Artificial Intelligence and Human-Computer Interaction
Y. Ye and P. Siarry (Eds.)

doi:10.3233/FAIA240138

Refinement of Trimap for Portrait Matting

Yangjun XU [a,1], Xiaodong HUI [a], Kun ZHANG [b], Xin JIN [a]
[a] *Southern Power Grid Digital Grid Technology Co., Guangzhou, China*
[b] *China Southern Power Grid Company Limited, Guangzhou, China*

Abstract. Portrait matting refers to separating the portrait part from the background in an image. The difficulty of the problem lies in accurately identifying the pixels of the person and also maintaining the contour details. In this paper, we propose a fully automatic deep learning approach to achieve portrait matting. Firstly, semantic segmentation is used to predict the probabilities of pixels belonging to portrait, background, and unknown region, then a trimap is obtained. In order to remove the misclassification of pixels, we refine the portion of head contour for the trimap. The method used is to introduce the result of facial landmark detection, and erosion operation is performed on the head region while maintaining the integrity of the facial contour of the portrait. After that, we use deep matting method to predict the alpha value in the image to get the matting results at the details. We then propose a novel framework that integrates the optimised trimap, the deep matting result, and the original image to obtain the final matting result. Both qualitative and quantitative experiments verify the effectiveness of the proposed method.

Keywords. Matting, Segmentation, Trimap, Deep learning

1. Introduction

Portrait matting aims at extracting human portraits from natural images with various applications, such as human-computer interaction, video image synthesis, and so on. Therefore, it is of great practical importance to design a method that enables fully automatic portrait matting. It is crucial to have a method that guarantees quality and efficiency in extracting the target person.

Portrait matting is full of challenges. Before deep learning techniques became popular, most matting algorithms were based on sampling [1–4] or propagation [5–8]. While such algorithms were able to roughly matting out the foreground, they only took into account low-level features and lacked contextual information that did not take into account high-level as well. When the colour spatial distributions of the foreground and background overlap, this class of methods can confuse similarly coloured foregrounds and backgrounds, making the matting much less effective. Another class of methods is semantic segmentation, which can roughly separate the target foreground, but often blurs out the details of structuring and transparency.

In recent years, with the continuous improvement of computer hardware performance, deep learning methods have been highly valued and applied to various fields to

[1]Corresponding Author: Yangjun Xu, Southern Power Grid Digital Grid Technology Co., Guangzhou, China. E-mail: 245873393@qq.com.

solve problems, also including semantic segmentation [9-10] and image matting [9–15]. However, many existing deep learning-based matting methods still require the input of manually labelled trpmaps, which is not convenient to use. Therefore, we would like to develop a deep learning matting method that can automatically acquire the trimaps. At the same time, we consider combining deep learning methods with traditional methods, applying low-level features as well as high-level context, maintaining the structure and details of the target, and further improving the effectiveness of portrait matting.

In this paper, a deep matting method is proposed, which consists of five steps. The first step is to perform semantic segmentation of the image to coarsely separate the foreground from the background. The second step converts the semantic segmentation result into a trimap suitable for matting, i.e., it is divided into foreground, background and unknown regions. The third step is to get the face region by extracting the faical landmarks. The fourth step combines the original image and the optimised trimap to predict the alpha value. The fifth step is to fuse the matting results from different methods.

Because semantic segmentation is a kind of coarse segmentation, it will treat some backgrounds between the hair strands as portrait, which will lead to an inaccurate trimap, greatly affecting the subsequent matting results. Therefore, this paper proposes that the trimap needs to be optimised. That is, the trimap of the head region is subjected to an erosion operation to ensure the integrity of the portrait face. Among them, the number of corrosion is judged using the face feature points. Experiments will prove that a simple optimisation of the trimap will improve the final portrait matting results very significantly.

