Synthesising Security Camera Images for Face Recognition

Joost Loohuis
University of Twente
P.O. Box 217, 7500AE Enschede
The Netherlands
j.e.loohuis@student.utwente.nl

ABSTRACT

The field of face recognition sees a lot of development, but its use in forensic settings has lagged behind. One of the reasons for this is the lack of face image sets of sufficient size for training neural networks. This paper looks at the creation of a system to generate such a set of face images in a simulated forensic setting, since none exist for this purpose. The viability of using those images is tested with the FaceNet face recognition network.

Keywords

face recognition, forensic face recognition, face generation, image synthesis, morphable model

1. INTRODUCTION

Face recognition has been a constant area of development for decades. While modern systems can produce some impressive results they often rely on controlled conditions or the input of multiple sensors [10]. Input images where the subjects are not aligned and lit consistently and in clear view of the camera may still pose a problem [15, 17]. The training of neural networks for face recognition also relies on large data sets that can require modelling, photographing and labelling by professionals [17, 29].

A field where this technology is starting to see more use is that of forensics, both as tool for assisting investigators and to provide evidence. Not all courts accept it to the same degree as existing techniques, such as finger print matching [12, 26]. One of the factors contributing to this is that, while large image sets for regular use are readily available [7], the size of those that are close to the adverse conditions in surveillance and security camera situations is relatively small [5, 27]. The use of sets of synthetic images where few are available to train face recognition networks has shown promising results [14].

This paper seeks to create a system for generating a training set of images that mimic these conditions. It builds on existing research in the generation of face images and combines this with rendering techniques to simulate the conditions encountered in forensics. This system can then be used to train and evaluate existing network architectures on their suitability for this challenging use case.

34th Twente Student Conference on IT Jan. 29th, 2021, Enschede, The Netherlands.

To test the quality of the system it is tested with a state of the art face recognition neural network and compared with an existing forensic image set.

2. RESEARCH QUESTIONS

The questions this paper aims to answer are as follows:

RQ1 Can existing techniques for face generation and rendering be used to create images similar to those created by a security camera?

RQ2 Can this image set be used for the training and evaluation of face recognition systems for forensic settings?

- **RQ2.1** How do existing face recognition systems perform on images generated from a ground truth as opposed to real security camera images?
- **RQ2.2** Are the faces in generated images seen as the same person as the ground truths they were generated from?

3. RELATED WORK

Research has already been done as to how images sets of faces can be synthesised, since the neural networks that are used for face recognition generally see an increase in performance when trained on a larger data set, as well as for other purposes, such as use in the movie industry. Both the direct generation of 2D face images [13, 19] and, since quite recently, renderings of 3D morphable face models have been looked at to increase the size of available data sets [23, 25, 3, 4]. Another method that has been used for this purpose is the reverse rendering of existing 2D image sets into reposable 3D models [11, 8].

The field of forensics, while starting to increase its research into face recognition, has not yet produced a golden standard for face recognition systems. Existing systems are used to produce a measure of the probability that two pictures are of the same person [1, 12]. Generally face recognition can be done through the comparison of extracted features or though deep neural networks, of which deep convolutional networks show the most promise for pose invariant recognition [6].

4. BACKGROUND

Face Model Generation

There are two main ways of acquiring 3D face models: face capture and direct 3D generation. Face capture uses imaging of a real face as a basis for the generation of a 3D model. In state of the art systems this is done by collecting depth information by using images from multiple views or by using information from additional sensors

[21, 10]. While it produces less accurate results a lot of research in this area has also looked at producing models from more practical monocular images [30]. A large part of these systems is based around the extraction of a feature vector, containing a numerical representation of the face, which is then fed into a morphable 3D model, based on an average encoding of many face models, to estimate the general shape of the face [3, 2]. Such statistical face models are limited in the face shapes that they can represent, as these are bound by its latent space [4]. An alternative technique that gets around this limitation is the use of volumetric regression, where the 3D shape is matched directly without a reference model [11]. Direct generation of 3D models uses the same techniques as face capture. A parametrised face model can be used without the encoding of an existing face through giving it a random feature vector [3]. Recent works have also looked at direct generation of 3D meshes without morphing a predetermined model, since these models are still limited in the face shapes they can represent.

Forensic Face Recognition

Face recognition can generally be used in two ways: one to one image comparison for identity verification, and one to many to make an identification in a set of images. When making an identification the set can both be closed—when it is certain that the subject is in it—or open. In either case the systems in forensic use return a ranking of the likelihood ratio [1, 12]. The likelihood ratio is calculated from the face recognition system output using eq. (1), giving a higher score when it is unlikely the image depicts someone else and likely to match the subject. This should give a more robust measure of how probable it is that we found the right person than just the similarity score given by the face recognition system.

$$\frac{P(S|Y,I)}{P(S|N,I)} \tag{1}$$

Where:

S: is the score given by the face recognition system

I: is the background information, like witness testimony

Y: is the hypothesis that the subject of the security camera image is the person we found

N: is the hypothesis that the subject of the security camera image is someone else

Many techniques for face recognition have been developed, both with neural networks and without. Those techniques that are not based on neural networks often use feature extraction, such as principle component analysis, to compare numerical representations of the subjects of images [16]. Systems that use convolutional neural networks give some of the best results, often pre-processing the image inside the system to improve face alignment for improved performance [16, 20].

