Sketch-to-Photo Matching Enforcing Realistic Rendering Generation

Leonardo Capozzi[1,2]
leonardo.g.capozzi@inesctec.pt
João Ribeiro Pinto[1,2]
joao.t.pinto@inesctec.pt
Jaime S. Cardoso[1,2]
jaime.cardoso@inesctec.pt
Ana Rebelo[2]
arebelo@inesctec.pt

[1] Faculdade de Engenharia
Universidade do Porto
Porto, Portugal
[2] INESC TEC
Porto, Portugal

Abstract

The use of forensic sketches to locate suspects is a challenging task. These sketches are posted on public spaces, social media and the news with the hope that someone recognizes the suspect. Recent methods present some limitations, as they do not use end-to-end networks or/and do not offer other alternatives in case the matching process fails, such as providing a photo-realistic representation of the sketch. This paper presents a method that combines a conditional generative adversarial network (cGAN) and a pre-trained face recognition network, optimised as an end-to-end model. This method is able to retrieve a list of potential suspects, and it simultaneously provides an intermediate realistic representation of the sketch. Evaluation on the CUFS and CUFSF databases shows that the proposed method outperforms state-of-the-art methodologies in most tasks, and that forcing a photo-realistic rendering of the sketch only results in a slight performance decrease.

1 Introduction

Over the years, the use of deep learning has brought a lot of advancements in pattern recognition and computer vision tasks, such as face recognition. Recent methodologies report significantly higher accuracy in the matching process when using deep learning [8, 10]. However, the use of real photos in face recognition is much easier than the use of a forensic sketch, since a forensic sketch might not be a very accurate representation of the suspect, as it was drawn using descriptions from eye witnesses [9].

Recent state-of-the-art methods have used convolutional neural networks (CNN) to perform sketch-to-face matching [2, 3, 5, 7], however many of these do not use end-to-end approaches, which can limit the performance of the model by causing dissonance between separately optimised blocks.

Other methodologies include the generation of a photo-realistic representation of the sketch, which can be useful for manual identification in case the matching process fails. They use adversarial approaches based on CycleGANs [5] and cGANs [2, 7].

This work tackles two important aspects that have not been addressed in the literature. The first aspect is the use of an end-to-end model which is jointly optimised, avoiding the performance limitations associated with the use of separate processes. The second aspect is enforcing the end-to-end model to generate an intermediate photo-realistic representation of the input sketch that can be used by law enforcement in manual matching processes.

The proposed methodology is composed of a cGAN and a matching CNN that are optimised in an end-to-end fashion. When trained, the model receives a sketch and returns a feature vector that can be used for matching using simple distance metrics. The model also generates an intermediate latent representation which is realistic and similar to the corresponding real photographs.

2 Proposed Methodology

The proposed methodology is composed of two main parts, which are jointly optimised: a sketch-to-render generator and a matching network (see Fig. 1). The sketch-to-render generator receives a sketch and outputs a photo-realistic representation of the sketch that is similar to the real face of the person. The matching network receives the intermediate photo-realistic representation and outputs a feature vector that can be used for the matching process.

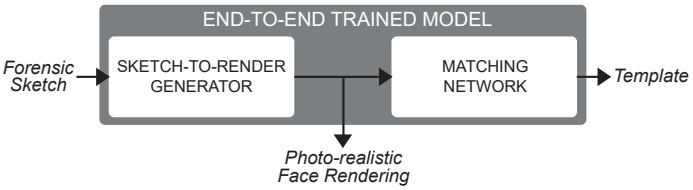

Figure 1: Overview of the proposed methodology (Taken from [1]).

2.1 Network Architecture

The generator consists of a U-Net, which is an encoder-decoder with skip connections that enable the transmission of information between corresponding levels of the network. It receives a sketch and outputs a photo-realistic rendering of the sketch [1].

The discriminator used in this work is a CNN adapted from the cGAN of pix2pix [4]. It receives as input the photo-realistic rendering of the sketch outputted by the generator and the corresponding sketch, and outputs a prediction on whether the photo-realistic rendering is a real image or generated image [1].

