Semantic Segmentation Based Automatic Two-Tone Portrait Synthesis

Zhuoqi Ma[1], Nannan Wang[2](✉), Xinbo Gao[1], and Jie Li[1]

[1] State Key Laboratory of Integrated Services Networks, School of Electronic Engineering, Xidian University, Xi'an 710071, China
[2] State Key Laboratory of Integrated Services Networks,
School of Telecommunications Engineering, Xidian University, Xi'an 710071, China
nnwang@xidian.edu.cn

Abstract. This paper presents a semantic segmentation based method for automatically synthesizing two-tone cartoon portraits in black-and-white style. Synthesizing two-tone portraits from photographs can be considered as a heterogeneous image transformation problem, of which the result should be vivid portraits with distinct freehand-like features, such as clean backgrounds and continuous lines. To achieve this goal, our system connects two separate subsystems together, namely semantic segmentation and portrait synthesis. In the semantic segmentation phase, photographs are segmented into background, hair and skin regions using multiple segmentations method. In the portrait synthesis phase, we treat different regions with different strategies. Experimental results demonstrate that our system can precisely segment the input photo and produce visually desired two-tone portraits.

Keywords: Semantic segmentation · Two-tone portraits
Component based modeling

1 Introduction

Two-tone portrait synthesis, as an artistic branch of face hallucination, has been favored by more and more people in social life or art craft manufacturing. For example, people would prefer characterized facial portraits as virtual appearances for online chatting or to be printed out on cards, clothes or cups as gifts for their friends, which would require the portrait to be two-tone not only for visual effects but also for the practicability to print it out. However, manually rendering such portraits usually requires professional skills and would cost certain amount of time, which is unrealistic for applications in social media with millions of users or craft manufacturing industry that requires automation and efficiency. Thus, it is of great use to have automatic systems for synthesizing two-tone portraits.

General face stylization strategies can be broadly categorized into three types: image processing, example-based synthesis, and deep learning based face stylization. Image processing based face hallucinations [6] treat the image as a whole

Y. Zhao et al. (Eds.): ICIG 2017, Part III, LNCS 10668, pp. 170–181, 2017.
https://doi.org/10.1007/978-3-319-71598-8_16

processing unit and would easily cause noise or lose some important details in the output. Example-based approaches could be further categorized into local patch based methods and component based models. Local patch based methods [21] can generate shading sketches with high quality and resemblance. However, these methods divide the face images into patches, which ignored the inherent structure consistency of the photo, thus would cause blurring effect in the result. Component based models are mostly used to synthesis portraits with specific artistic styles, like cartoons [20], line drawings [3] or paper-cuts [9]. Even though they are not affected by the aforementioned drawbacks, their over dependence on the templates make the results lose resemblance easily, especially for hair region. Apart from these, many headshot style transfer methods [12,18] based on deep neural networks have attracted much attention these days. These methods can handle multiple painting styles using same network, however, their methods can not be adopted to our problem because the two-tone portraits require clear distinct lines and parts.

For a two-tone portrait synthesis system of which the main application is in art craft manufacturing or entertainment, the rendered portrait should resemble the input photo and meanwhile maintain its desired and attractive appearance. Face photo actually has very clear composition and structure, so different regions should be dealt with respective methods when it comes to synthesize a portrait from a photo. Recently, some hair and face labeling approaches [2] have made significant progress. Different from the traditional KNN matting used in some portrait synthesis methods, which is neither automatic or comprehensive, they can label background, hair, skin even clothes at one time.

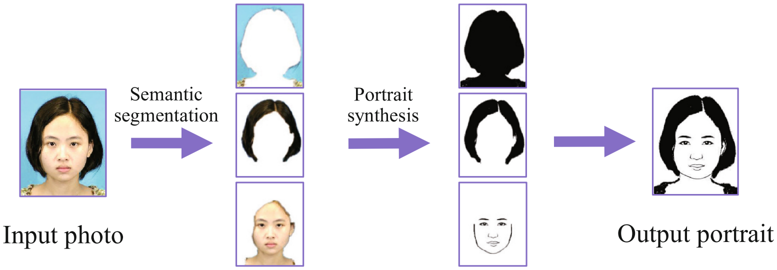

Fig. 1. Overview of our system.