Image Degradation in Security Cameras

In an uncontrolled environment the quality of images can vary widely. Low resolution, blurriness, unfavourable illumination and varying alignment or occlusion due to subject pose are just some of the most occurring problems [7]. When performing face recognition in forensic settings there are often even more challenges, since systems are often set up cheaply and subjects may actively be trying to not be captured on cameras. This exacerbates the problems through large camera angles, deliberate face obstructions, low sensor dynamic range, heavy compression and an increase in optical aberrations [26, 12], as shown in fig. 1. The optical aberrations are caused by imperfections in a cameras construction, specifically the lenses, which cause distortions and defocus, as well as chromatic and achromatic aberrations [24].

(a) Clean reference. (b) Security Camera.

Figure 1: An example of degradation from the SCface set

5. PIPELINE DESIGN

To imitate these low-quality images a pipeline is needed that takes a reference image or random identity representation as input, produces a model that can be re-posed, re-lit and otherwise manipulated, and finally creates low quality images from this model. The created images should match the general quality of the security camera images that we want to recreate and get a similar response from face recognition systems. To ensure it can be used to recreate many different scenarios and with a wide range of research setups it should ideally be easy to adapt or expand. To this end we propose a system combining the face reconstruction from Deng et al. [2], the Basel 2009 morphable model for which it creates embeddings, and the Blender 3D environment for rendering and post-processing.

Face Reconstruction

To recreate the subjects of an existing data-set as 3D models the information for creating such a model first needs to be extracted from reference images. Such an embedding could also be randomly generated, but this would make performance comparisons for face recognition systems more difficult. To get the embeddings the convolutional neural network designed in Deng et al.[2] is used. This model has been trained with both a photometric loss, which calculates the difference on a per-pixel basis, and a perception loss, which finds the difference between features extracted by a face recognition network. According to the authors their photometric loss function, which includes a probability model for pixel-level skin colour, ensures an accurate texture, while the perception loss function prevents loss of detail in both the texture and shape of the model. It is based on the ResNet-50 architecture [9], aided by a pose and illumination estimator to give pose, lighting, identity, texture and expression vectors from an image with five key points marked.

Morphable Model

The face embeddings contain all the information needed to reconstruct the faces as 3D models. They are encoded in such a way that they can be applied to a standard model to

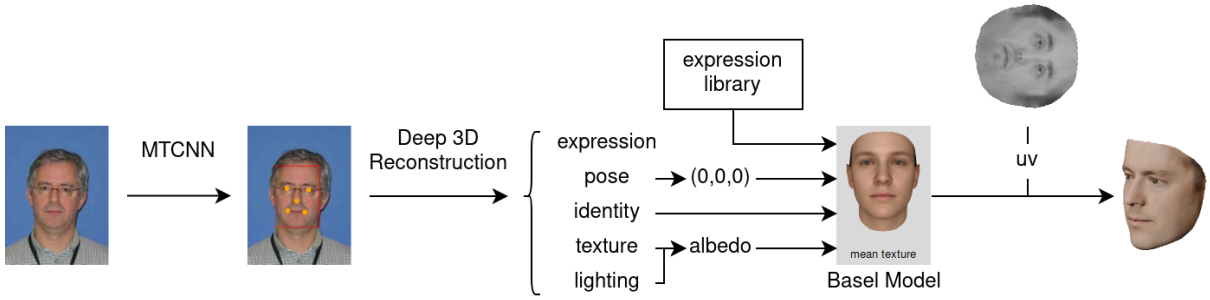


Figure 2: An illustration of the steps in generating a 3D face model from a reference.

morph it to the shape and texture of the original face. In Deng et al.[2] this is done according to eq. (2) as described in their paper. In this formula $\bar{S}$ and $\bar{T}$ are the mean shape and texture , and B_{id} and B_t are the shape and texture principle component base vectors from the Basel 2009 model [18]. This model consists of a parametrised 3D shape model and an albedo texture, containing colour information without shadows and other lighting effects, obtained from 3D scans of 100 male and 100 female individuals. Also included in the formula is B_{exp} as an expression principle component base for the face shape. In the model from Deng et al. [2] the Basel model was extended with expression deformations from [8] to enable the reconstruction of expressions.

$$\begin{aligned} Shape = S(\alpha, \beta) = \bar{S} + B_{id}\alpha + B_{exp}\beta \\ Texture = T(\delta) = \bar{T} + B_t\delta \end{aligned} \tag{2}$$

The implementation in the pipeline is a modification of the one published by the authors of Deng et al.[2] in Python[1] as displayed in fig. 2. In the modified version the position information is disregarded and set to zero to ensure a constant head pose in the exported 3D model. The variable β can either be passed on or replaced with an expression vector from a library of expressions, for example the zero vector to give a neutral expression on the 3D model. For the texturing the albedo information is stored for each vertex in the model, ensuring the texture contains no lighting information from the original photo. Each vertex is also given a uv-coordinate to support more advanced texturing in the 3D scene.

