Human Hair Segmentation In The Wild Using Deep Shape Prior

Yongzhe Yan, Anthony Berthelier, Stefan Duffner, Xavier Naturel, Christophe Garcia, Thierry Chateau

▶ **To cite this version:**

Yongzhe Yan, Anthony Berthelier, Stefan Duffner, Xavier Naturel, Christophe Garcia, et al.. Human Hair Segmentation In The Wild Using Deep Shape Prior. Conference on Computer Vision and Pattern Recognition Workshop (CVPR Workshop), Jun 2019, Long Beach, United States. hal-02891974

HAL Id: hal-02891974
https://hal.science/hal-02891974

Submitted on 7 Jul 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Human Hair Segmentation In The Wild Using Deep Shape Prior

Yongzhe Yan[1,2] Anthony Berthelier[1,2] Stefan Duffner[3] Xavier Naturel[2]
Christophe Garcia[3] Thierry Chateau[1]
[1]Université Clermont Auvergne, CNRS, SIGMA, Institut Pascal, F-63000 Clermont-Ferrand, France
[2]Wisimage [3]Université de Lyon, CNRS, INSA-Lyon, LIRIS, UMR5205, F-69621, France
yongzhe.yan@etu.uca.fr

Abstract

Virtual human hair dying is becoming a popular Augmented Reality (AR) application in recent years. Human hair contains diverse color and texture information which can be significantly varied from case to case depending on different hair styles and environmental lighting conditions. However, the publicly available hair segmentation datasets are relatively small. As a result, hair segmentation can be easily interfered by the cluttered background in practical use. In this paper, we propose to integrate a shape prior into Fully Convolutional Neural Network (FCNN) to mitigate this issue. First, we utilize a FCNN with an Atrous Spatial Pyramid Pooling (ASPP) module [2] to find a human hair shape prior based on a specific distance transform. In the second stage, we combine the hair shape prior and the original image to form the input of a symmetric encoder-decoder FCNN to obtain the final hair segmentation output. Both quantitative and qualitative results show that our method achieves state-of-the-art performance on the publicly available LFW-Part and Figaro1k datasets.

1. Introduction

In recent years, hair detection and segmentation plays an important role in AR applications such as virtual hair dying and facial animation [6]. Hair segmentation is also an essential prior step for 3D hair modeling from a single portrait image [1]. Human hair contains rich color, shape and textural information. It is generally related to gender, culture and personality. On the other hand, human hair appearance can be notably influenced by the environmental lighting conditions as well.

Hair segmentation *in the wild* consists in performing hair segmentation in an unconstrained view without any explicit prior face or head-shoulder detection [8]. We address this problem as a semantic segmentation problem by taking texture and shape constraints into account. Hair segmentation, especially under *in the wild* conditions, is challenging for the two following reasons in practical AR applications:

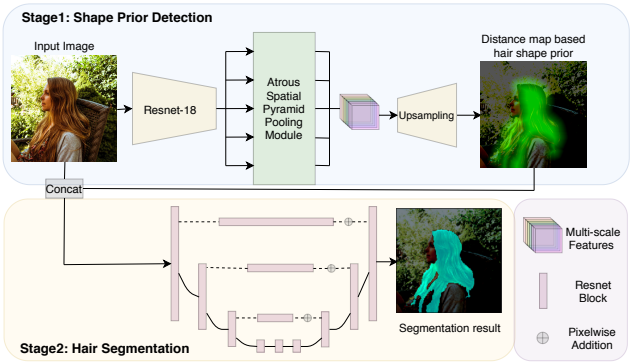
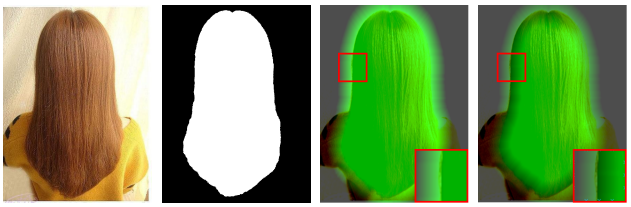


Figure 1: Shape prior integrated hair segmentation pipeline.

- **Cluttered background**: textures in the background can be similar to human hair, which introduce significant difficulties for hair segmentation *in the wild*.
- **Lack of rigid and consistent form**: the form of hair varies according to the head pose, different point of view and ambient environment such as wind. However, we believe that human hair, although in different situations, share implicit shape constraints.

