Synthesising unseen image conditions to enhance classification accuracy for sparse datasets: applied to chimpanzee face recognition

Roz Sandwell[1]
Roz.Sandwell@bristol.ac.uk
Alexander Loos[2]
Loos@idmt.fraunhofer.de
Tilo Burghardt[1]
Tilo@cs.bris.ac.uk

[1] Department of Computer Science, University of Bristol, Woodland Road, Bristol, BS8 1UB, UK
[2] Fraunhofer Institute for Digital Media Technology IDMT, 98693 Ilmenau, Germany

Abstract

To aid automated non-invasive population monitoring, we explore chimpanzee face recognition accuracy using a number of algorithms on images with pose and illumination variation, by synthesising images from a generic 3D model. The expense of expeditions and uncontrollable nature of this wild species and environment requires automated face recognition techniques to be robust to pose and illumination variance without incurring additional data collection or manual annotation costs. Unlike for humans, prior knowledge of chimpanzee face shape does not exist, leading us to synthesise 2D images from a custom-built generic 3D shape model for augmenting training and testing data. We use the resulting synthesised images to profile five existing face recognition algorithms. We show that synthetic data can be used to constructively augment training data, as three recognition algorithms have significantly increased accuracy for pose-offset data when augmenting the training data as compared to real data alone.

1 Introduction

By enhancing automated chimpanzee face recognition, we aim to reduce the manual load of collecting and analysing field data for estimating endangered species' population size. The elusive and uncooperative nature of chimps limits available data, guaranteeing neither frontal, well-lit examples nor sufficient variety of image parameters for realistic representation of individuals. Existing chimp recognition requires at least five near-frontal images per individual for training and results in more than half of the individuals captured in wild datasets being omitted from analysis due to insufficient data [7]. We propose that increased robustness to pose and lighting variation can be achieved without requiring more real data, and instead synthesising unseen image conditions, which saves manual effort and reduces data discarding.

We demonstrate that automatically annotated synthetic images generated in controlled conditions can be used to supplement real data to improve accuracy in non-frontally posed

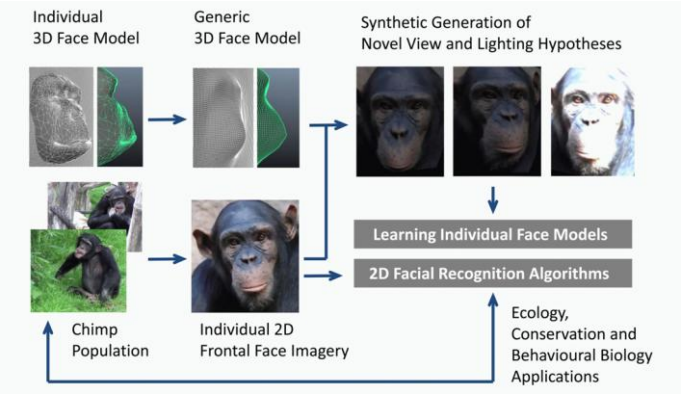

Figure 1: Proposed mixed 2D-3D approach to enhancing chimp face recognition. Projecting individual 2D facial textures of chimps gathered from known individuals onto a generic 3D model allows for the generation of synthetic appearance hypotheses. We show that these can be used to complement sparse sample sets for enhancing the profiling, training, and application of face recognition algorithms for chimps given sparseness and limited access to samples from real world populations.

test instances. This allows more complete coverage of image parameter variability for training and could allow inclusion of more individuals as fewer real images are required. Challenges of transferring synthesis methods developed for humans include the impracticality of obtaining a 3D scanned model and the 2D images being of an uncontrollable subject – in terms of pose and expression – from an uncontrolled environment, with natural lighting. Our synthesis is therefore constrained by limited subject information, however remains sufficiently representative to profile recognition algorithms and improve robustness to pose variation.

We make two contributions using synthetic data to explore the effects of pose and lighting on five existing face recognition algorithms (detailed in Section 2). Firstly we quantify the relationship between the controlled parameters and algorithm accuracy. Extensively annotated synthetic test images highlight the limitations of algorithm robustness to image variance, identifying the operating constraints and focusing future development. Secondly, we increase the generalisation capacity of the algorithms to horizontal pose variation. Augmenting available data with synthetic images in unseen conditions allows training on individuals in image conditions for which fewer real training images are available. We propose a mixed 2D-3D approach: using an approximate 3D model to augment 2D datasets for training and testing. We isolate and normalise the least complex parameters, working with pose and lighting. An overview is shown in Figure 1: we describe our data generation method and present experimental results indicating sensitivity to these parameters for all algorithms, and enhance the generalisation capacity and increase accuracy by augmenting the training sets.