Scene Generation

The 3D models of the faces have been generated with expressions baked in, but they still require posing and lighting before they can be rendered. This is done within Blender, a very complete open-source 3D suite, in combination with an add-on for automatic batch processing. The add-on uses the internal Blender API, so all of its original features can be used to recreate almost any security camera setting. The add-on imports the face models, randomises the camera and lighting position and tilts the face. All random values get picked again for each individual image between a minimum and maximum value set by the user and can be turned off if they are not required. For the camera the height, distance and yaw are set together, so the face is always at the centre of the image. The lighting is set through the location and angles of the chosen lamp. The pose of the face is varied independently from the camera and light setup by rotating it around each axes independently. All generated images in this paper were rendered using the Eevee physically based rendering engine, but the Cycles engine can be used for more advanced options at a cost of speed.

[1]https://github.com/Microsoft/Deep3DFaceReconstruction

Post Processing

Most of the post-processing, like lowering the dynamic range, can be done inside Blender, since it includes many photography filters. Since it is normally used to create high quality images it does not handle low resolution images like a security camera, instead giving us blurry images. To remedy this sub-sampling is done outside of Blender by dividing the image in squares of a given size and averaging the colour of four random pixels for each square, as shown in fig. 3. The resulting image is scaled down by the size of the square, where 10 times down sampling uses a 100 pixel square, and has harsher colour transitions between pixels. The final image is then saved with lossy jpeg compression.

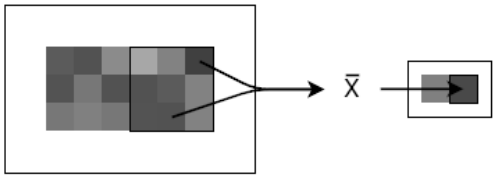


Figure 3: An illustration of three times down-sampling.

6. METHOD

To answer the research questions, and thereby measure the useful quality of the pipeline, the outputs of the system were compared against an existing data set. A subset of the ForenFace image set [27] , as seen in fig. 4, was used for this purpose, with the image set c1a7 used for medium-quality comparison and c3a3 used for low-quality comparison. The details about these images can be found later in this chapter. For the recreation of both sets the a set, consisting of passport photos, was used as a reference to generate the face models from. Since many subjects smiled to some degree in their photo two sets of face models were generated: both with the original expression of the reference image or with the zero vector for a neutral expression. Because pose, lighting and post-processing influence each other, and since any small change can drastically change the final image, any loss function would have given very different results for each image in a set, so most steps in recreating the scenes were done at least partially empirically. The camera position was matched in Blender by calculating the distance and height to the centre of the head, assumed to be at 165cm, using the information from the ForenFace paper. Light positions had to be estimated

from the shadows and highlights in the reference images. Post-processing filters were built empirically to create a visually similar degradation as in the original image sets. The high resolution 'mug shots' were created with a frontal sun lamp and no post-processing. In all cases the setup was created so the generated images matched the look of all reference images in a set. Especially for the very low quality c3a3 set this proved to be difficult, as the pose and resulting lighting varied greatly between subjects.

The generated images were cropped to the bounding box around the face, as detected by the MTCNN landmark detector [28], and scaled to 160x160 pixels. In outliers where the detector found no face the image was not used in the measurements. The FaceNet neural network [22], which is described later in this chapter, was used to encode these pictures into embedding vectors. A matrix of similarity scores was calculated by taking the squared Euclidean distances from each subject to the reference embedding of each other subject and then subtracting this distance from the highest possible similarity score. For every unique score the amount of true and false positives was calculated by using it as a threshold. Through these values we can see how separated the identities of the generated images are, and thereby how distinctively identifiable the faces seen in those images.

