PORTRAIT SEGMENTATION BY DEEP REFINEMENT OF IMAGE MATTING

Carlos Orrite, Miguel Angel Varona, Eduardo Estopiñán and José Ramón Beltrán

Aragon Institute of Engineering Research, University of Zaragoza, Spain

ABSTRACT

Portrait segmentation is becoming a hot topic nowadays. In this paper we propose a novel framework to cope with the high precision requirements that portrait segmentation demands on boundary area by deep refinement of the portrait matting. Our approach introduces three novel techniques. First, a trimap is proposed by fusing information coming from two well-known techniques for image segmentation, i.e., Mask R-CNN and DensePose. Second, an alpha matting algorithm runs over the previous trimap generate. From this mate result we generate a couple of masks, one of them boundary-sensitive kernel, called boundary and the other one inside-sensitive kernel called leftover. Third, we refine the portrait by a pre-trained CNN-based model, followed by a transposed convolution. We have evaluated our approach on the PFCN dataset as well as the portrait images collected from COCO dataset. Experimental results demonstrate the better performance of our algorithm over previous methods.

Index Terms—portrait, FCNs, trimap, matting

1. INTRODUCTION

Recently, portrait segmentation is becoming a hot topic due to the fact that is widely used in many applications such as portrait stylization, background replacement, augmented reality, etc. Portrait segmentation is generally regarded as a sub-problem of semantic segmentation: However, it differs from traditional segmentation in two aspects. On the one hand, the foreground object is focused only on people. On the other hand, portrait segmentation requires higher precision on boundary area, which it is challenging.

One approach to reach such precision is by means of image matting. Matting refers to the problem of accurate foreground estimation in images and video and it is one of the key techniques in many image editing and film production applications. However, Natural image matting is originally ill-posed. To make the problem tractable, user specified strokes or trimap are used to sample foreground and background colors. In addition, matting generally exhibits a poor performance when an image has similar foreground and background color or complicated textures.

In this paper, we propose a new approach for portrait segmentation using some deep networks for human detection and trimap generation, alpha matting for extracting the first portrait and a deep refinement based on a Fully Convolutional Network and a regression, as shown in Fig. 1.

2. RELATED WORK

In recent years, convolutional neural network (CNN) based methods have been successfully applied to semantic segmentation. In 2014, Fully Convolutional Networks (FCN) by Long et al. [1], popularized CNN architectures for dense predictions without any fully connected layers. This allowed segmentation maps to be generated for image of any size and was also much faster compared to the patch classification approach.

In contrast to the segmentation-first strategy of FCN-based methods, Mask R-CNN [2] is based on an instance-first strategy. Mask R-CNN, extends other CNNs by adding a branch for predicting segmentation masks on each Region of Interest (RoI), in parallel with the existing branch for classification and bounding box regression. This method is generic for instance segmentation, including people.

DensePose [3] is dedicated to dense human pose estimation using discriminative trained models. Authors adopt the architecture of Mask R-CNN with the Feature Pyramid Network features, and ROI-Align pooling so as to obtain dense part labels and coordinates within each of the selected regions. Results are impressive for whole body segmentation but not so good for portrait segmentation.

While previous methods as Mask R-CNN and DensePose can provide a rough segmentation, they generally fail to generate precise segmentation around subject boundaries. One technique widely used to tackle this problem is matting. Image matting aims at extracting foreground elements from an image (I) by means of color and opacity (alpha) estimation, decomposing it into background B and foreground F. Due to the highly ill-posed nature of the matting problem, most existing approaches require additional constraints in the form of user input, either as trimaps or scribbles. There are quite a few matting methods, categorized according to color sampling and propagation. A survey is given in [4].