In this paper, we aim at improving hair segmentation *in the wild* by correctly distinguishing hair texture from similar texture in the background. Previous CNN-based methods generally adopt a single stage network [6, 7], which is insufficient under such conditions. We propose a two-stage pipeline (see Fig. 1.) consisting of a shape prior detection stage and a hair segmentation stage. Our contributions can be summarized as: before segmentation, we propose to first estimate a hair shape prior which is based on a specific distance transform map. The results show that it helps to improve the robustness on the cluttered background.

Project page: https://yozey.github.io/Hair-Segmentation-in-the-wild



(a) (b) (c) (d)

Figure 2: An illustration of our distance map transformation. From left to right: (a) Original image (b) Ground truth hair mask (c) Clipped distance transform map overlaid on original image (d) Clipped distance transform map with "erosion" overlaid on the original image. With "erosion", an uncertain region is created on both sides of the hair boundary.

2. Proposed approach

We decouple the hair segmentation task into two important steps: (a) find the general hair shape prior and (b) segment the hair region based on the shape prior. In the first stage, inspired by the hair occurrence probability mask used in previous methods [11, 12, 8] and soft segmentation, we aim at finding a coarse hair mask that indicates hair texture presence and hair shape constraints regardless of their exact border. In the second stage, we aim at identifying the exact hair border by a symmetric encoder-decoder FCNN.

2.1. Hair Shape Prior Detection

Distance map regression: In most of the previously proposed FCNN models for semantic segmentation, object shape constraints are not explicitly imposed. We propose to introduce a coarse hair mask without precise boundary information as a shape prior for segmentation. We transform the binary ground truth hair mask to a boundary-less coarse hair mask by using a specific type of distance transform.

An illustration of our distance map transform is shown in Fig. 2. Consider a binary ground truth hair mask $I(x, y)$ in Fig. 2(b). Hair pixels and non-hair pixels can be denoted respectively as $I^+ = \{I(x, y) = 1\}$ and $I^- = \{I(x, y) = 0\}$. We define a clipped distance transform map dt_{mask} on the image positions $p(x, y)$ as:

$$dt_{mask}(p) = d_{max} - \min(d_{max}, \min_{p^+ \in I^+} \|p^+ - p\|) \quad (1)$$

where d_{max} denotes the maximum clipping threshold for distance values (see Fig. 2(c)). And, similarly, we define a clipped inverse distance transform map with respect to the background pixels:

$$dt_{inv}(p) = e_{max} - \min(e_{max}, \min_{p^- \in I^-} \|p^- - p\|) \quad (2)$$

where $e_{max} (< d_{max})$ denotes the second clipping threshold. Then, the final distance transform map dt is obtained by:

$$dt = dt_{mask} - dt_{inv} \quad (3)$$

which is then normalized between -1 and +1 to form the regression target (see Fig. 2(d)). The use of dt_{inv} "erodes" the initial distance transform dt_{mask} and produces an uncertain hair boundary region for the target image. e_{max} can be considered as the magnitude of "erosion". We do not use traditional mathematical morphology to do this because some small hair regions on the binary mask might be ignored while small holes might be filled. We use HardTanh as final activation function, and L1 loss to train our distance map regression. In our implementation, we empirically set the values of d_{max} to 25 and e_{max} to 10.

Atrous Spatial Pyramid Pooling (ASPP) encoder: Although texture is considered as a local information, in the setting of hair segmentation *in the wild*, the scale of the hair region varies considerably. ASPP with different atrous rates effectively captures multi-scale information to learn the presence of hair texture. We use DeeplabV3 [2] structure with Resnet18 [4] pre-trained on ImageNet as backbone encoder in our hair detection network. Finally we upsample the multi-scale feature map to obtain the final distance transform map in the original image size.

2.2. Hair Segmentation

Symmetric encoder-decoder: In hair segmentation stage, we implement a symmetric encoder-decoder structure with skip connections, which renders refined object borders. At each level, we use a ResNet [4] block in both of the encoder and the decoder part. Additionally, as in [9], we add a ResNet block in the skip connections to process the low-level information transferred from the decoder.

3. Experiments

3.1. Datasets

We conducted our experiments on both the LFW-Part dataset [5] and the newly-released Figaro-1k [8] dataset. The LFW-Part dataset is a face parsing dataset with hair annotation which consists of 2927 images. To the best of our knowledge, Figaro-1k is the only hair analysis dataset *in the wild* with precise hair annotation. It consists of 1050 images (210 for validation) and manually annotated ground truth hair masks, which varies in different hair styles, hair colors, length and levels of background complexity.