Evaluation

To make the obtained results more understandable they are displayed as receiver operating characteristic (ROC) curves. Such an ROC curve is created by sweeping trough a range of thresholds and then noting for how many subjects the similarity to their ground truth is above that threshold, as well as how many incorrect ground truths get above that threshold. These numbers give us the true positives (TP) and false positives (FP) respectively, as well as false negatives (FN) and true negatives (TN), allowing us to calculate the true positive rate as $TPR = \frac{TP}{TP+FN}$ and false positive rate as $FPR = \frac{FP}{FP+TN}$. Plotting these against each other gives us an indication of the performance of the classifier for many thresholds, and thereby how separated the identities are for each subject. For a perfect face recognition network there would always only be true positives and no false positives, since the identities are perfectly separated, giving a right angle in the upper left corner. A random classifier would give as high a true positive rate as a false positive rate, which is shown as a linear baseline between the bottom left and top right. To make it easier to directly compare ROC curves the area under the curve (AUC) can be calculated, which would be 1.0 and 0.5 for the perfect and random classifier respectively. All AUC values in this paper were approximated using the trapezoidal sum, since the data points are do not form a continuous line.

ForenFace

The ForenFace data set consists of images mimicking security camera footage, inc, meant for forensic research. It contains stills of 97 subjects filmed by six cameras standing at four positions, facing set directions at each position. The subjects were filmed both with and without a baseball cap, but only the images without obstruction were used in this paper. The c3a3 image set, as seen in fig. 4c, was taken with the Panasonic WVP480, at a distance of 4.20m and angle of 30° to the subject, looking down at an angle of 25° The orientation of the faces varies in the set, which was accounted for in the generated images by allowing a 10° variance in the pitch and 5° variance in the roll of the faces, along with a 10° variance in yaw from the line to the camera, as this resulted in visually similar poses. For post-processing the green and blue channels were translated by a few pixels in opposite directions, the dynamic range was decreased and colours were corrected. The final images were down-sampled by a factor ten except when indicated differently. The c1a7 set, as seen in fig. 4b, was taken with a Watec WAT-230A in greyscale, at a distance of about 80cm. Since the subjects were instructed to look directly in the camera, the face and camera face each other directly in the generated images. Since most subjects look above or below the camera at a varying height the pitch of the generated faces varies by 10°, combined with a variation of 2° in yaw and roll. The only post processing done on the generated images is greyscale conversion and three times down-sampling.

(a) Passport (b) c1a7 (c) c3a3

Figure 4: Examples of the images used from the ForenFace data set.

FaceNet

FaceNet is a deep convolutional neural network designed to give a Euclidean embedding of the face in an image, which means there is a direct correspondence between the squared Euclidian distance between two normalised embeddings and their similarity, in this case on a scale from zero to four. On the Labeled Faces in the Wild data set it is one of the best scoring networks with an accuracy of 99.60±0.09 percent [16], which should make it a good representative of the state of the art in face recognition. The system takes a square cropped image face of a set size as an input and produces a 128 dimensional numerical representation of the identity of that face. For the measurements the implementation by Hiroki Taniai[2] in was used with an implementation of MTCNN[3] for cropping the faces.

7. RESULTS

Image Comparison

To get an idea of how well the quality of the generated images matches the ground truths we can visually compare them for one of the subjects. Since the original lighting could only roughly be matched and the pose of each subject was slightly randomised these can differ somewhat from the original and between each subject. The aim was to match the general look and quality for the whole image series more than to create exact matches. Only one subject in the ForenFace data set can be shown here for privacy reasons, so these images are only an average representation of how well the generated images match the references.
The high quality 'mugshot' images give a good overview of the quality of the 3D models, since the high resolution and ideal setup prevent any divergence due to post-processing or camera and light placement. The reference face looks

[2]https://github.com/nyoki-mtl/keras-facenet
[3]https://pypi.org/project/mtcnn/

(a) Original Image (b) Generated Image

Figure 5: A comparison of high quality images.

(a) Original Image (b) Generated Image

Figure 6: A comparison of medium quality c1a7 images.

wider, with a smaller nose. This can partially be explained by the virtual camera using a shorter focal length than the original cameras, but the morphable model also seems to make large nose tips. The texture matches quite well for most subjects, but darker skins are lightened significantly and the resolution is limited due to using vertex colours. Another problem of the morphable model is that all the details in the face, like laugh lines, and sharp features, like cheekbones, get smoothed out.

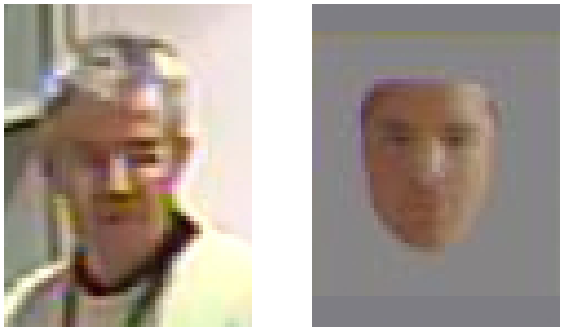
Due to the c1a7 images being black and white they generally match the skin tone better. Because the light is coming from quite a sharp angle both the original and generated images shown a lot of self-shadowing in the faces. These shadows highlight the details and sharper features in the faces, once again showing how these got smoothed out in some of the subjects.