2. Proposed method

We uses a deep learning based portrait matting framework. The model PSPNet [16] is used to semantically segment the image, predicting for each pixel in the image the probability that the pixel belongs to the portrait. Then the image is transformed into a trimap by threshold segmentation, i.e., dividing the image into foreground, background, and unknown regions. After producing the preliminary trimap, we carry out a unique optimisation strategy for the trimap. Firstly, the foreground in the trimap is extracted, and then the face part in the foreground is extracted. The erosion is operated on the extracted facial region to remove the pixels around the face that are wrongly classified as foreground. At the same time, the results of facial landmark detection are introduced to control the number of erosion steps, which maintains the integrity of the face contour. Optimising the trpmap with this method results in a higher quality trimap suitable for matting. The optimised trimap as well as the original image are inputted into the second network, and the alpha value of the whole image is obtained using the depth matting method [9], where the transparency value of the image edge details can be obtained. Finally, the fusion strategy proposed in this paper is used to obtain the final matting results. This strategy first replaces the unknown region in the trimap with the result of deep matting to get the first matting result. Then the second matting result is obtained by using the trimap and the original image using the closure matting method [5]. Finally, we replace the head of the portrait with the first matting result to get the final result. The overall framework of the proposed method is shown in Figure 1.

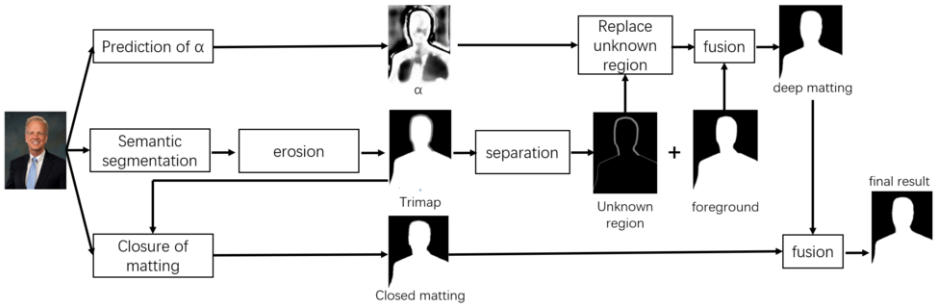

Figure 1. Overall flowchart of the proposed method.

2.1. Formulation

Portrait matting aims to extract humans from natural images. An image can be viewed as a linear combination of foreground and background, which can be represented by Equation (1):

$$I = \alpha F + (1 - \alpha)B, \alpha \in [0, 1], \tag{1}$$

where I denotes the colour image, B denotes the background of the image, F denotes the foreground of the image, and α denotes the value of the alpha channel on the image. The problem to be solved in this paper is how to obtain α and F from I. According to Equation (1), only the input colour image I is known while B, F and α are unknown. Thus for each pixel there are 7 unknown variables and 3 known variables which makes the original matting problem unsolvable. Therefore, most of the existing matching algorithms require the user to provide a trpmap or the user to sketch on the image as a constraint of the problem. The matting algorithms in this paper are more concerned with algorithms that can automatically generate a trimap of the image.

2.2. Generation of Tripmap

Generating a trimap corresponds to semantically segmenting an image and roughly extracting the foreground regions. Specifically, a trimap indicates an image into three regions, i.e., foreground, background and unknown regions. In this paper we have chosen PSPNet [16]. PSPNet provides an effective global context prior for pixel-level scene parsing. The main feature of this network is that the pyramid pooling model it contains is then able to use the information perceived by its sub-neighbourhoods to make judgments. In other words, the pyramid pooling module connects the different levels of the feature maps produced by the pyramid pools into a single fully connected layer for classification. As far as computational cost is concerned, PSPNet does not increase much compared to the original expanded FCN network. In end-to-end learning, global pyramid pool modules and local FCN features can be optimised simultaneously. In this paper, a pre-trained ResNet model is used to extract the feature maps. Thresholding is performed instantly for the output F of PSPNet to obtain the trimap. The thresholding value set in this paper is used as shown in Equation (2):

$$P = \begin{cases} 1, & if F > \theta_1 \\ 0.5, & if \theta_1 \leq F \leq \theta_1 \\ 0, & if F < \theta_2 \end{cases} \quad (2)$$

where P represents the value of the trimap and P=0 represents the background region, P=0.5 represents the unknown region, and P=1 represents the foreground region. In this paper, we take θ_1=0.95 and θ_2=0.5 as thresholds. From this we transform the binary classification segmentation result into a trimap suitable for matting.