The matching network used in this work is a VGG-16 with pre-trained weights from VGG-Face [8]. It receives as input the photo-realistic rendering of the sketch outputted by the generator and outputs a feature vector. The sketch is matched to real photos from a database by computing the cosine distance between the respective feature vectors.

2.2 Loss

The loss function used for training is composed of several losses. The first component comes from the loss of the cGAN, and it is responsible for generating photo-realistic images. It can be written as the following:

$$\mathcal{L}_{cGAN}(G,D) = \mathbb{E}_{x,y}[\log D(x,y)] + \mathbb{E}_{x}[\log(1 - D(x,G(x)))], \quad (1)$$

where x is a sketch and y is the corresponding ground-truth photo. The generator (G) tries to minimize the loss and the discriminator (D) tries to maximize it.

The generator should also try to mimic the real photograph of the person in the sketch, therefore we add a second loss term:

$$\mathcal{L}_{L1}(G) = \mathbb{E}_{x,y}[||y - G(x)||_1], \quad (2)$$

which minimizes the $L1$ norm of the difference between the real image (y) and the generated image ($G(x)$).

The identity of the generated realistic rendering needs to match the identity of the ground-truth image. Hence, we add a third component to the loss function:

$$\mathcal{L}_{match}(G) = \mathbb{E}_{x,y}[||V(y) - V(G(x))||_2]. \quad (3)$$

The weights of the VGG-Face network (V) are frozen, since we are using the pretrained weights. The matching loss only adjusts the weights of the generator.

Combining all the loss components, the final loss function becomes:

$$\mathcal{L}(G,D) = \min_G \max_D \mathcal{L}_{cGAN}(G,D) + \lambda_1 \mathcal{L}_{L1}(G) + \lambda_2 \mathcal{L}_{match}(G). \quad (4)$$

Table 1: Matching accuracy on the CUFSF and CUFS datasets, using different methods to enhance the sketch (r.r.g.: realistic rendering generation) (Taken from [1]).

Method	CUFSF			CUFS		
	R-1	*R-5*	*R-10*	*R-1*	*R-5*	*R-10*
Sketch	49	77	88	47	80	90
pix2pix	53	83	92	52	79	86
IPMFSPS [6]	**74**	94	97	**80**	**95**	**97**
HFFS2PS [2]	-	-	-	36	69	-
Proposed (with r.r.g)	54	87	96	44	77	86
Proposed (without r.r.g)	**74**	**98**	**99**	59	85	91

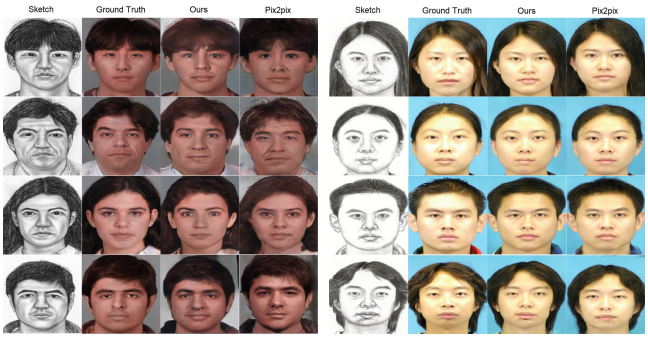

Figure 2: Images generated by our method using the CUFSF dataset (on the left); Images generated by our method using the CUFS dataset (on the right). (Taken from [1]).

3 Experimental Settings

The proposed methodology was trained using sketch-photo pairs from the CUHK Face Sketch database (CUFS) [11] and the CUHK Face Sketch FERET dataset (CUFSF) [11, 12].

The photos of CUFSF from the FERET database were colorized using the DeOldify API (Available on: https://github.com/jantic/DeOldify). This was done to uniformize the datasets and to allow the model to generate color images.

The sketches and photos were transformed so that the position of the eyes was consistent in each sketch-photo pair, in order to improve the quality of the generated images and the accuracy of the matching process.

The model was trained on two experimental settings. The first one used the previously mentioned loss function, in order to generate intermediate realistic renderings and to have a high matching accuracy. The second setting measured the impact that removing the loss terms $\mathcal{L}_{cGAN}$ and $\mathcal{L}_{L1}$, which promote the generation of a realistic image, had on the matching accuracy. For more details refer to [1].