The c3a3 images are of a very low quality, which hides most discrepancies in the details. In both the real and generated images the eyes and mouth are reduced to coloured smears in many images. The distortion of colours is a little bit larger in some of the reference images, but the colours match quite well overall. The loss of sharp features in the 3D model causes the highlights and shadows to be smooth in the generated images, wheres these are very sharp in the original images. The graininess and compression artefacts are visually similar in both sets of images.

Face Recognition on Generated Images

Intra-Domain

Before a comparison can be made between the identities that FaceNet sees in real and generated images we first need a baseline for how well the face recognition handles synthesised faces in the first place. This is done by comparing the similarity of high and low resolution images within the real and generated image sets separately. The results of these measurements are plotted in fig. 9 and fig. 10 for the c3a3 and c1a7 sets respectively. All the generated images used for these measurements use the neutral expression override unless otherwise specified.



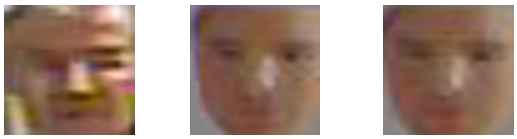
(a) Original Image (b) Generated Image

Figure 7: A comparison of very low quality c3a3 images.

In fig. 10 we can see that the general performance is the same for both the real and generated images, with only about a 1% difference between the AUC values. The curve for the real faces does not reach a full true positive rate for a large range of false positive rates, indicating that a few faces were confusing the network, which explains the lower area under this curve. The top left corner of the real curve is sharper though, which means that a slightly larger number of real faces was identified correctly without confusion than the generated faces.

In the curves for the c3a3 image sets a large difference can be seen between between the images that were down sampled to different sizes, with 10% less down-sampling giving about a 15% improvement, and the real images performing about 5% better than images that were downsampled to a tenth of their original size. When the images are scaled to the size used by FaceNet, shown in fig. 8, only a minor difference in the loss of detail can be seen, notably around the mouth. A possible explanation is that this lost detail happens to include those features that FaceNet looks at, as it may move the colour contrast that distinguishes these features past a point where FaceNet can clearly distinguish them.

A similar range of curves is seen when different combinations of the original expression and the neutral expression are used for both the low and the high resolution images, with the own expression in the low resolution image causing a large increase (16% and 19%) in performance regardless of the expression in the high resolution image. A possible explanation for this is that since some subjects smiled in their passport photo, using their own expression creates a clearer mouth in the low resolution images, which could help FaceNet, but comparing the individual scores of smiling and non-smiling subjects does not confirm this.



(a) Reference Image (b) 9× Down-sampled (c) 10× Down-sampled

Figure 8: A comparison of downsampled c3a3 images, resized to 160 by 160 pixels.