2 Related work

Related work is primarily drawn from human face recognition, though there are several major differences for application to a wild animal. These include the limited number of available images, their uncontrolled nature and the absence of an anatomically correct 3D chimp model. Training set augmentation is in part motivated by the assertion that the

accuracy of face recognition in unseen poses is greater if there is a greater variety of poses present in the training data [13]. These existing face recognition algorithms are expected to have limitations in their representation of chimp face image variability, using limited available data and also respond to an increased amount of (synthetic) training data.

Face recognition algorithms. Four of the five face recognition algorithms considered are community-standard techniques developed for human face recognition: Eigenfaces [9], Laplacianfaces [4], Randomfaces [10] and Gabor-based Sparse Representation Classification (GSRC) [11]. The fifth has been developed specifically for recognition of great apes, and has been shown to outperform the other approaches for application to chimps [5]. Referred to here as LPP+GSRC, this last method combines Laplacianfaces with GSRC and uses Locality Preserving Projections (LPP) for the feature space transformation instead of Principal Component Analysis (PCA). Eigenfaces are a well-studied benchmark, and are known to be susceptible to pose variation, which leads to misalignment of the image pixels [13]. It has been reported that most experiments for holistic approaches are limited to twenty degree rotations [13], which may indicate the approximate limit of applicability of these methods. Eigen-, Laplacian-, and Random-faces differ in their dimensionality reduction techniques, but all use simple grey-level pixel-based information as features. GSRC and LPP+GSRC are based on Gabor features, which are known to perform well in face recognition due to their robustness against difficult lighting conditions.

Synthesising for recognition. Synthesised 2D images from 3D models have been used successfully to improve face recognition rates for humans in non-frontal poses. Synthesising rendered views from scanned 3D models has resulted in increased recognition accuracy [1, 2], however relies on prior knowledge of the individuals, requiring their cooperation to be scanned. Cylindrical textures have been constructed from five images per individual and explored to increase recognition accuracy without a more complex 3D structure [8], with overall results showing higher recognition accuracy for a consistently lit dataset. Both types of synthetic face representations above have improved recognition accuracy with pose variation, and their results suggest lighting control could also be informative. Many approaches using 3D information require prior knowledge of the face shape or control of training data. Methods exist which use 3D modelling and synthesis without the use of scanned images, and with a reduced initial training set requirement. These do however usually stipulate conditions on the gallery images used for training, for example requiring several images in controlled poses [8]; frontal and profile mugshots [12]; or additional sensor data such as the depth maps of 2.5D images [3]. The unavailability of head scans or controlled training images for chimps therefore leads us to use a generic shape model to approximate the 3D geometry of a chimp's face.

We aim to profile and enhance accuracy for the five recognition algorithms without controlled images or 3D data, and to separate the effects of lighting and pose by generating controlled synthetic datasets from a generic model.

3 Data, synthesis method and experimental approach

We perform our experiments using datasets collected from a zoo environment, and our synthetic images are generated using the following generic 3D model and method throughout our investigation.

Initial data. Our main dataset consists of 572 chimp face images containing 24 individuals. The faces are all in a frontal position – only minor vertical pose variation is

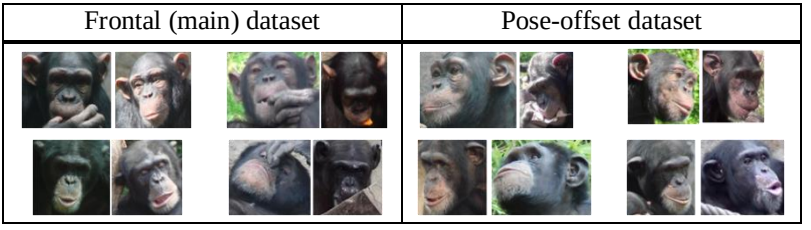

Figure 2: Example images for two individuals, (one per row). Lighting and pose (limited vertical offset permitted) examples are from the main dataset; semi-left and semi-right from the additional pose-offset profiling datasets.