In this paper, we present a segmentation based system for automatically synthesizing two-tone portraits of input photos. As show in Fig. 1, our system could be divided into two parts, the semantic segmentation part and the portrait synthesis part. We illustrate the high segmentation accuracy on the labeled CUHK dataset. Experimental results show that our system can render two-tone freehand-like portraits with desired visual appearances.

The main contributions are twofold. (1) Semantic segmentation is introduced to deal with the face portrait synthesis problem, which take the internal structure consistency of the photo into consideration and enable different processing methods on one photo. (2) A global model for synthesizing component templates balancing the resemblance and attractiveness.

2 Related Work

Creating digital artistic portraits have drawn growing attention in computer vision and graphics. Examples include human facial illustrations [6], portrait generation [3] and face style transfer [12]. More related to our work, however, are researches that focus on generating portraits with two-tone style and rendering facial appearances from photos. We briefly go through the existing portraits rendering methods of the aforementioned two categories: image-based methods and example-based methods.

Image-based portraits rendering methods. A lot of scientific research works and commercial products have been proposed and made with image-based methods to generate two-tone portraits or sketches. For instance, Gooch et al. [6] use brightness and luminance thresholding methods to create black-and-white facial illustrations from photos. Gastal et al. [5] use edge-aware filters to transform a photo into a pencil drawing. These image-based methods can create portraits only with a little computation and maintain high resemblance with the original photos. As straightforward and simple as it is, the holistic handling of the input photo would cause noise in the output, especially in the background and skin region. Additionally, this method could lose some important details, such as face contour, which would not be neglected by human artists.

Example-based portraits synthesis methods. Due to the drawbacks of rendering portraits via image-based approaches, many example-based methods were proposed in terms of their ability to handle styles that are difficult to represent using image-based methods. These methods can be further categorized as shading sketch synthesis, profile sketch synthesis and other special artistic stylization strategies.

Shading sketch synthesis began with global methods. Tang and Wang [13] proposed global eigen-transformation based method. However, it is difficult to obtain clear face sketches via these methods when they regard the photo as a whole. In view of these shortcomings, many patch based methods were proposed for their ability to express the local features of the face [14–16, 19, 21]. Although these methods can generate shading sketches that closely match the input photo, they ignore the inherent structure of photos. Furthermore, it is difficult for their frameworks to generate two-tone style portraits with freehand-like features.

Profile sketch synthesis methods mainly handle strokes to form a line drawing. Chen et al. [3] introduced a interactive computer system with two subsystems to compose face components and hair sketch in a Japanese cartoon style. Some works based on and-or graph [17] also caught certain attention in sketch synthesis field. Nevertheless, these methods depend on templates too much especially for hair regions, and their segmentation strategy can not reach full automation.

Furthermore, the profile sketches has hollow hair, which should be painted black in real freehand portraits.

Various strategies have been studied to render portraits with special artistic styles. Meng et al. [9] created artistic paper-cut of human portraits by binarization and templates matching methods. Rosin and Lai [11] generated cartoon portraits in puppet style and in the style of Julian Opie using non-photorealistic rendering pipeline and fitting facial models. These works were specifically designed according to their objective artistic styles so their strategies can not be directly applied to our task.

3 Semantic Segmentation for Region Labeling

In this section, we introduce the segmentation part of our whole system as shown in Fig. 2, which is the pre-processing stage for next synthesis phase. We present an initial coarse segmentation method based on multiple segmentation [2] and then a refining process by image matting [8] to precisely label the background (and clothes), hair and skin regions of the photo.

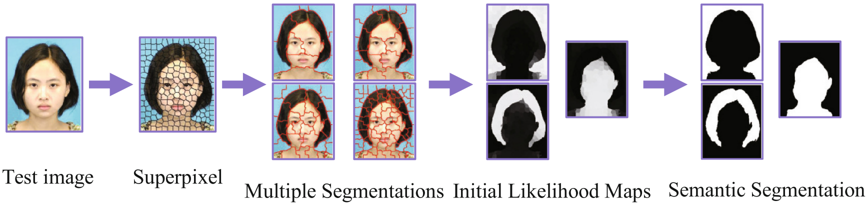

Fig. 2. Pipeline of the segmentation method.