4 Results and Discussion

To measure the performance of the matching process we computed the rank-N accuracy (R-N) on the test sets of CUFS and CUFSF. In this case, we consider $N \in \{1, 5, 10\}$.

The results in Table 1 show that the proposed method is superior or aligned to the alternative methods on all considered ranks. However, when the photo-realistic rendering is dismissed, the matching accuracy increases significantly, showing a trade-off between matching performance and the realism of the rendering, which could be tuned for specific scenarios.

Examples of photo-realistic renderings, the corresponding sketches, ground-truth photos and images generated using the pix2pix method can be seen in Fig. 2. Visually, we can confirm that the generated renderings look realistic, and similar to the ground truth photos. Considering the performance of the model, maintaining the photo-realism is a positive aspect of the proposed methodology.

5 Conclusion

This paper proposes an end-to-end-method for sketch-to-photo matching that enforces a photo-realistic rendering of the input sketch. Upon evaluation, the matching process showed that the proposed method is superior or aligned to state-of-the-art methodologies, and the generation of intermediate face renderings offered realistic results. The matching results improved when disregarding realistic face renderings, showing a trade-off between matching accuracy and the realism of the generated image. Further efforts should be devoted to improve the realistic rendering generation, in order to allow for a more diverse range of face characteristics, such as hair, eyes, and skin color. We believe these efforts would improve the results of both the realistic renderings and the matching process.

Acknowledgements

This work was partially funded by the Project TAMI - Transparent Artificial Medical Intelligence (NORTE-01-0247-FEDER-045905) financed by ERDF - European Regional Fund through the North Portugal Regional Operational Program - NORTE 2020 and by the Portuguese Foundation for Science and Technology - FCT under the CMU - Portugal International Partnership, and within the PhD grants "SFRH/BD/137720/2018" and "2021.06945.BD". Portions of the research in this paper use the FERET database of facial images collected under the FERET program, sponsored by the DOD Counterdrug Technology Development Program Office.

References

[1] Leonardo Capozzi, Jaime S. Cardoso, Ana Rebelo, and Joao Pinto. End-to-end deep sketch-to-photo matching enforcing realistic photo generation. In *Proceedings of the IEEE International Symposium on Biomedical Imaging (CIARP'21)*, 2021.

[2] W. Chao, L. Chang, X. Wang, J. Cheng, X. Deng, and F. Duan. High-fidelity face sketch-to-photo synthesis using generative adversarial network. In *ICIP*, pages 4699–4703, 2019.

[3] S. M. Iranmanesh, H. Kazemi, S. Soleymani, A. Dabouei, and N. M. Nasrabadi. Deep sketch-photo face recognition assisted by facial attributes. In *IEEE BTAS*, pages 1–10, 2018.

[4] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017.

[5] H. Kazemi, M. Iranmanesh, A. Dabouei, S. Soleymani, and N. M. Nasrabadi. Facial attributes guided deep sketch-to-photo synthesis. In *WACVW*, pages 1–8, 2018.

[6] Y. Lin, S. Ling, K. Fu, and P. Cheng. An identity-preserved model for face sketch-photo synthesis. *IEEE Signal Processing Letters*, 27: 1095–1099, 2020.

[7] Uche Osahor, Hadi Kazemi, Ali Dabouei, and Nasser Nasrabadi. Quality guided sketch-to-photo image synthesis. *arXiv*, 2020. 2005.02133.

[8] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *British Machine Vision Conference*, 2015.

[9] Sourav Pramanik and Dr. Debotosh Bhattacharjee. An approach: Modality reduction and face-sketch recognition. *arXiv*, 2013.

[10] Mei Wang and Weihong Deng. Deep face recognition: A survey. *arXiv*, 2018.

[11] X. Wang and X. Tang. Face photo-sketch synthesis and recognition. In *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, volume 31. IEEE, 2009.

[12] W. Zhang, X. Wang, and X. Tang. Coupled information-theoretic encoding for face photo-sketch recognition. In *CVPR*, 2011.