permitted. Each individual is represented by at least five images and the left- and right-eye and mouth centres have been manually annotated for all images. This annotation does not require an expert, and could be replaced with a detection algorithm in future [6]. The following experiments using this frontal dataset have been ten-fold cross-validated by initially splitting the whole dataset into ten subsets: each time nine sets are used for training; one for testing. Each synthetic image is associated with one image 'seed': they only ever appear in the train or test set in which the seed is present. An additional dataset is used only for profiling the algorithms with respect to manually annotated horizontal pose variation. These are chimp faces controlled as above, however allowing moderate left- and right-turned poses. It retains topological equivalence, (i.e. retaining visibility of both eyes and mouth), and contains 490 semi-left and 506 semi-right images. An example of the variation of the image conditions present in both the frontal and manually annotated pose-offset data is shown for two individuals in Figure 2. The frontal images form the basic dataset for baseline testing and from which images are synthesised, and the pose-offset dataset provides a manually annotated real benchmark for our testing.

Synthetic images method. The 3D model has generic shape, a rigid alignment of projected textures and a straightforward reflectance model. Pose and lighting are varied independently, and synthetic images are annotated with the conditions under which they were generated. Each source image used to generate a set of synthetic images is aligned to the 3D model using the minimum Procrustes distance based on the three manually annotated key-points from the original image. This results in a map that is scale, shift and rotation normalised, and provides a simple and efficient texture projection. A simple shape model avoids extensive warping of the projected texture. It has Lambertian reflectance, as more complex reflectance models developed for human skin [14] are not directly transferrable to this species. The same 3D model is used for all synthetic images, allowing independent variation of lighting and pose for all projected textures.

Synthetic dataset creation. Synthetic images are created from each of the 572 images by altering the pose and lighting conditions for the 3D model. Horizontal and vertical pose angles are each varied between +/-30° in 10° increments. The lighting conditions include an ambient setting and can include a spotlight. In the images where the spotlight is present, it is at one of three intensities: high, mid or low. Its position is varied between +/-60° in 30° increments, leading to five different spotlight positions. An example set of synthesised images in Figure 3 displays the variation in synthesised images generated from a single chimp face input image. Only one parameter is varied at a time, the other parameter remains in a 'neutral' state of frontal pose or ambient lighting. Rendering under these controlled conditions results in 11,440 unique synthetic images: each containing either horizontal or vertical pose variation or spotlight exposure or position variation.

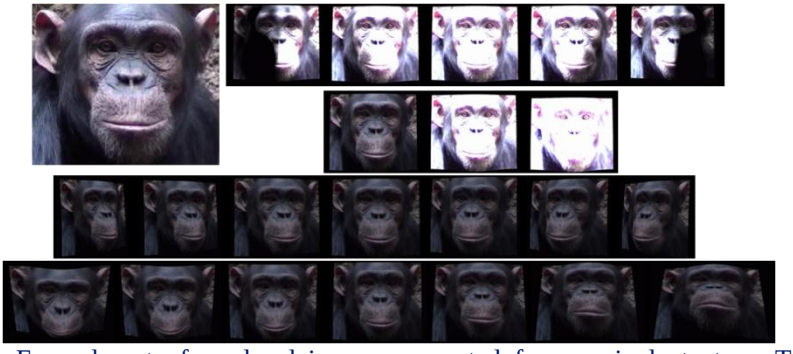

Figure 3: Example set of rendered images generated from a single texture. Top left: original image. Row 1: spotlight position variation (mid-intensity); Row 2: spotlight intensity variation; Row 3: horizontal pose variation; Row 4: vertical pose variation. Twenty images are generated per texture (discounting the two duplicates: frontal neutral at zero degrees pose offset and mid-frontal spotlighting).

Further images can be synthesised by combining parameter variations, however we have restricted the synthetic conditions to explore the effects of lighting and pose variation on face recognition accuracy separately. These are used as test images to quantify the effects of pose and lighting variance on face recognition algorithm accuracy and separately to augment the training sets to enhance generalisation capacity when tested on real frontal and pose- offset data.

Experimental overview. In Section 4 we introduce our face recognition algorithm parameters and use real train and real test data only. Section 5 uses synthetic test sets for the same real-trained models, to analyse their accuracy at a range of horizontal and vertical pose angles and spot lighting conditions. We also present synthetic data in previously unseen conditions to the learning stage of the algorithms, using purely synthetic trained models on the same synthetic test sets. Finally, in Section 6, models are built using synthetically-augmented training data (including the real seeds) to compare to the real-trained models on real test conditions.

