1, 2, \ldots, m) \right\}$$

6 Use regression to learn 3 different sets of regression models $\mathscr{R}_{\mathscr{N}}$, $\mathscr{R}_{\mathscr{C}}$, $\mathscr{R}_{\mathscr{P}}$ for predicting the normalisation, centroid and point vector respectively, where

$$\mathscr{R}_{\mathscr{N}} = \left\{ R_{N_h}\ ,\ R_{N_v} \right\}$$
$$\mathscr{R}_{\mathscr{C}} = \left\{ R_{C_i} \mid i \in (1, 2, \ldots, 12) \right\}$$
$$\mathscr{R}_{\mathscr{P}} = \left\{ R_{P_i} \mid i \in (1, 2, \ldots, 2n) \right\}$$

Here, R represents the regression model learned over a particular training set. GPR [13] is used for the experiments in this paper.

and $\vec{y} = \left(\frac{\sum_{i=1}^{c} y_i}{c}\right)$, where (x_i, y_i) is the location of each landmark point representing this facial feature in the normalised frame.

$$\mathbf{C} = [\vec{x_1}; \vec{y_1}; \ldots; \vec{x_6}; \vec{y_6}]^T \tag{2}$$

Point Vector ($\mathbf{P}$) - 1D vector containing the location of each of the n landmark points in the normalised frame.

$$\mathbf{P} = [x_1; y_1; \ldots; x_n; y_n]^T \tag{3}$$

Once $\mathbf{N}$, $\mathbf{C}$ and $\mathbf{P}$ have been extracted, we learn the regression model for the construction of synthetic images as explained in Algorithm 1.

With the location of the landmark points, extracted by the AAM (see Section 2.1) from every gallery (frontal) image, at hand, the landmark points are normalised to remove any minor variation in the *roll* present in the gallery images (e.g. samples in rows 2-5 of Figure 3). This is simply done by applying a Euclidean transformation to the landmark points, taking the landmark points representing the corner of the eyes and tip of the nose as reference points. These normalised landmark points can now be used to generate the synthetic images by predicting the new landmark locations via the above learnt regression model and warping the texture from the frontal gallery image. We use *Piecewise Affine Warping* (PAW) [5, 8] for warping the texture from the frontal images to the new landmark locations. Next, all invalid pixel locations inside the convex hull of the canonical shape are filled by their nearest-neighbours (NN) and the background pixels are filled with the mean value of texture

Algorithm 2: Constructing the Synthetic Images

Require: Landmark points for the frontal gallery image, *Img*, from *AAM*.

1 Normalise *Img* to remove the variation in *roll*.
2 Extract $\mathbf{N}_{test}$, $\mathbf{C}_{test}$, $\mathbf{P}_{test}$.
3 Use $\mathscr{R}_{\mathscr{N}}$, $\mathscr{R}_{\mathscr{C}}$ and $\mathscr{R}_{\mathscr{P}}$ (Algorithm 1) to predict $\mathbf{N}'_{test}$, $\mathbf{C}'_{test}$ and $\mathbf{P}'_{test}$.
4 $\mathbf{P}'_{test}$ contains the new location of each of the n landmark points w.r.t. the origin in the normalised frame. Arrange the landmark points in the normalised frame accordingly.
5 $\mathbf{C}'_{test}$ contains the new location of six centroids, representing six individual facial features, w.r.t. the origin in the normalised frame. Arrange these centroids in the normalised frame accordingly.
6 Transform the group of landmark points representing each of the six individual facial features (arranged in Step 4) to the new centroid locations drawn in Step 5.
7 De-normalise the constructed shape using the $\mathbf{N}'_{test}$.
8 **for** $i = 1$ *to* n **do**
9 $\quad x_i^{real} = x_i^{'}.(N_h^{'})$ where . denotes multiplication
10 $\quad y_i^{real} = y_i^{'}.(N_v^{'})$
11 Warp the texture from *Img* to the new locations using PAW.
12 Fill invalid pixel locations with NN and background pixels with mean texture.

to complete the synthetic image (the mean value is used to reduce the sharp transition from face edges to the background, in order to obtain a degree of resemblance with real non-frontal faces present in the FERET dataset). The step-by-step procedure is given in Algorithm 2. Figure 2 shows the construction steps for a sample image at a pose of 45°.

Figure 3 shows synthetic images constructed for a subset of the gallery images. The high quality of the constructed synthetic images is maintained throughout by virtue of the non-linear predictor utilised by the regression-based learning approach to predict the new locations for the landmark points.