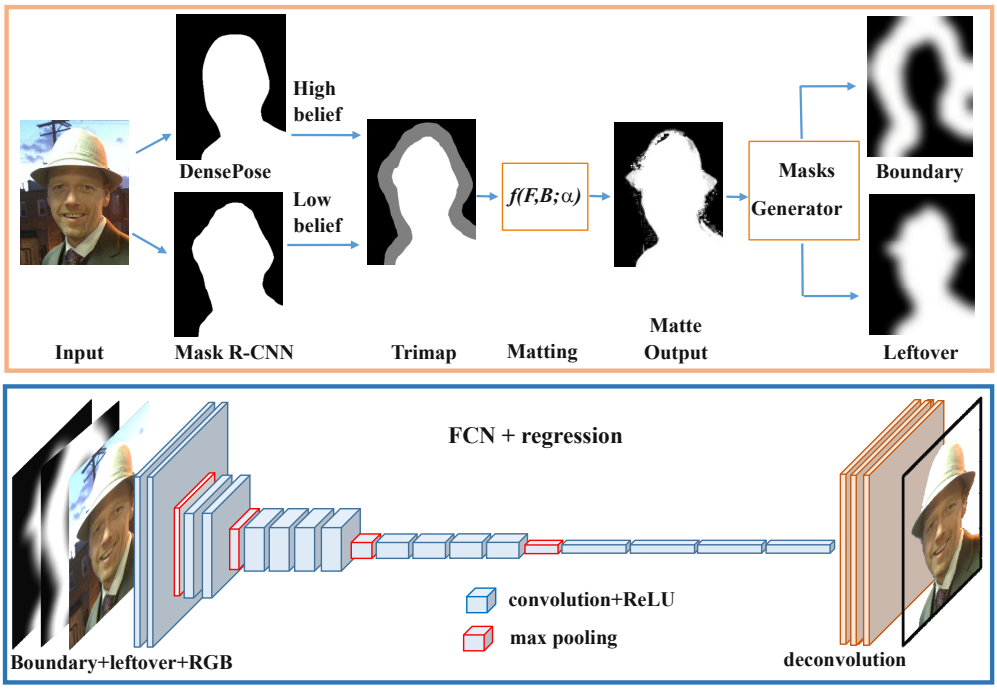


Fig. 1. Deep refinement of matting for portrait segmentation. On the top alpha matting based on DensePose and Mask R-CNN algorithms. On the bottom, the matting refinement by a FCN plus regression.

In [5] authors propose an automatic image matting method for portrait images. The proposal is based on an end-to-end CNN taking the input of a portrait image given the matte result as output.

As portrait segmentation requires a higher precision on boundary area, some authors propose a boundary-sensitive deep neural network (BSN) for portrait segmentation [6]. An individual boundary-sensitive kernel is proposed by dilating the contour line and assigning the boundary pixels with multi-class labels. Afterwards, a global boundary sensitive kernel is employed as a position sensitive prior to further constrain the overall shape of the segmentation map.

3. DEEP REFINEMENT OF MATTING

We define four stages for the overall method. 1) Trimap generation. 2) Alpha matting. 3) Boundary and leftover masks generation. 4) FCN plus regression. See Fig. 1.

3.1. Trimap generation for matting

First of all, we use Mask R-CNN and DensePose for people detection in an image. We use the Mask R-CNN implementation of on Python 3, Keras, and TensorFlow provided by [2]. The model generates bounding boxes and segmentation masks for each instance of an object in the image. DensePose implementation is carried out by [3].

The next step consists of forming a trimap combining the outputs of mask R-CNN and DensePose. Specifically, the masks resulting from both methods are eroded to establish the values of Foreground, and in turn, they are expanded to obtain the Unknown values. Both for erosion and for dilation, the Densepose mask uses a "kernel_size" of half the value used for Mask R-CNN. In this way, we give more confidence to the exit of Densepose, which is more reliable to identify the inside of the human figure. After several tests, the optimal value of 35 pixels is selected.

3.2. Alpha matting

We follow the approach provided by [7]. Basically, the technique takes an input image and its corresponding trimap, previously generated. Finally, the matte output is binarized, scaled and cropped into size 600x800.

3.3. Boundary and leftover masks

We now use the matte output to generate a pair of masks, the one centered on the contour, named boundary and another one on the inside, called leftover. We use a couple of threshold, th1 and th2, for morphological operations.

Algorithm 1. Boundary mask generation
Input: binarized matte image **I**, kernel size **th1**
dilate_layer = dilate (I, th1)
erode_layer = erode (I, th1)
boundary_mask = dilate_layer – erode-layer
boundary = GaussianBlur (boundary_mask, (2*th1)+1)
Output: boundary

Algorithm 2. Leftover mask generation
Input: binarized matte image **I**, kernels size **th1** and **th2**
leftover_mask = erode(I, th2)
leftover = GaussianBlur (leftover_mask, (2*th1)+1)
Output: leftover

3.4. FCN plus regression

The input of the FCN is the Boundary and Leftover masks and the portrait RGB channels. FCNs can be described as the example depicted in Fig. 1: an encoder (a pre-trained model) followed by a decoder (transposed convolutions). For the encoder we used the VGG16 model pretrained on ImageNet for classification [8]. For the decoder transposed convolution is used to upsample the input to the original image size. We train the FCN as a normal classification CNN. In the case of our FCN, the goal is to assign each pixel to the appropriate class (background or foreground). To do so, cross entropy loss is used as the loss function.

(a) (b) (c) (d) (e) (f)
Fig. 2. (a, d) Input; (b, e) original COCO ground truth; (c, f) COCO+ ground truth refinement.