4 Pre-processing and real data benchmark

Firstly we establish a benchmark for the five algorithms applied to chimp faces, trained and tested on real data only.

Face recognition algorithms. Real and synthetic images are pre-processed before being used to train or test the five face recognition algorithms. Images are aligned by their annotated eye and mouth coordinates: firstly by rotation; then warped using a projective transformation such that these features are at a uniform position throughout the entire dataset; and finally a histogram equalisation is used for lighting normalisation before the face is scaled to 64x64 pixels. To obtain a fair comparison between algorithms, 160 features are used for all applied feature space transformation techniques. For Randomfaces, GSRC, and LPP+GSRC we use SRC (Sparse Representation Classifier, where the feature vector is represented as a linear combination of training images [10]) for classification and a nearest neighbour classifier for the remaining two, as used in the original publications. We use five scales and eight orientations for the generation of the Gabor kernels. After convolving the input image with the resulting forty Gabor wavelets, we down-sample the magnitude-matrix by a factor of eight using bilinear interpolation, and transform the

Figure 4: Examples of incorrect recognition instances for real-trained LPP+GSRC.

resulting feature vector to a 160-dimensional subspace. These face recognition parameters are kept constant throughout, and real and synthetic images are similarly processed before being used for training or testing.

Real, frontal data for training and testing. Testing and training using only the real images provides a baseline against which we compare our later results for synthetically augmented training data. The rank-1 accuracies of the five face recognition algorithms for real train and test data (the frontal dataset only) are shown as black-line plots in Figure 6. LPP+GSRC achieves the highest accuracy of 90.24%, and GSRC obtains the second highest accuracy, 83.74%. As there is limited pose variation present in these images, this indicates that the use of Gabor features may well result in improved illumination invariance as compared to the pixel based methods. Failure case examples for LPP+GSRC are shown in Figure 4 – pose, lighting and expression are represented. This provides baseline results indicating the superiority of the custom chimp method LPP+GSRC and the Gabor based methods with accuracies for a fairly controlled (subjectively manually annotated frontal, well-lit) real dataset.

Additional real test data with horizontal pose variation. The same real-trained models are used to obtain accuracies from the pose-offset datasets. This is to characterise the real-trained models' robustness to manually annotated pose variation. The entirety of the offset face dataset is used for testing each of the ten cross validated models for each of the five algorithms. All five algorithms suffer a significant decrease in accuracy with this horizontal offset in pose as compared to frontal – by as much as 25-38 percentage points, as seen in Figure 6. We later quantify this drop in accuracy with specific head angles using synthetic test datasets. Lighting conditions have not been manually annotated and are difficult to objectively quantify in natural images, so we have been unable to similarly investigate baseline lighting effects. This significant decrease in accuracy caused by offset horizontal poses highlights the weakness of using only frontal images to train the algorithms, and the extent of their inability to generalise to pose variation.

5 Profiling the face recognition algorithms using synthetic test data

The synthetic images and their associated annotations are first used to quantify the relationship between the controlled parameters (pose and illumination) and the accuracy of the face recognition algorithms. Four synthetic test sets are used: one each for horizontal and vertical pose variation, lighting exposure and lighting position variation. In each case, every image from the original real frontal test sets is replaced by its seven, three or five synthetically varied counterparts. Firstly these test images are used for the real-trained models for each algorithm built using the cross validation sets as previously. Then new models are trained using similarly constructed synthesised training sets and profiled in the same way.

Pose variance. The algorithms' accuracy clearly suffers beyond a pose-offset of more than ten degrees in either direction, as can be seen in the black-line plots of Figure 5(a) and (b). The decrease is more marked in horizontal than vertical pose variation. This is likely as a