3 Experiments and Discussion

We used the CMU PIE database [18] for training the regression model. A total of 970 images across 50 people (35 males and 15 females), with a face cropped area of approximately 140×150 pixels, were manually annotated with 69 landmarks each. There were 100 images each for the frontal pose, $-22.5°$, $\pm 45°$ and $\pm 67.5°$, and 170 images for $+22.5°$, exhibiting various facial expressions. Six learners were trained (see Section 2.2), one each for predicting the new landmark locations.

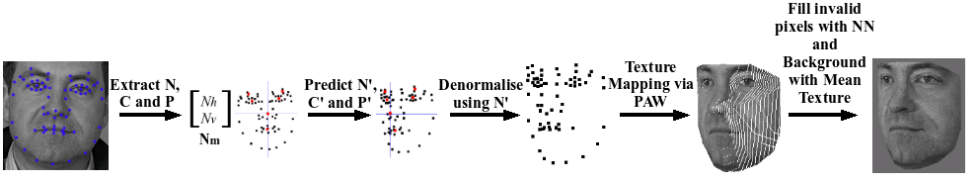


Figure 2: Face synthesis at a pose of 45° from a frontal view.

Figure 3: Synthetic images generated from sample gallery (frontal) images.

We used the *b* subset of the FERET database [11] to evaluate the usefulness of the synthetic images in a face recognition scenario. The *b* subset contains 1800 images of 200 unique subjects. Specifically, each person has 9 pose views: *ba* (frontal), *bb* (+60°), *bc* (+40°), *bd* (+25°), *be* (+15°), *bf* (−15°), *bg* (−25°), *bh* (−40°) and *bi* (−60°). The 200 frontal images were used as the gallery images and 6 synthetic images per subject were constructed and added to the gallery. The synthetic images were generated for non-frontal views at the same angles as present in the PIE database. The images from the remaining views (i.e. *bb* to *bi*) were used for testing (i.e. as probes). A conceptual example of the overall synthesis and augmentation approach was given in Figure 1.

The advantage of using two different databases (i.e. the PIE database for learning the regression model and the FERET database for evaluating the performance of the synthetic images) is two-fold. Firstly, constructing the synthetic images from the FERET database using the regression model learnt on the PIE database demonstrates the high generalisability of the construction approach. Secondly, it allows testing of the robustness of the face recognition algorithms for minor pose variations and their suitability to follow the approach presented in this paper. This is particularly important from a real-world application point of view. As stated earlier, the new gallery consists of 7 poses, if we consider the out-of-plane rotation along the x-axis. However in a real-world scenario, it is not necessarily the case that

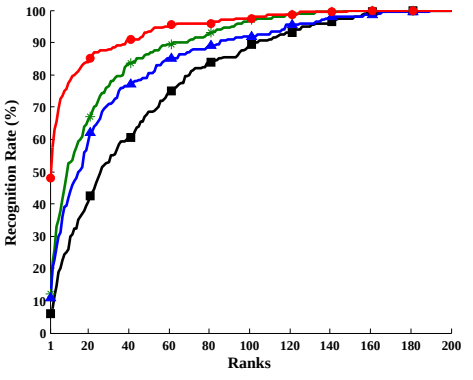
the pose of the probe image will be identical to one of the poses. In this case, one would have to rely on the robustness of the face recognition algorithm w.r.t. minor pose variations. The experiments presented in this paper simulate this condition throughout by using two independent databases with minor pose differences between the gallery and probe images. The 7 poses in the gallery are separated by $\Delta = 22.5^\circ$, starting from -67.5° to $+67.5^\circ$, while the probe images from the *b* subset of the FERET database have poses of $\pm 15^\circ$, $\pm 25^\circ$, $\pm 40^\circ$ and $\pm 60^\circ$.

To evaluate the utility of the synthetic images in a face recognition scenario, we used a baseline PCA-based recogniser [3] and the recently proposed *Multi-Region Histograms* (MRH) approach [14]. The MRH approach is motivated by the concept of 'visual words' used in image categorisation [10], which can be briefly described as follows. A given face is divided into several fixed and adjacent regions that are further divided into small overlapping patches. Each patch is represented by a low-dimensional feature vector, followed by representing each feature vector as a high-dimensional probabilistic histogram. Each entry in the histogram reflects how well a particular feature vector represents each 'visual word', where the dictionary of visual words is in effect a set of prototype feature vectors. For each region, the histograms of the underlying patches are then averaged. Two faces are compared through an L_1-norm based distance between corresponding histograms.