3.2. Experimental Settings

For quantitative evaluation, we adopted several standard measures e.g. mean Intersection over Union (mIoU), accuracy and F1-score. The images are resized to 256 $\times$ 256 for training but the evaluation is performed in their original size.

Table 1: Hair Segmentation Results on Figaro1k.

Method	Precision(%)	F1 score(%)	mIoU(%)	Acc(%)
U-Net [10]	95.63	94.39	89.69	96.36
DeeplabV3+ [3]	96.86	95.05	91.11	97.07
Muhammad et al. [8]	-	84.90	-	91.50
Only Stage2	95.64	94.53	89.91	96.56
Stage1 + Stage2	**97.25**	**95.09**	**91.15**	**97.20**

Table 2: Hair Segmentation Results on LFW-Part.

Method	Precision-hair(%)	F1-hair(%)
U-Net [10]	89.11	87.66
DeeplabV3+ [3]	91.66	88.36
Liu et al. [7]	-	83.43
Only Stage 2	89.13	88.07
Stage1* + Stage2*	**98.24**	**88.94**

3.3. Results

We compared our method with the encoder-decoder fully convolutional neural network U-Net [10], the state-of-the-art semantic segmentation approach DeeplabV3+ [3] based on ImageNet pre-trained ResNet18 and the previous work on hair analysis *in the wild* [8] on the Figaro-1k dataset. The result is reported in Table 1 and Figure 3. Our approach outperforms all the previous methods for hair segmentation *in the wild*. By adding a detection stage, a gain of more than 1% point on IoU and F1-score can be achieved. The larger improvement on precision shows that our method is effective for removing false positives on the background. On the LFW-Part dataset (see Table 2), by adding a hair detection stage, our method outperforms other methods by nearly 1% point on the hair F1-score.

4. Conclusions

In this paper, we presented a two-stage pipeline for hair segmentation *in the wild* for AR application. We trained a distance map based hair shape prior, and then estimate the final segmentation by a symmetric FCNN. Our approach outperforms previous state-of-the-art methods, being more robust to cluttered background.

Acknowledgement: This work is sponsored by Région Auvergne-Rône-Alpes. We are grateful to the NVIDIA corporation for a Titan Xp GPU donation.

Figure 3: Challenging examples in Figaro1k. First row: Input image. Second row: Segmentation results without detection stage (only Stage2). Third row: Segmentation results by ImageNet pre-trained DeepLabV3+ [3]. Fourth row: Segmentation results by our two-stage model. Many tiny isolated false positives can still be observed on the man's shirt in the first image of DeepLabV3+ results.

References

[1] Menglei Chai, Tianjia Shao, Hongzhi Wu, Yanlin Weng, and Kun Zhou. Autohair: fully automatic hair modeling from a single image. *ACM Transactions on Graphics*, 35(4), 2016.

[2] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2018.

[3] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[5] Andrew Kae, Kihyuk Sohn, Honglak Lee, and Erik Learned-Miller. Augmenting CRFs with Boltzmann machine shape

priors for image labeling. 2013.

[6] Alex Levinshtein, Cheng Chang, Edmund Phung, Irina Kezele, Wenzhangzhi Guo, and Parham Aarabi. Real-time deep hair matting on mobile devices. *arXiv preprint arXiv:1712.07168*, 2017.

[7] Sifei Liu, Jianping Shi, Ji Liang, and Ming-Hsuan Yang. Face parsing via recurrent propagation. In *British Machine Vision Conference (BMVC)*, 2017.

[8] Umar Riaz Muhammad, Michele Svanera, Riccardo Leonardi, and Sergio Benini. Hair detection, segmentation, and hairstyle classification in the wild. *Image and Vision Computing*, 71:25–37, 2018.

[9] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision*, pages 483–499. Springer, 2016.

[10] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[11] Dan Wang, Xiujuan Chai, Hongming Zhang, Hong Chang, Wei Zeng, and Shiguang Shan. A novel coarse-to-fine hair segmentation method. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 233–238. IEEE, 2011.

[12] Nan Wang, Haizhou Ai, and Feng Tang. What are good parts for hair shape modeling? In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 662–669. IEEE, 2012.