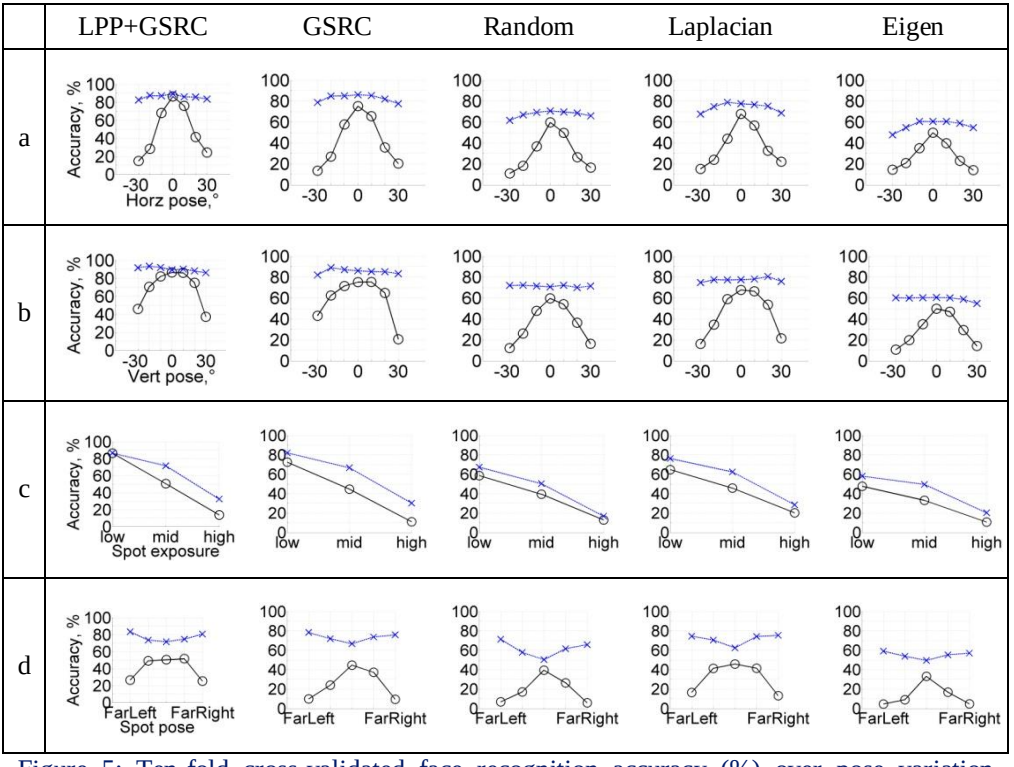

Figure 5: Ten-fold cross-validated face recognition accuracy (%) over pose variation (degrees deviation from frontal) and spotlighting variation using synthetic test sets for five recognition algorithms. Black with circle markers are results from models trained on real data; blue with crosses are from models trained on relevant synthetic training data. Test set variation by row: (a) horizontal pose; (b) vertical pose; (c) spot intensity; (d) spot position.

result of the real training data being more stringently controlled for horizontal variation than it was vertical. Especially notable at a vertical head angle of -30° (tilting downwards), the greater accuracy of the LPP+GSRC and GSRC algorithms as compared to others suggests a more flexible representation of individuality by using Gabor features rather than pixel-based methods, which allows higher accuracy at more extreme pose variation. The asymmetric shape of the accuracy curve against vertical variation is interpreted as greater invariance to a head tilting downwards than upwards. All five algorithms demonstrate a decline in accuracy for poses that deviate from frontal, indicating a limited generalisation capacity beyond that present in the training data.

Illumination variance. Spotlighting adversely affects accuracy across all five face recognition algorithms, as seen in the black-line plots of Figure 5 (c) and (d). At low spot intensity (still retaining ambient lighting), frontal spot-lit test faces exhibit similar accuracies to ambient lit images. Increasing the intensity causes a decline that reflects an increasing "white-out" of the facial features, as visible in the example on Row 2 of Figure 3. Spotlight at mid-intensity and varying horizontal angle with respect to the frontal face also indicates declining accuracy at more extreme angles. This is likely due to the introduction of shadowed blacked-out patches that become present in the synthetic face images, amounting to occlusion of the opposite side of the face from the spotlight. This increasing 'occlusion' affects the algorithms differently: LPP+GSRC and Laplacianfaces

do not suffer as great a decline at the 'mid-right' spot position as the remaining three. All five algorithms again suffer with increasing intensity and extreme position of spotlighting.

Summary of synthetic test results. Testing on synthetic data has revealed sensitivities of all algorithms to pose and illumination variation. Varying horizontal angle causes the sharpest decline in accuracy with pose variation, and spotlighting also reduces accuracy. LPP+GSRC and GSRC display greater robustness to vertical pose variation as compared to the other algorithms, and LPP+GSRC is generally the best performing overall when faced with image variation.