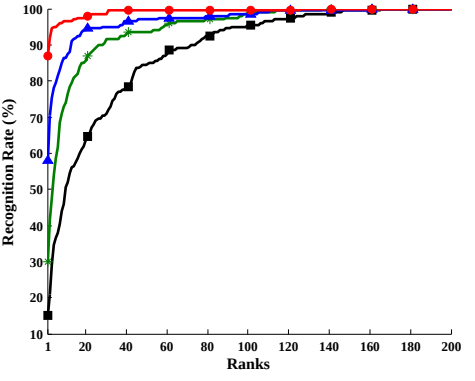
For the PCA-based system, each face was represented by a 128-dimensional feature vector (based on preliminary experiments). The MRH-based face recognition system follows the 3×3 region configuration of [14], with 1024-dimensional histograms for each region. For training, 1000 images from the *fa* subset of FERET (frontal images) were used for constructing the PCA-based dimensionality reduction matrix and the visual dictionary for the MRH approach.

Figure 4 shows the overall performance of the MRH and PCA-based approaches for each pose variation with and without the use of synthesised faces. Figures 4(a)-4(d) show the cumulative recognition rates for views at -60°, -40°, $+40^\circ$ and $+60^\circ$, respectively. Rank-1 recognition rates, shown in Figure 4(e), suggest an overall increase in the performance for both MRH and PCA-based systems (*See the supplementary material for the full table of results*). For the MRH-based recogniser, a significant increase in the recognition rate is obtained for the pose variation of $\pm 40^\circ$ and $\pm 60^\circ$. For other pose variations, a high recognition accuracy is either maintained or increased (e.g. $+25^\circ$). For the PCA-based recogniser the recognition rates are noticeably increased through the use of synthetic faces. However, the overall performance is still poorer when compared to the MRH-based recogniser.

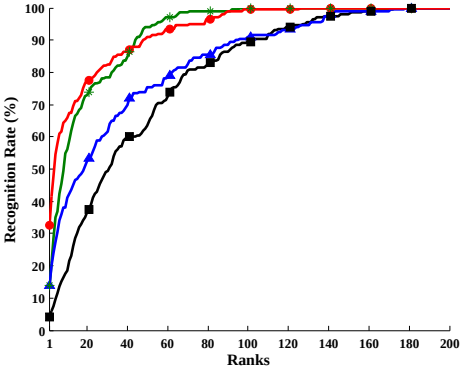
It should also be noted that the increase in the overall accuracy for the MRH-based approach is relatively larger than that for the PCA-based recogniser. This is due to the minor pose variation between the probe images and the pose of the faces in the extended training set, as stated earlier. The MRH-based approach appears to handle minor pose variations well, whereas the baseline PCA approach appears to be very sensitive to out-of-plane rotations. This is further compounded by the fact that the effect of occlusion increases at the extreme poses closest to profile images. Therefore, the amount of change in the appearance induced by the pose variations (say of $5-10^\circ$) at extreme poses, such as $\pm 60^\circ$, is relatively more than that induced by the same amount of variation at the poses close to frontal, such as $\pm 15^\circ$.



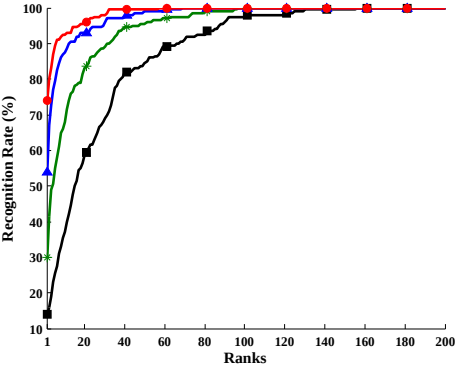
(a) Cumulative recognition rates for $-60°$



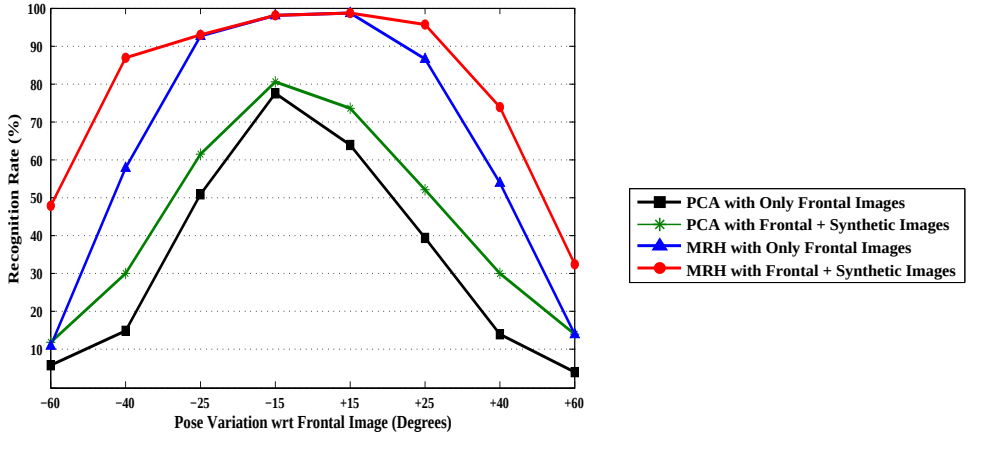
(b) Cumulative recognition rates for $-40°$



(c) Cumulative recognition rates for $+60°$



(d) Cumulative recognition rates for $+40°$



(e) Rank-1 recognition rates

Figure 4: Comparison of recognition rates obtained by PCA and MRH face recognition systems on the *b* subset of FERET database, with and without the use of synthesised faces.