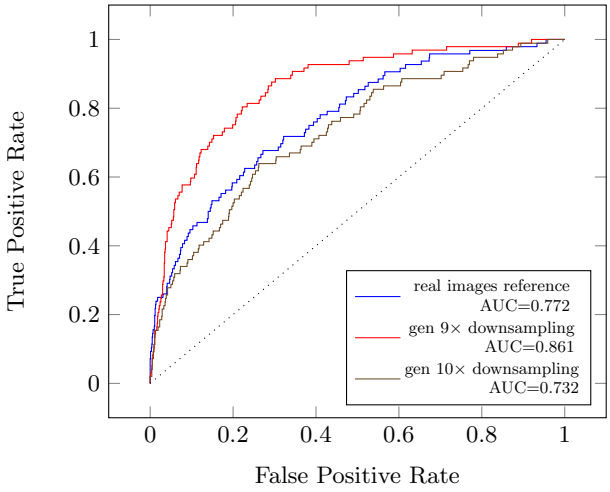


Figure 9: ROC curves for face recognition in the very low quality c3a3 images

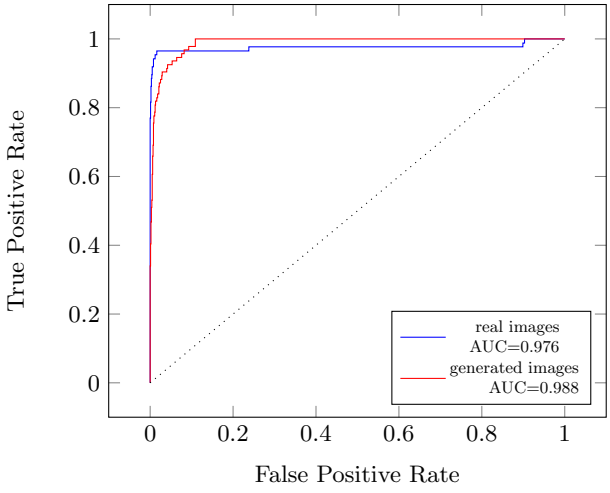


Figure 10: ROC curves for face recognition in the medium quality c1a7 images

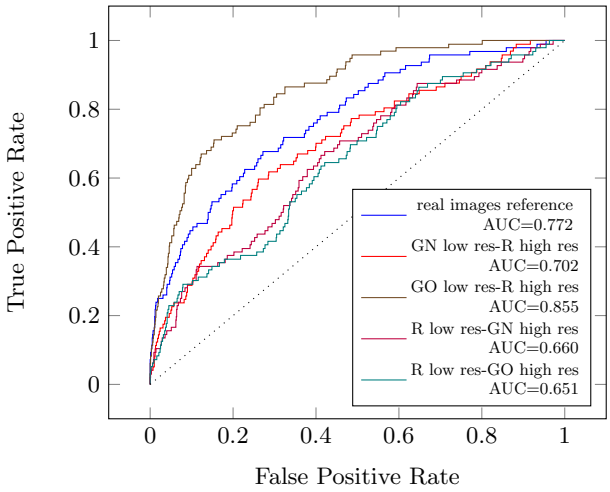


Figure 11: A comparison of different combinations of expressions and domains for c3a3 images. R indicates a real reference photo and G a generated photo. For the generated images N indicates a neutral expression and O the original expression from the passport photo.

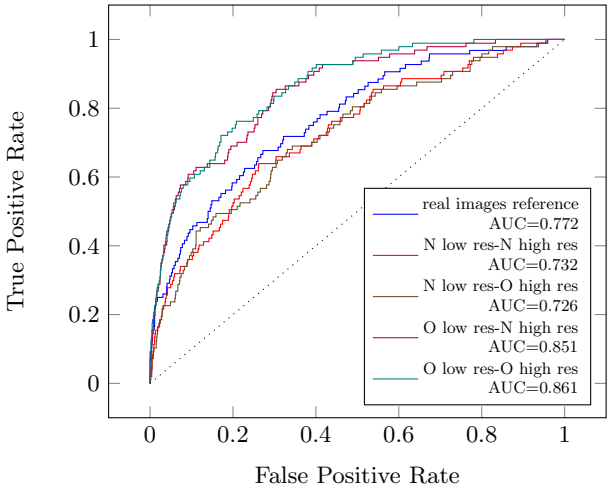


Figure 12: A comparison of different expression combinations for c3a3 images at 10x down sampling. N indicates a neutral expression and O the original expression from the passport photo.

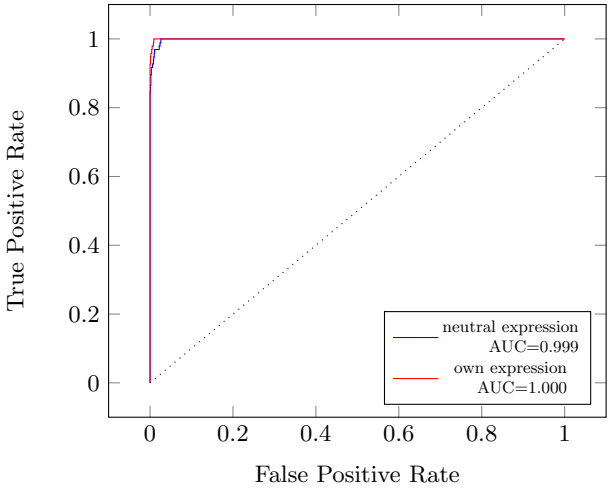


Figure 13: ROC curves for the similarity of high resolution generated images and the references they were generated from.

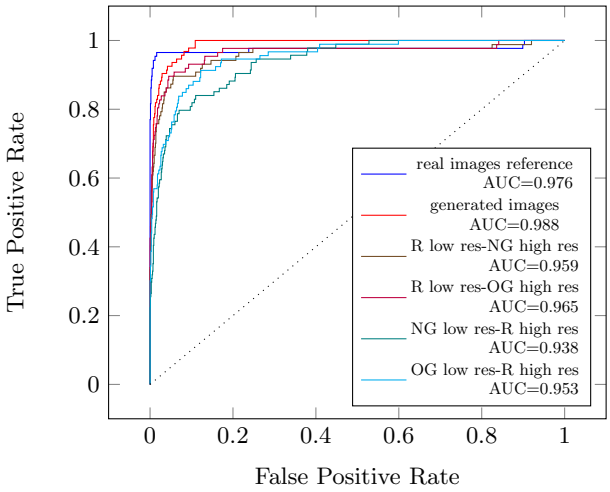


Figure 14: A comparison of different combinations of expression and domain for c3a3 images. R indicates a real reference photo and G a generated photo. For the generated images N indicates a neutral expression and O the original expression from the passport photo

The high variance between the different sets illustrates the big difference any small adjustment can make in such low quality images, which makes it difficult to get a similar level of degradation as in a reference image. This also makes it difficult to tell how much of a factor the fact that the faces are 3D models is, though FaceNet generally performs about as well or even better on them than on just real faces at the same resolution.