To address this decline in accuracy with image variation, the following sequence of experiments is conducted using algorithms trained on purely synthetic image sets (real seed images excluded) with only one parameter varied throughout its entire stated range. For example the horizontal-pose model is trained on sets of seven synthetic images in each horizontal pose in place of the real training image. Each of the new synthetically-trained models (ten for each algorithm, as before) is then tested on its respective synthetic test set – by presenting a richer training set, we provide a more dense representation of the possible image variability.

Synthetic train and test. As expected, training using 100% synthetic train and test sets produces more uniform accuracies across the pose and illumination variation conditions, as seen by the blue-line plots of Figure 5. This indicates that as we present datasets containing increased parameter variation for training, declines in accuracy become increasingly due to insufficiencies in the algorithms' underlying techniques for extraction and representation of features. Overall, the accuracies are much more uniform across parameter variation as compared to the real-trained algorithms - maintaining over 80% accuracy across horizontal and vertical pose variations of +30° to -30° for LPP+GSRC - although there remains sensitivity to strong spotlighting. This uniformity of accuracy across parameter variation is likely due to a good alignment of the variance exhibited in training and testing data: the training set is representative of the variance of the testing data. The flattening of the accuracies across the parameter ranges indicates an ability (especially of LPP+GSRC and GSRC) to generalise and represent the variety present in the training data. It does appear to be beneficial to learn from these synthetic images containing unseen or more extreme parameter variation, rather than real frontal data alone to achieve greater consistency.

6 Synthetically augmenting training data

Finally we present results for the five algorithms' models built using a synthetically-augmented training set and tested on the original real datasets; both frontal and offset. Algorithms trained on sets augmented with pose-offset synthetic data and tested on real data provide a comparison to the real-trained algorithms, indicating the increase by using synthesis to populate the sparse data representation. Maintaining the same cross validation sets, four synthetic images are added for every training image: one each with +20° or -20° horizontal offset or +20° or -20° vertical offset. The resulting training sets are therefore 80% synthetic data, and are five times as large as the seeding real-only datasets. Models are trained on these sets, are then tested as before on the real datasets. The augmentation includes horizontal offsets as the aim is to increase the recognition accuracy on offset-posed datasets, and includes vertical offsets to reflect the moderate tilt variation permitted in the original data.

Accuracy for the horizontal pose-offset real datasets is at least maintained by all five algorithms, and increased for LPP+GSRC, GSRC and Randomfaces, as compared to the

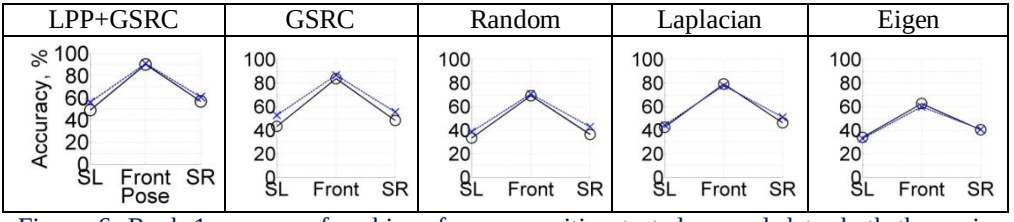

Figure 6: Rank-1 accuracy for chimp face recognition tested on real data: both the main frontal dataset (ten-fold cross-validated) and the horizontal pose offset datasets (SL: semi-left; SR: semi-right). Black with circle markers is real frontal data trained model; blue with crosses is synthetically augmented train data. Note the similarity between the frontal accuracies, and the increased accuracies of three algorithms for the semi-offset test data when trained on augmented as compared to real, (LPP+GSRC; GSRC; Randomfaces).

real-trained algorithms (Figure 6): the additional synthetic images provided for training have led to greater robustness to pose variation exhibited in these offset datasets. The accuracy on the frontal dataset is comparable to the real-trained model indicating that the inclusion of our synthetic images does not harm any of the algorithms' performance, not reducing the peak accuracy as when using synthetic training data only. As expected, the synthesis does not extract any additional individuality from the images: the maximal amount of individuality for frontal poses was present in the original training set, so the accuracies have remained roughly the same. Illumination is kept ambient throughout, though there may be residual lighting differences resulting from the original illumination present in the real training images. Additional incorporation of synthetically lit images may further increase recognition accuracy. The increased accuracy of the three algorithms indicates robustness to pose variation introduced by the synthetic images in place of additional data collection and ecologist expeditions.