3.1 Initial Coarse Segmentation

We employ the label probability estimation method [2] to perform initial segmentation. In training phase, the SLIC algorithm [1] is employed to partition the training photo into superpixels. Then we compute multiple segmentations by semi-supervised spectral clustering scheme [2]. After the multiple segmentations were generated, we need to evaluate the homogeneity of each segments and estimate its label likelihood. We train two classifiers, the segment label classifier and the homogeneity classifier. The segment label classifier outputs the class(background, hair and skin) likelihood for segments and the homogeneity classifier estimates the unity of the labels of the contained superpixels.

In test phase, for a given image, we first generate multiple segmentations m_j as aforementioned, and then extract features from each segment. The segment label likelihood $P(\tilde{y}_{ij} = k\,|I, m_j\,)$ and its homogeneity likelihood $P(m_j|I)$ are estimated using the trained label classifier and homogeneity classifier. The label likelihood for the i_{th} superpixel is obtained through the combination of segments label likelihood and homogeneity likelihood over all levels of multiple segmentations:

$$P(y_i = k \,|I\,) = \sum_{j}^{n_s} P(\tilde{y}_{ij} = k \,|I, m_j\,) \times P(m_j \,|I\,), \tag{1}$$

where y_i and $\tilde{y}_{ij}$ denote the label of the i_{th} superpixel and the label of the segment in j_{th} multiple segmentation level containing the i_{th} superpixel. I represents the image, k is the class labels (background, hair and skin in this work). n_s denotes the number of multiple segmentations with different levels, here n_s is set to 15 ($n_s \in \{5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 60, 70, 80, 90, 100\}$). If the segments number is too small in one image, it is possible that every segment has low homogeneity and include them into computation would not have positive effect for label estimation.

3.2 Improve Segmentation Accuracy by Image Matting

The initial likelihood map generated from the former phase is not accurate enough to be applied to the synthesis stage, so we use image matting method to refine the segmentation result. Given the initial likelihood maps, we first choose the superpixels with high label probabilities for each label:

$$\begin{array}{ll} i & P\,(y_i = hair) > \theta_h \\ j & P\,(y_j = background) > \theta_{bg} \\ k & P\,(y_k = skin) > \theta_s \end{array} \tag{2}$$

here, θ_h, θ_{bg} and θ_s denotes the threshold values for hair, background, and skin respectively and are set to 0.8, 0.9 and 0.9 in our work. After these high confidence regions are selected, we use white scribbles to indicate foreground region, and black scribbles background region. As the methods are similar, we use hair region labeling as an example. The inscribed rectangles of the superpixels whose hair label likelihood $P\,(y_i = hair)$ surpasses the threshold value θ_h in set i are selected as the foreground regions. The inscribed rectangles of the superpixels from the other two sets are selected as the background regions. Then the fine hair segmentation result is obtained via employing automatic alpha matting [8]. The background and face region can be segmented in the same way by simply set the superpixels with high confidence of the target label as the foreground and others as the background.

4 Two-Tone Portrait Synthesis

This section is a detailed introduction of our two-tone portrait synthesis method. After the background, face and hair regions have been segmented, we handle different regions with different strategies and combine them together to form the final portrait.

Background region. Background region includes not only the shooting scene but also the clothes. In our target style, portraits should have clean and white background, and the clothes region should be painted in black and white colors to depict its patterns. We first compute the per-plane composite gradient of the background region to obtain the edge map. Taking the two-tone style into consideration, we then take inversion of the edge map to generate a sketch in which deep colors denotes the edge of the image. We further enhance the sketch to darken the lines and colors using histogram equalization method. Then we use dynamic thresholding method to obtain the binary image.