4 Conclusion

In this paper, we presented a framework for generating high quality synthetic face images at various poses from a single frontal image, based on Active Appearance Models and Gaussian Process Regression, that can be used to increase the robustness of any 2D face recognition algorithm to variations in head pose. To objectively verify the quality of the synthesised images and to demonstrate an application of the proposed framework, we augmented two frontal face recognition systems by extending the training set for each person with synthesised non-frontal faces. Experiments on the FERET dataset show that the robustness of a baseline PCA approach as well as the recent Multi-Region Histograms method can be considerably increased when dealing with faces at $\pm 40^\circ$ and $\pm 60^\circ$ views, while maintaining high recognition rates for $\pm 15^\circ$ and $\pm 25^\circ$ views. Moreover, the approach works in 2D and is entirely data-driven, computationally inexpensive and can easily be combined with other face recognition algorithms than the ones shown here without requiring any major modification.

We are currently working on extending the approach into the 3D domain to enable better handling of occlusions as well as the construction of synthetic faces at extreme poses (such as profile views).

5 Acknowledgements

The authors would like to thank Jason Saragih, Carnegie Mellon University for the use of the DeMoLib software library [16, 17]. NICTA is funded by the Australian Government via the Department of Broadband, Communications and the Digital Economy, as well as the Australian Research Council through the ICT Centre of Excellence program. The work presented in this paper was in part supported by the ARC grant TS0669874.

References

[1] A. Asthana, R. Goecke, N. Quadrianto, and T. Gedeon. Learning Based Automatic Face Annotation for Arbitrary Poses and Expressions from Frontal Images Only. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2009)*, pages 1635–1642, June 2009.

[2] S. Baker, R. Gross, and I. Matthews. Lucas-Kanade 20 years on: A unifying framework - Part 3. Technical report, Robotics Institute, Carnegie Mellon University, Pittsbugh, USA, 2003.

[3] C.M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[4] V. Blanz and T. Vetter. Face Recognition Based on Fitting a 3D Morphable Model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(9):1063–1074, September 2003.

[5] V. Blanz, P. Grother, P.J. Phillips, and T. Vetter. Face Recognition Based on Frontal Views Generated from Non-Frontal Images. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2005)*, pages 454–461, 2005.

[6] G. Edwards, C.J. Taylor, and T.F. Cootes. Interpreting Face Images Using Active Appearance Models. In *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition FG'98*, pages 300–305. IEEE, April 1998.

[7] S. Lucey and T. Chen. A Viewpoint Invariant, Sparsely Registered, Patch Based, Face Verifier. *International Journal Computer Vision (IJCV)*, 80(1):58–71, December 2007.

[8] I. Matthews and S. Baker. Active Appearance Models Revisited. *International Journal of Computer Vision*, 60(2):135–164, November 2004.

[9] J. L. Crowley N. Gourier, D. Hall. Estimating Face Orientation from Robust Detection of Salient Facial Features. In *Proc. of Pointing 2004, ICPR, Int. Workshop on Visual Observation of Deictic Gestures, Cambridge, UK*, 2004.

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

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

[12] S.J.D. Prince, J. Warrell, J.H. Elder, and F.M. Felisberti. Tied Factor Analysis for Face Recognition across Large Pose Differences. *IEEE Transactions Pattern Analysis and Machine Intelligence (TPAMI)*, 30(6):970–984, June 2008.

[13] C.E. Rasmussen and C.K.I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA, 2006.

[14] C. Sanderson and B.C. Lovell. Multi-Region Probabilistic Histograms for Robust and Scalable Identity Inference. In *Proc. International Conference on Biometrics (ICB)*, volume 5558 of *LNCS*, pages 199–208, June 2009.

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

[16] J. Saragih and R. Goecke. Learning AAM fitting through simulation. *Pattern Recognition*, 42(11):2628–2636, November 2009.

[17] J.M. Saragih. *The Generative Learning and Discriminative Fitting of Linear Deformable Models*. PhD thesis, RSISE, Australian National University, Canberra, Australia, January 2008.

[18] T. Sim, S. Baker, and M. Bsat. The CMU Pose, Illumination, and Expression Database. *IEEE Transactions Pattern Analysis and Machine Intelligence (TPAMI)*, 25(12):1615–1618, December 2003.