7 Conclusions

We conclude from our experiments that simulating additional training data under different pose and illumination conditions does indeed increase the robustness of some face recognition algorithms when applied to chimpanzee faces, without introducing additional annotation or data requirements. Testing a real-image benchmark in controlled synthetic conditions has allowed quantification of greater robustness to vertical pose variation than horizontal, which corresponds to the variability permitted in the training set. Experiments using purely synthetic data have revealed an ability of the algorithms to generalise beyond neutral images to more extreme pose and lighting situations. Finally, generalisation capacity has been increased on real test images using synthetically augmented training sets, resulting in increased robustness to horizontal pose variation for three of the five algorithms.

Despite already resulting in improvements on real test data, our system has not yet reached the limit of its retention of individuality – the synthetic profiling suggests that the algorithms are capable of representing horizontally offset poses with greater accuracy still. Recommended future work therefore includes determining to what extent augmented-data trained models represent the variation present in the underlying real datasets, for example by including a greater degree of synthetic pose or lighting variation in the training sets. The generic shape model can be verified and optimised by performing similar experiments using simple affine plane transformations or more complex shape approximations. Mixed

approaches will also be necessary to move beyond the current topological constraints – for example faces for which the eyes or mouth are not present. We intend to further explore the application of our technique to one-shot learning: using only one real image of an individual to seed synthesis for recognition in a range of image situations.

Acknowledgements. AL funded by the BMBF (Germany). RS funded by EPSRC and BBSRC (UK). We thank Zoo Leipzig; Wolfgang Köhler Primate Research Center; Josep Call; Josefine Kalbitz; Laura Aporius for images and annotation; Professor Mike Mendl; Thales Research and Technology UK for additional funding. This work was supported by the Max Planck Society.

References

[1] Blanz, V., & Vetter, T. Face recognition based on fitting a 3D morphable model. *Pattern Analysis and Machine Intelligence*, 25(9), 1063-1074, 2003.

[2] Dahm, N., & Gao, Y. A Novel Pose Invariant Face Recognition Approach Using A 2D-3D Searching Strategy. In *Proceedings of the International Conference on Pattern Recognition, (ICPR)*, 3967-3970, 2010.

[3] Hajati, F., Raie, A. A., & Gao, Y. Pose-invariant 2.5 D face recognition using Geodesic Texture Warping. In *International Conference on Control Automation Robotics & Vision (ICARCV)*, 1837-1841, 2010.

[4] He, X., Yan, S., Hu, Y., Niyogi, P., & Zhang, H. J. Face recognition using Laplacianfaces. *Pattern Analysis and Machine Intelligence*, 27(3), 328-340, 2005.

[5] Loos, A. Identification of great apes using gabor features and locality preserving projections. In *Proceedings of the international workshop on Multimedia analysis for ecological data, (ACM Multimedia)*, 19-24, 2012.

[6] Loos, A., & Ernst, A.. Detection and Identification of Chimpanzee Faces in the Wild. In *Proceedings of the International Symposium on Multimedia (ISM)*, 116-119, 2012.

[7] Loos, A., & Pfitzer, M. Towards automated visual identification of primates using face recognition. In *Proceedings of the International Conference on Systems, Signals and Image Processing (IWSSIP)*, 425-428, 2012.

[8] Rama, A., Tarres, F., Onofrio, D., & Tubaro, S. Mixed 2D-3D Information for pose estimation and face recognition. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2006.

[9] Turk, M. A., & Pentland, A. P. Face recognition using Eigenfaces. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 586-591, 1991.

[10] Wright, J., Yang, A. Y., Ganesh, A., Sastry, S. S., & Ma, Y.Robust face recognition via sparse representation. *Pattern Analysis and Machine Intelligence*, 31(2), 210-227, 2009.

[11] Yang, M., & Zhang, L. Gabor feature based sparse representation for face recognition with gabor occlusion dictionary. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 448-461, 2010.

[12] Zhang, X., Gao, Y., & Leung, M. Recognizing rotated faces from frontal and side views: An approach toward effective use of mugshot databases. *Information Forensics and Security*, 3(4), 684-697, 2008.

[13] Zhang, X., & Gao, Y. Face recognition across pose: A review. *Pattern Recognition*, 42(11), 2876-2896, 2009.

[14] Zhang, X., & Gao, Y. Heterogeneous Specular and Diffuse 3-D Surface Approximation for Face Recognition Across Pose. *Information Forensics and Security*, 7(2), 506-517, 2012.