Hair region. Hair region is one of the most identifying part in the automatic synthesis of portrait, so the detection and representation of hair have great influence on resemblance. As hair region only contains one component and artists normally paint hair in black in two-tone style portraits, we simply binarize the extracted RGB hair region image. However, the binarized image usually comes with sawtooth effect around the edge, which is highly unlikely to appear in a freehand portrait. In order to alleviate this noise, we use gaussian filter to smooth the image and then posterize the image with specific levels to create fluidity for the edge.

Face region. Face has very regular yet complicate structure, so we can not directly use binarization strategy to produce two-tone portrait as it would cause severe noise and lose some details, such as lines depicting chins and cheeks. For synthesis of face region, our method proceeds in the following steps: firstly, the face region is decomposed into components; then, these components are matched to the templates in the dictionary according to respective features; finally, we use a global model to organize the composition of the components.

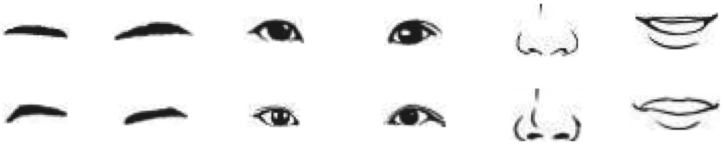

Fig. 3. Example two-tone facial components templates

Two-tone templates. We have asked several professional artists to create black-and-white portraits for some photos using tablets, then manually decompose them into facial components, as shown in Fig. 3. Keypoints of these components are also manually labeled on the portraits. We then construct a dictionary including the example photo components and the corresponding templates pairs. Templates are usually drawn based on various distinct photos for a greater power of representation.

A. Face decomposition

After face region is segmented, we first use the active appearance model (AAM) [4] to localize the facial structure keypoints. These facial landmarks are used to extract face regions of eyebrows, eyes, nose, mouth and face contour. These components are clipped from the image in a rectangular shape, while face contour is represented by the polygon defined from the landmarks.

B. Component matching

For each component extracted from the earlier step, we search the template-component dictionary for the most similar realistic photo components. Since different facial components exhibit diverse visual features, we handle each kind of component with different matching strategies. After matching the input components with the dictionary components, we use their corresponding two-tone templates for the subsequent composition step.

Eyebrows, Eyes and Nose. These components could be differentiated from each other in the perspective of shape and texture. The shape feature is represented using the length, width of the bounding box and its ratio. For eyebrows and eyes, the texture feature is obtained via using SIFT feature descriptor at the facial landmarks. For nose, we use RGB-SIFT descriptor to describe the color pattern change in this place. We denote the shape vector as $S = [s_{lbrow}, s_{rbrow}, s_{nose}, \cdots]$, the texture vector as $T = [t_{lbrow}, t_{rbrow}, t_{nose}, \cdots]$. For a component of the input photo c and one of the dictionary c', we can define the distance function of their appearances as

$$d\left(c, c'\right) = \lambda \left\|S - S'\right\| + (1 - \lambda) \left\|T - T'\right\| \tag{3}$$

where λ is the weight for shape and texture when measuring different components.

After we compute the shape vector S and the texture vector T for different components, we then find the most alike candidate from the template dictionary by minimizing the distance:

$$\min_j d\left(c, {c_j}'\right) \tag{4}$$

where ${c_j}'$ is the jth component example in the template-component dictionary.

Mouth. The appearance of mouth in portrait is largely determined by the subject's expression. The mouth templates in our dictionary could be classified into 2 primary classes: open and shut, and then be subdivided into 4 classes: laugh and no expression under open category, smile and no expression under shut category. We train a two layer classifier to determine the expression on the subject's face. Then we use feature matching to match the component to template in corresponding categories.

Face contour. The shape of face contour can be perfectly represented simply by its landmarks. Besides shape, the texture feature should also be considered to distinguish single-chin from double-chin. The Hausdorff distance $d_H\left(V_I, V_D\right)$ between the input face contour landmarks V_I the components' face contour landmarks V_D from the dictionary are used to measure the shape dissimilarity.

$$d_H\left(V_I, V_D\right) = \max\left\{\sup_{x \in V_I} \inf_{y \in V_D} d\left(x, y\right), \sup_{y \in V_D} \inf_{x \in V_I} d\left(x, y\right)\right\} \tag{5}$$