4. EXPERIMENTAL RESULTS

4.1. Datasets

We evaluate our proposal on the PFCN dataset, which is to be considered the largest publicly available portrait segmentation dataset [9]. Authors collected 1800 portrait images from Flickr and manually labeled them with variations in pose, appearance, age, background, illumination, hairstyle, etc. All the training images are scaled and cropped into size 600 x 800. We follow the same protocol as the authors who split the 1800-labeled images into 1500 training images and 300 testing images. In one portrait image, the pixels are labeled as either "foreground" or "background".

To show the effectiveness of our method, we further test on the portrait images collected from COCO [10]. We have selected 100 images in total with ground truth (GT) segmentation maps. This selection is more challenging than the PFCN data in various ways such as large pose variations, large occlusions, large portion of background, different kinds of accessories, etc. As we noticed that the labelling process was not good enough, we run a matting algorithm and manually refined by pixel-level stuff annotations on these 100 testing pictures given as a result COCO+, the refinement GT of COCO. Fig. 2 shows some examples.

4.2. Training and evaluation settings

We have used for boundary and leftover masks the following thresholds:
th1=0.2*sqrt(area of matte output)
th2=0.05*sqrt(area of matte output)
To train the FCN we use 45k iterations with a learning rate of 10e-4, and other 14k iterations with a learning rate of 10e-5.

The segmentation error is measured by the standard metric Interaction-over-Union (IoU) accuracy which is computed as the area of intersection of the output with the GT, divided by the union of their areas as:

$$IoU = area(output \cap GT) / area(output \cup GT)$$

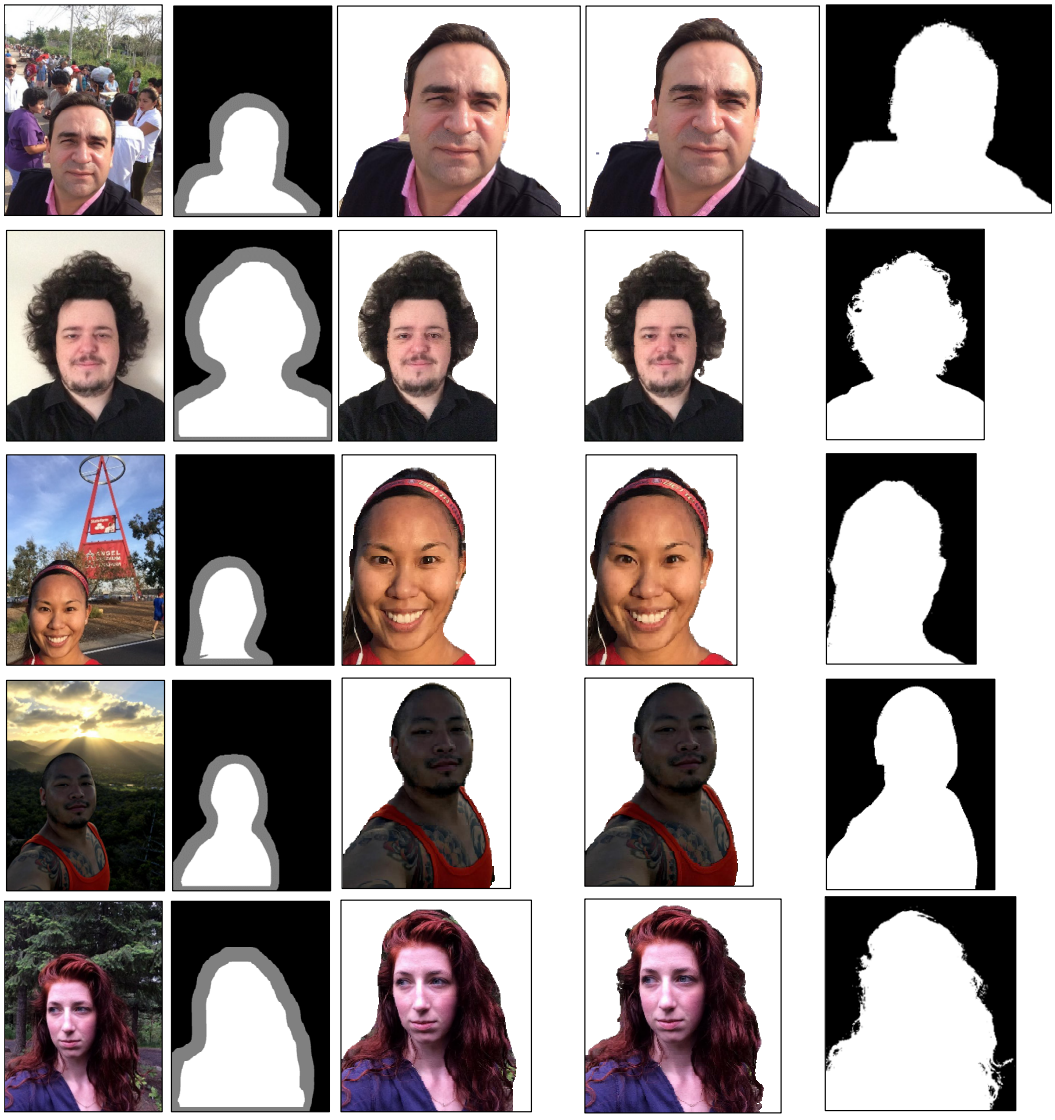
4.3. Results analysis

Table 1: Portrait segmentation performance for different methods and datasets.

Method	PFCN	COCO	COCO+
Shared Matting	94.0%	88.1%	94.0%
Portrait FCN+	95.9%	68.6%	
BSN	96.7%	77.7%	
ours	**97.0%**	**89.6%**	**95.2%**

We compare our method with PortraitFCN+ [9] and BSN [6], which can be considered as the state-of-the-art. Additionally, we provide some results by the combination suggested in this paper of fusing Mask R-CNN and DensePose to generate a trimap used by the alpha matting algorithm [10], which we will refer to shared matting.

Our method outperforms any of the methods under analysis for PFCN dataset, reaching a mean IoU at 97.03%., as well as for COCO dataset, reaching a mean IoU at 89.6%. 89.6%. This score is even higher for COCO+, obtaining a mean IoU at 95.2%. It is worth mentioning we do not provide data for PortraitFCN+ and BSN as the code is not publicly available. Fig. 3 shows how our method is able to segment the portrait in clutter environments, difficult lighting conditions and provide a fairly good approximation to hair segmentation.



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

Fig. 3. (a) Input; (b) trimap; (c) shared matting result; (d) regression after FCN; (e) ground truth.

5. CONCLUSIONS

We have proposed a framework based on two key ideas: 1) using state-of-the-art algorithms, such as Mask R-CNN and DensePose for person detection and fusing both of them to generate a trimap to be used in an alpha matting algorithm 2) to refine the matte output by a pre-trained CNN model followed by transposed convolution to portrait segmentation.

Preliminary results show that our proposal is in the state-of-the-art for portrait segmentation. However, hair is not completely solved, mainly because there are not enough dataset where hair had been perfectly segmented to be used for training.

As future work, we propose an end-to-end fully convolutional network taking as input the masks provided by Mask R-CNN and DensePose to provide the matte result. In addition, we would like to extend our method to obtain the whole silhouette of the subject not just the portrait.

6. REFERENCES

[1] Evan Shelhamer, Jonathan Long, and Trevor Darrell, "Fully convolutional networks for semantic segmentation," IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 4, pp. 640–651, 2017.

[2] Waleed Abdulla, "Mask r-cnn for object detection and instance segmentation on keras and tensorflow," 2017. [3] Iasonas Kokkinos Riza Alp G¨uler, Natalia Neverova, "Densepose: Dense human pose estimation in the wild," 2018.

[3] Riza Alp Güler, Natalia Neverova, Iasonas Kokkinos "DensePose: Dense human pose estimation in the wild," 2018.

[4] JueWang and Michael F. Cohen, "Image and video matting: A survey," Found. Trends. Comput. Graph. Vis., vol. 3, no. 2, pp. 97–175, Jan. 2007.

[5] Xiaoyong Shen, Xin Tao, Hongyun Gao, Chao Zhou, and Jiaya Jia, "Deep automatic portrait matting," in Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part I, 2016, pp. 92–107.

[6] Xianzhi Du and Larry S. Davis, "Boundary-sensitive network for portrait segmentation," CoRR, vol. abs/1712.08675, 2017.

[7] Eduardo S. L. Gastal and Manuel M. Oliveira, "Shared sampling for real-time alpha matting," Computer Graphics Forum, vol. 29, no. 2, pp. 575–584, May 2010.

[8] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," CoRR, vol. abs/1409.1556, 2014.

[9] Xiaoyong Shen, Aaron Hertzmann, Jiaya Jia, Sylvain Paris, Brian L. Price, Eli Shechtman, and Ian Sachs, "Automatic portrait segmentation for image stylization," Comput. Graph. Forum, vol. 35, no. 2, pp. 93–102, 2016.

[10] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick, "Microsoft COCO: common objects in context," CoRR, vol. abs/1405.0312, 2014.