2.3. Detection of facial landmarks

Facical landmark detection is carried by an algorithm based on regression tree ensemble [17]. The algorithm predicts the location of face landmarks directly from a subset of sparse pixel intensities and achieves real-time performance through high quality prediction results. In addition to this, the algorithm uses a gradient-based enhancement method for learning the regression tree ensemble. Missing or partially labelled data is handled by optimising the sum of squared error losses.

2.4. Optimising the Trimap by erosion

We design an iterative corrosion method to optimise the trimap obtained in the first part. Firstly, the facial region is determined based on the result of facial landmark detection, and then the region of the head is subjected to an erosion operation, in which the structural elements of the erosion are shown in Figure 2. Before each erosion, it is judged whether the kandmarks of the face are in the foreground region. If they are in the foreground region, the erosion continues, otherwise, the erosion stops. This is done to minimise the domain of mis-segmentation while keeping the face correctly matted. Suppose the foreground part of the original trimap is F_o and the foreground part after erosion is F_α. The part of F_o that does not contain F_α is labelled as an unknown region.

2.5. Fusion

The deep matting network is able to predict the transparency of the whole image, the structure of the image and the image details very well, but the network does not work well with semantic information. Coupled with the fact that the trimap generated earlier has mispredicted pixels in the body part, the misclassified pixels need to be patched with a closure matting method to enhance robustness, and the resultant fusion strategy will be described below.

2.5.1. Fusing deep matting results with semantic information

Based on the trimaps obtained from the prediction, it can be known that the results of the deep matting network need to be used only in the unknown area, because most of this part is the place that possesses more complex structures, such as hair, transparent clothes, etc. In areas that clearly belong to the foreground or background can be completely matted using only the results of the semantic segmentation obtained by PSPNet. So this step is to replace the unknown regions directly with the predicted alpha values obtained from deep matting to get a matting result whose contour details are generated by deep matting.

2.5.2. Closure

After obtaining the matting result that incorporates the deep matting and semantic information, it is found that there are still misclassified pixels in the body part, which is due to the misclassification of the foreground pixels of the body as the unknown pixels. In general, the misclassified pixels are very close in colour to the surrounding pixels. Therefore we have chosen a matting closure scheme [5] to repair the misclassified pixels in the body. The algorithm is designed based on the assumption that the local colour space follows a linear model and is suitable for repairing such mis-segmented pixels.

2.5.3. Fusion of matting results

After obtaining the results produced by the deep matting results and the closure scheme respectively, this paper is going to combine the advantages of these two methods to obtain the final image. The initial fused image is first obtained using Equation (3).

$$I_{tem} = max(I_{closed}, I_{deep}) \tag{3}$$

where I_{tem} represents the intermediate result, I_{closed} represents the matting result generated using the closure scheme, and I_{deep} represents the matting result generated using the depth matting method. In this step, a more complete matting result can already be obtained, but simply taking the maximum of the two predictions will increase the foreground of the details that are wrongly divided, such as part of the background between the hair will be given a larger alpha value, thus reducing the quality of the matting. Considering that the human head has more such details, the face detection result is introduced, and the alpha value of the face region is replaced with the matting result produced by the depth matting method.

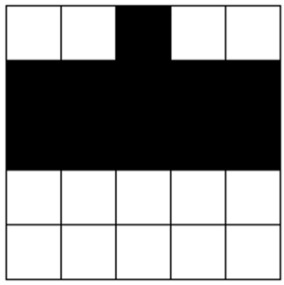

Figure 2. Structural elements of erosion.

3. Experiment

3.1. Experimental Setup

The test set used in this paper is derived from the alphamatting.com dataset, in which images containing portraits were selected. Three metrics will be used to evaluate the quality of keying, namely absolute error, mean square error, and the gradient error. The values of all three metrics are as smaller as better. In this paper, the following two algorithms will be used as benchmark algorithms for comparison: the PSPNet [16] and the closure scheme [5]. Since the closure matting scheme [5] is required to input the trimap, all trimap will be generated using the algorithm in this paper to ensure fairness.