The texture distance $d_T(t_I, t_D)$ are represented using SIFT descriptor, t_I and t_D denotes SIFT feature computed at keypoints for input face contour and the face contour from the dictionary. We then select the most similar face contour component by minimizing the distance:

$$\min d_H(V_I, V_D) + d_T(t_I, t_D) \tag{6}$$

C. Composition of facial component

After obtaining the best matching template for each input component, we compose them together to synthesize the face region for the final output. We use a global model to organize the arrangement of each component on the canvas. Our global model includes a set of control points $\{P_{lbrow}, P_{leye}, P_{rbrow}, P_{nose}, \cdots\}$ for arranging the templates and resizing and warping parameters for templates.

Previous studies have demonstrated that humans have a preference for symmetric faces. In the meantime, tests have revealed that the original slightly asymmetric faces were rated as more attractive than completely symmetric versions [10]. Previous work [3,20] focus on learning the styles that the artists use to arrange the face components, however, here we emphasize more on the balance between high resemblance to the input photos and the attractiveness of the portraits by building symmetry. For a symmetric structure, our model would set up a coordinate system for the face and model the relative placement of the input face components and then adjust them using axis-symmetry protocols to determine where to put the templates. To reach the high resemblance of the original slightly asymmetric face, we warp and resize the templates in accordance with the input components.

D. Additional Component

In compositional models, ears are always neglected. The existence of ears could be detected by examining whether there is a sudden change area of the segmented face region mask, and if there is an outward curve with certain length in edge map, we set the ear template there.

By this step, the face region is synthesized. Then we combine the processed background region and hair region together to synthesize the final output portrait.

5 Experimental Results

In this section, we first show that our segmentation method can reach a very high accuracy on labeled CUHK dataset (Section-A). We then demonstrate that our method can generate vivid two-tone portraits with high resemblance and freehand-like features (Section-B).

A. Semantic segmentation

Datasets. We use the public dataset from Kae et al. [7], consisting of 2927 LFW (funneled) images, and supplement the dataset with manually labeled photos from the CUHK student database (see Fig. 4).

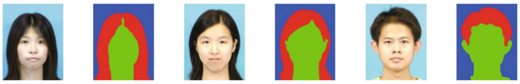

Fig. 4. Labeled CUHK dateset.

Evaluation. The segmentation accuracy of our method tested on LFW (funneled) images has reached 94.25%, compared to 90.7% [7] and 92.69% [2]. The segmentation accuracy tested on CUHK photos with classifiers trained on our manually labeled CUHK student database are over 96%. Figure 5 shows some segmentation results on CUHK dataset. The comparison of automatically labeled results and manually labeled results shows that our method can segment background, hair and skin region accurately for the next synthesis phase.

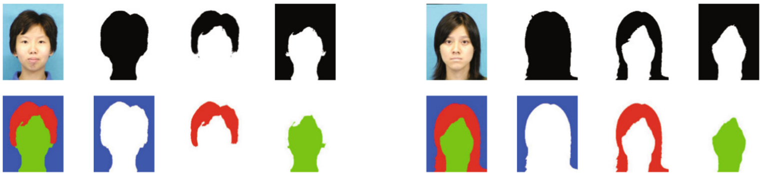

Fig. 5. Segmentation result for labeled CUHK dataset. Top row: segmentation result. Bottom row: manually labeled result.

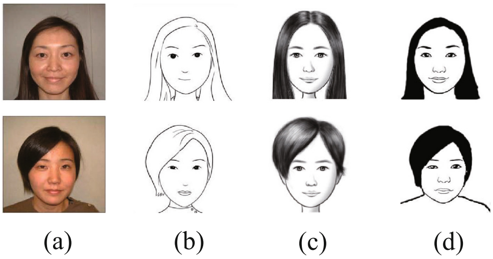

Fig. 6. Comparison with previous methods for sketch and cartoon synthesis. From left to right: (a) input photo; (b) the result of [3]; (c) the result of [20]; (d) our result.