Inter-Domain

Generated faces do not seem inherently problematic for FaceNet, so we can also measure whether it sees the same person in a generated image as in the reference image it was generated from. For this purpose the same image sets were compared again, but this time real and generated faces against each other.
When comparing the generated and real faces under ideal conditions, namely a high resolution and frontal lighting and camera alignment, FaceNet has no trouble recognising the subjects of the images as the same person, as can be seen in fig. 13. When using the same expression the separation of identities is almost perfect and using the neutral expression only causes confusion for a few subjects.
On the c1a7 set comparing between domains decreases the performance by up to 5% with the generated low resolution images performing the worst. While the comparisons within the domains mostly had a few subjects that were hard to identify, with a large amount of true positives without false negatives, the curves for the generated low resolution images indicate some confusion happens for a much larger set of faces. Overall this means that the distinction between identities is less clear when comparing between domains for the c1a7 images.
The curves for the c1a3 images get even closer to the chance line, with generated images with the original expression once again outperforming even the real images by 10%. The lowest scoring comparisons are those were the real low resolution images are used, performing 16% and 15% worse. This may be because the real faces have sharp self-shadowing and highlighting, accentuating features that are not present in the smoothed out generated faces.

8. CONCLUSIONS

Using morphable models to generate faces from reference images and then posing and rendering those faces with the techniques supported by Blender allows us to recreate images like those seen in security camera footage. Recreating very low quality images still proves challenging, as it can be hard to tell how the original images were degraded, but most forensic quality degradation can be mimicked.

The state of the art FaceNet face recognition network seems to perform quite well on the synthesised faces, but it is difficult to compare its performance on very low quality images, since those produce varying results. Recognising the generated faces as the same person that they were generated from works well under ideal circumstances. If the conditions are not ideal performance can vary quite a bit, but it is generally worse than when comparing identities within the real or generated domain. This is likely caused by limitations in the used morphable model, as it smooths out important details and sharper facial features. Especially in very low resolution images facial expressions also seem to have a great influence, though the cause of this is unclear.

Future Work

Comparisons between real and generated faces show that the 3D models created from reference faces can still be improved quite a bit. Further work could look at alternative methods for generating these models, or step away from reference images and synthesise the faces completely.
Further research could also look at the influence of facial expression at low resolutions on facial recognition systems, FaceNet specifically, as it is unclear why changing the expression of generated faces caused such a variance in the recognition performance.
While this paper looked at efficacy of using synthesised face models to test face recognition systems under non-ideal conditions further work could research the influence of training on such models, since the large increase in the number of training images could have a large influence on the performance on low quality photos.

References

[1] T. Ali. Biometric Score Calibration for Forensic Face Recognition. June 2014.

[2] Y. Deng, J. Yang, S. Xu, D. Chen, Y. Jia, and X. Tong. Accurate 3D Face Reconstruction with Weakly-Supervised Learning: From Single Image to Image Set. *arXiv:1903.08527 [cs]*, Apr. 2020. arXiv: 1903.08527.

[3] B. Egger, W. A. P. Smith, A. Tewari, S. Wuhrer, M. Zollhoefer, T. Beeler, F. Bernard, T. Bolkart, A. Kortylewski, S. Romdhani, C. Theobalt, V. Blanz, and T. Vetter. 3D Morphable Face Models – Past, Present and Future. *arXiv:1909.01815 [cs]*, Apr. 2020. arXiv: 1909.01815.

[4] B. Gecer, A. Lattas, S. Ploumpis, J. Deng, A. Papaioannou, S. Moschoglou, and S. Zafeiriou. Synthesizing Coupled 3D Face Modalities by Trunk-Branch Generative Adversarial Networks. *arXiv:1909.02215 [cs]*, Sept. 2020. arXiv: 1909.02215.

[5] M. Grgic, K. Delac, and S. Grgic. SCface – surveillance cameras face database. *Multimedia Tools and Applications*, 51(3):863–879, Feb. 2011.

[6] P. J. Grother, M. L. Ngan, and K. K. Hanaoka. Ongoing Face Recognition Vendor Test (FRVT) Part 2: Identification. Nov. 2018. Last Modified: 2018-11-27T15:11-05:00.

[7] G. Guo and N. Zhang. A survey on deep learning based face recognition. *Computer Vision and Image Understanding*, 189:102805, Dec. 2019.

[8] Y. Guo, j. zhang, J. Cai, B. Jiang, and J. Zheng. CNN-Based Real-Time Dense Face Reconstruction with Inverse-Rendered Photo-Realistic Face Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(6):1294–1307, June 2019. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

[9] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. pages 770–778, 2016.

[10] A. Inc. Electronic device having a vision system assembly held by a self-aligning bracket assembly, 2018. U.S. Patent 15/914,956.