3.2. Performance Comparison

In this section the proposed method of this paper will be compared with other methods and the experimental results are shown in Table 1. Using the semantic segmentation method for direct matting has the worst results in terms of data. Since in the portrait matting task not only requires the ability to separate the foreground and background, but also requires maintaining the complex structure of the foreground as well as some of the details. From the data, the absolute error and the mean square error of semantic segmentation are smaller than those of other matting methods, which shows that semantic segmentation is good enough for classification at the pixel level. From the view of the gradient error, which is an index that indicates the degree of smoothing between the predicted photographs and the true photographs, the gap of semantic segmentation is larger than that of other methods, which shows that it is difficult for this kind of method to learn the structure and details of the image.

The method in this paper performs best in pixel level classification and also in the smoothing part between the detailed pixels and the background. The key reason is that the proposed method results in a better smooth transition of the details in the detailed region.

Figure 3 shows the comparison of the four methods with the groundtruth, from which it can be seen that the method in this paper not only ensures semantic accuracy at the pixel level, but also smoothes over the parts with more complex details at the hair.

Table 1. Comparison of different matting methods.

Method	Absolute error	Mean square error	Gradient error
PSPNet[16]	1.3	1.6	5.8
Close[5]	1.2	1.4	4.0
deep matting[9]	1.2	1.4	3.0
Ours	1.1	1.2	2.9

Figure 3. Visual comparison of different matting methods.From left to right are the PSPNet[16], Close[5], Deep Matting[9], Ours, and the the groundtruth.

4. Conclusion

A deep learning based fully automatic portrait matting method is designed. Compared with the existing deep learning based methods, the proposed method does need a input of

the trimap but focuses on the correction of the automatically extracted trimap, and also the integration of the deep learning methods with the traditional methods. Experiments verify the effectiveness of the proposed method.

References

[1] Chuang YY, Curless B, Salesin DH, Szeliski R. A bayesian approach to digital matting. In: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001. vol. 2. IEEE; 2001. p. II-II.

[2] Gastal ES, Oliveira MM. Shared sampling for real-time alpha matting. In: Computer Graphics Forum. vol. 29. Wiley Online Library; 2010. p. 575-84.

[3] He K, Rhemann C, Rother C, Tang X, Sun J. A global sampling method for alpha matting. In: CVPR 2011. Ieee; 2011. p. 2049-56.

[4] Shahrian E, Rajan D, Price B, Cohen S. Improving image matting using comprehensive sampling sets. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2013. p. 636-43.

[5] Levin A, Lischinski D, Weiss Y. A closed-form solution to natural image matting. IEEE transactions on pattern analysis and machine intelligence. 2007;30(2):228-42.

[6] Chen Q, Li D, Tang CK. KNN matting. IEEE transactions on pattern analysis and machine intelligence. 2013;35(9):2175-88.

[7] Sun J, Jia J, Tang CK, Shum HY. Poisson matting. In: ACM SIGGRAPH 2004 Papers; 2004. p. 315-21.

[8] Grady L, Schiwietz T, Aharon S, Westermann R. Random walks for interactive alpha-matting. In: Proceedings of VIIP. vol. 2005. Citeseer; 2005. p. 423-9.

[9] Xu N, Price B, Cohen S, Huang T. Deep image matting. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2017. p. 2970-9.

[10] Shen X, Tao X, Gao H, Zhou C, Jia J. Deep automatic portrait matting. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14. Springer; 2016. p. 92-107.

[11] Zhu B, Chen Y, Wang J, Liu S, Zhang B, Tang M. Fast deep matting for portrait animation on mobile phone. In: Proceedings of the 25th ACM international conference on Multimedia; 2017. p. 297-305.

[12] Cho D, Tai YW, Kweon I. Natural image matting using deep convolutional neural networks. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14. Springer; 2016. p. 626-43.

[13] Chen Q, Ge T, Xu Y, Zhang Z, Yang X, Gai K. Semantic human matting. In: Proceedings of the 26th ACM international conference on Multimedia; 2018. p. 618-26.

[14] Park K, Woo S, Oh SW, Kweon IS, Lee JY. Mask-guided Matting in the Wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2023. p. 1992-2001.

[15] Huang WL, Lee MS. End-to-end Video Matting with Trimap Propagation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2023. p. 14337-47.

[16] Zhao H, Shi J, Qi X, Wang X, Jia J. Pyramid scene parsing network. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2017. p. 2881-90.

[17] Savran A. Multi-timescale boosting for efficient and improved event camera face pose alignment. Computer Vision and Image Understanding. 2023;236:103817.