B. Portrait Synthesis

Our system aims at synthesizing two-tone styles portraits with both attractiveness and resemblance. In Fig. 6, we compare with the component-based methods for synthesizing profile sketch and cartoon face. These methods can generate stylized portraits yet not balance the resemblance and attractiveness of the output result. For example, The nose and mouth parts in Fig. 6(b) fail to reflect the distinct feature of the original input photo. Figure 6(c) has very stylistic appearance, but their over dependence on templates causes the lack of similarity in hair region.

In Fig. 7, we compare with the shading sketch synthesis result. We can see that these methods would easily cause noise and blurring in the result, which should not be occurred in freehand portraits.

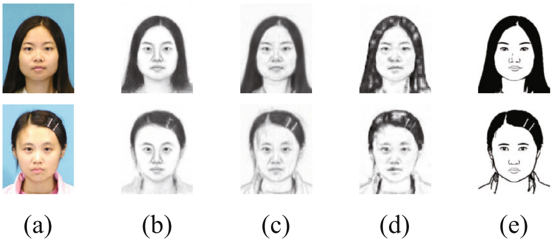

Fig. 7. Comparison with previous methods for shading sketch synthesis. From left to right: (a) input photo; (b) the result of [21]; (c) the result of [18]; (d) the result of [19]; (e) our result.

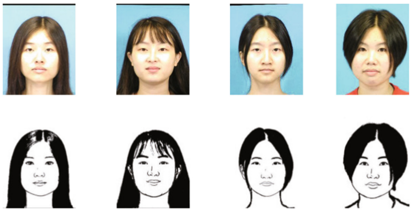

Fig. 8. Two-tone portraits generated by our system.

Figure 8 demonstrates some two-tone portrait results rendered by our system. We can see that our method can render two-tone portraits with distinct freehand features such as clear backgrounds and continuous lines. Also, it can maintain distinct figure features such as hair style, expression and face contour, which guaranteed the resemblance between the input photo and the output synthesis portraits. Third, our global model can synthesize portraits with symmetric facial structure meanwhile preserve the similarity to the original slightly asymmetrical face. More results are available online: http://www.ihitworld.com/ZQM/T2P.zip.

6 Conclusion and Future Work

In this paper, we have presented a framework for synthesizing two-tone portraits based on semantic segmentation. In the segmentation phase, our method based on multiple segmentations and image matting can precisely divide the input photo into background, hair and skin regions. For the after-segmentation synthesis stage, we handle different regions with appropriate methods respectively. For background and hair regions, our processing strategy can render clean and smooth results with freehand-like features. For face region, we select the best match for each input component and use a global model to warp and arrange them to reach the balance between resemblance and attractiveness. Our experimental results show that our system is capable of synthesizing visually desired two-tone portraits with distinct freehand-like features.

The main limitations of this method are the segmentation accuracy and the lack of diversity for component templates. In future work, we plan to improve the segmentation accuracy via existed deep neural networks and enrich the component dictionary with more distinctive templates. We are also trying to extend this after-segmentation synthesis method to other portrait styles, such as shading sketch, by exploring proper strategies to process hair and background region into pencil sketch style. Besides, synthesizing two-tone portraits with complex backgrounds is another worthy topic, for example, for photographs taken at landscapes, people would like to have the complete photo in two-tone style, not only the face part. Lastly, to realize the application of this framework in industrial production, the speed of segmentation is another place to improve in the future.

Acknowledgement. This work was supported in part by the National Natural Science Foundation of China (under Grants 61501339, 61671339, 61432014, U1605252, 61601158, and 61602355), in part by Young Elite Scientists Sponsorship Program by CAST (under Grant 2016QNRC001), in part by Young Talent fund of University Association for Science and Technology in Shaanxi, China, in part by the Fundamental Research Funds for the Central Universities under Grant JB160104, in part by the Program for Changjiang Scholars, in part by the Leading Talent of Technological Innovation of Ten-Thousands Talents Program under Grant CS31117200001, in part by the China Post-Doctoral Science Foundation under Grants 2015M580818 and 2016T90893, and in part by the Shaanxi Province Post-Doctoral Science Foundation.