[11] A. S. Jackson, A. Bulat, V. Argyriou, and G. Tzimiropoulos. Large Pose 3D Face Reconstruction from a Single Image via Direct Volumetric CNN Regression. *arXiv:1703.07834 [cs]*, Sept. 2017. arXiv: 1703.07834.

[12] M. Jacquet and C. Champod. Automated face recognition in forensic science: Review and perspectives. *Forensic Science International*, 307:110124, Feb. 2020.

[13] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila. Analyzing and Improving the Image Quality of StyleGAN. *arXiv:1912.04958 [cs, eess, stat]*, Mar. 2020. arXiv: 1912.04958.

[14] A. Kortylewski, A. Schneider, T. Gerig, B. Egger, A. Morel-Forster, and T. Vetter. Training Deep Face Recognition Systems with Synthetic Data. *arXiv:1802.05891 [cs]*, Feb. 2018. arXiv: 1802.05891 version: 1.

[15] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang. Interactive Facial Feature Localization. In A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, editors, *Computer Vision – ECCV 2012*, Lecture Notes in Computer Science, pages 679–692, Berlin, Heidelberg, 2012. Springer.

[16] E. Learned-Miller, G. B. Huang, A. RoyChowdhury, H. Li, and G. Hua. Labeled Faces in the Wild: A Survey. In M. Kawulok, M. E. Celebi, and B. Smolka, editors, *Advances in Face Detection and Facial Image Analysis*, pages 189–248. Springer International Publishing, Cham, 2016.

[17] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep Learning Face Attributes in the Wild. *arXiv:1411.7766 [cs]*, Sept. 2015. arXiv: 1411.7766.

[18] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter. A 3D Face Model for Pose and Illumination Invariant Face Recognition. In *2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 296–301, Genova, Italy, Sept. 2009. IEEE.

[19] S. Pidhorskyi, D. Adjeroh, and G. Doretto. Adversarial Latent Autoencoders. *arXiv:2004.04467 [cs]*, Apr. 2020. arXiv: 2004.04467.

[20] R. Ranjan, S. Sankaranarayanan, A. Bansal, N. Bodla, J.-C. Chen, V. Patel, C. Castillo, and R. Chellappa. Deep Learning for Understanding Faces: Machines May Be Just as Good, or Better, than Humans. *IEEE Signal Processing Magazine*, 35(1):66–83, 2018.

[21] J. Riviere, P. Gotardo, D. Bradley, A. Ghosh, and T. Beeler. Single-shot high-quality facial geometry and skin appearance capture. *ACM Transactions on Graphics*, 39(4):81:81:1–81:81:12, July 2020.

[22] F. Schroff, D. Kalenichenko, and J. Philbin. FaceNet: A Unified Embedding for Face Recognition and Clustering. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, June 2015. arXiv: 1503.03832.

[23] R. Slossberg, G. Shamai, and R. Kimmel. High Quality Facial Surface and Texture Synthesis via Generative Adversarial Networks. In L. Leal-Taixé and S. Roth, editors, *Computer Vision – ECCV 2018 Workshops*, volume 11131, pages 498–513. Springer International Publishing, Cham, 2019. Series Title: Lecture Notes in Computer Science.

[24] B. Steinert, H. Dammertz, J. Hanika, and H. P. A. Lensch. General Spectral Camera Lens Simulation. *Computer Graphics Forum*, 30(6):1643–1654, 2011. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-8659.2011.01851.x.

[25] L. Tran, F. Liu, and X. Liu. Towards High-Fidelity Nonlinear 3D Face Morphable Model. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1126–1135, Long Beach, CA, USA, June 2019. IEEE.

[26] C. G. Zeinstra, D. Meuwly, A. C. C. Ruifrok, R. N. J. Veldhuis, and L. J. Spreeuwers. Forensic Face Recognition as a Means to Determine Strength of Evidence: A Survey. page 13, 2018.

[27] C. G. Zeinstra, R. N. J. Veldhuis, L. J. Spreeuwers, A. C. C. Ruifrok, and D. Meuwly. ForenFace: a unique annotated forensic facial image dataset and toolset. *IET Biometrics*, 6(6):487–494, 2017. Conference Name: IET Biometrics.

[28] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint Face Detection and Alignment using Multi-task Cascaded Convolutional Networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, Oct. 2016. arXiv: 1604.02878.

[29] X. C. Zhang, J. T. Barron, Y.-T. Tsai, R. Pandey, X. Zhang, R. Ng, and D. E. Jacobs. Portrait shadow manipulation. *ACM Transactions on Graphics*, 39(4):78:78:1–78:78:14, July 2020.

[30] M. Zollhöfer, J. Thies, P. Garrido, D. Bradley, T. Beeler, P. Pérez, M. Stamminger, M. Nießner, and C. Theobalt. State of the Art on Monocular 3D Face Reconstruction, Tracking, and Applications. *Computer Graphics Forum*, 37(2):523–550, 2018. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.13382.