References

1. Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Süsstrunk, S.: Slic superpixels compared to state-of-the-art superpixel methods. IEEE Trans. Pattern Anal. Mach. Intell. **34**(11), 2274–2282 (2012)
2. Ahn, I., Kim, C.: Face and hair region labeling using semi-supervised spectral clustering-based multiple segmentations. IEEE Trans. Multimedia **18**(7), 1414–1421 (2016)
3. Chen, H., Liu, Z., Rose, C., Xu, Y., Shum, H.Y., Salesin, D.: Example-based composite sketching of human portraits. In: Proceedings of the 3rd International Symposium on Non-Photorealistic Animation and Rendering, pp. 95–153. ACM (2004)
4. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active appearance models. IEEE Trans. Pattern Anal. Mach. Intell. **23**(6), 681–685 (2001)
5. Gastal, E.S., Oliveira, M.M.: Domain transform for edge-aware image and video processing. In: ACM Transactions on Graphics (ToG), vol. 30, p. 69. ACM (2011)
6. Gooch, B., Reinhard, E., Gooch, A.: Human facial illustrations: creation and psychophysical evaluation. ACM Trans. Graph. (TOG) **23**(1), 27–44 (2004)

7. Kae, A., Sohn, K., Lee, H., Learned-Miller, E.: Augmenting CRFS with boltzmann machine shape priors for image labeling. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2019–2026 (2013)
8. Levin, A., Lischinski, D., Weiss, Y.: A closed-form solution to natural image matting. IEEE Trans. Pattern Anal. Mach. Intell. **30**(2), 228–242 (2008)
9. Meng, M., Zhao, M., Zhu, S.C.: Artistic paper-cut of human portraits. In: Proceedings of the 18th ACM International Conference on Multimedia, pp. 931–934. ACM (2010)
10. Mentus, T., Marković, S.: Effects of symmetry and familiarity on the attractiveness of human faces. Psihologija **49**(3), 301–311 (2016)
11. Rosin, P.L., Lai, Y.K.: Non-photorealistic rendering of portraits. In: Proceedings of the Workshop on Computational Aesthetics, pp. 159–170. Eurographics Association (2015)
12. Selim, A., Elgharib, M., Doyle, L.: Painting style transfer for head portraits using convolutional neural networks. ACM Trans. Graph. (TOG) **35**(4), 129 (2016)
13. Tang, X., Wang, X.: Face sketch synthesis and recognition. In: Proceedings of the Ninth IEEE International Conference on Computer Vision, pp. 687–694. IEEE (2003)
14. Wang, N., Gao, X., Sun, L., Li, J.: Anchored neighborhood index for face sketch synthesis. IEEE Trans. Circ. Syst. Video Technol. (2017)
15. Wang, N., Gao, X., Sun, L., Li, J.: Bayesian face sketch synthesis. IEEE Trans. Image Process. **26**(3), 1264–1274 (2017)
16. Wang, N., Tao, D., Gao, X., Li, X., Li, J.: Transductive face sketch-photo synthesis. IEEE Trans. Neural Netw. Learn. Syst. **24**(9), 1364–1376 (2013)
17. Xu, Z., Chen, H., Zhu, S.C., Luo, J.: A hierarchical compositional model for face representation and sketching. IEEE Trans. Pattern Anal. Mach. Intell. **30**(6), 955–969 (2008)
18. Zhang, L., Lin, L., Wu, X., Ding, S., Zhang, L.: End-to-end photo-sketch generation via fully convolutional representation learning. In: Proceedings of the 5th ACM on International Conference on Multimedia Retrieval, pp. 627–634. ACM (2015)
19. Zhang, S., Gao, X., Wang, N., Li, J., Zhang, M.: Face sketch synthesis via sparse representation-based greedy search. IEEE Trans. Image Process. **24**(8), 2466–2477 (2015)
20. Zhang, Y., Dong, W., Ma, C., Mei, X., Li, K., Huang, F., Hu, B.G., Deussen, O.: Data-driven synthesis of cartoon faces using different styles. IEEE Trans. Image Process. **26**(1), 464–478 (2017)
21. Zhou, H., Kuang, Z., Wong, K.Y.K.: Markov weight fields for face sketch synthesis. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1091–1097. IEEE (2012)