From Segmentation to Matting : Foreground and Background Separation Algorithms

Wu Zebin
21821034
Industrial Design Engineering Major, College of Computer Science and Technology
Wuzb1995@zju.edu.cn

Abstract

Nowadays, high quality extraction of foreground from natural images, is crucial for a wide variety of applications. Image matting is a fundamental computer vision problem. Different from segmentation, Matting can get a segmentation with transparency channel, which makes the fusion of foreground and background more perfect. This article will introduce some traditional methods of segmentation and matting, and then elaborate on the main process and network architecture of matting using deep learning. This article will help beginners or designers to clarify their learning ideas, make choices about different methods and make better use of them.

Keyword: Matting; Segmentation; Foreground;

1. Introduction

Image matting plays an important role in computer vision, which has a number of applications, such as virtual reality, augmented reality, interactive image editing, and image stylization. Given an input image I, image matting problem is equivalent to decomposing it into foreground F and background B in assumption that I is blended linearly by F and B:

$$I = \alpha F + (1-\alpha) B \qquad (1)$$

Where α is used to evaluate the foreground opacity (alpha matte). However, the formulation of image matting is still an ill-posed problem. Since in natural image matting, all quantities on the right-hand side of the composting Eq(1) are unknown.

To solve this problem, some special methods are thought out. They can be divided into two species: matting with contours and matting with energy optimization. And with the development of neural network, some methods with deep learning come out naturally.

1.1. Segmentation with contours

A simple thought of segmentation is to find the contour of the foreground, and then the problem will be simplified to deal with the contour instead of Eq(1).

The contour of image foreground can be obtained by edge detection. Existing edge detection algorithms include Ostu Algorithm[1], Canny Operator[2], Laplace Operator[3] and so on. Canny Operator is still a widely used edge detection method. But these algorithms always bring poor robustness when backgrounds are not simple. The specific advantages and weakness would be given in section 2.

1.2. Segmentation with energy optimization

For the poor robustness in edge detection used for segmentation, Matting with energy optimization emerges as the times require. The most famous algorithm is Grab Cut[4]. It comes from Normalized Cut[5] and Graph Cut[6], and is widely used for interactive matting. It still keeps its good robustness these days, but when applying in matting, the boundary can't be so natural.

Another famous algorithm is Level Set[7]. It can acquire complete contours and avoid pre/post processing in traditional image segmentation methods, but it is sensitive and easy to fall into local extremum.

1.3. Matting with deep-learning

Because the equation is too difficult to solve, some methods turn to divide this problem into two parts. Firstly, it is necessary to classify pixels in images into three types: foreground, background and uncertain areas. The processed picture containing only the above three types of pixels is called trimap. Subsequently, uncertain areas are further processed with $\alpha \in [0,1]$

With the rise of deep learning, segmentation algorithm has entered the era of neural networks. So trimaps can be generated with FCN[8] and other deep learning segmentation algorithms, and the uncertain areas will be processed by other networks like VGG16[9]. Due to the emergence of a large number of matting data sets, it come up with several matting algorithms with deep learning. Some would be introduced in section 2.

2. Performance of traditional methods

In Section 2, the traditional methods mentioned in Section 1 will be introduced in more detail, and they will be used for comparison to illustrate the advantages and disadvantages of these methods.

2.1. Ostu

Ostu algorithm[1], also known as the maximum inter-class variance method, achieves the automatic selection of global threshold T by counting the histogram characteristics of the whole image.

Ostu algorithm divides original image into foreground and background by threshold. Foreground is represented by the number of points, moments of mass and average gray level of the foreground under the current threshold. Background is represented by the points, moments of mass and average gray levels of the background under the current threshold. When choosing the best threshold, the background should be the most different from the foreground. The key is how to choose the criterion to measure the difference. In Otsu algorithm, the criterion to measure the difference is the maximum inter-class variance.

Using OpenCV library, Ostu algorithm can be easily implemented. However, inter-class variance method is very sensitive to noise and target size. It only produces good segmentation results for images with single peak of inter-class variance. So when it applies in the natural situation, its disadvantages in robustness become visible.

The result of Ostu algorithm is as follows. It can only classify the color distribution, but it has a general effect on the foreground with large contrast.

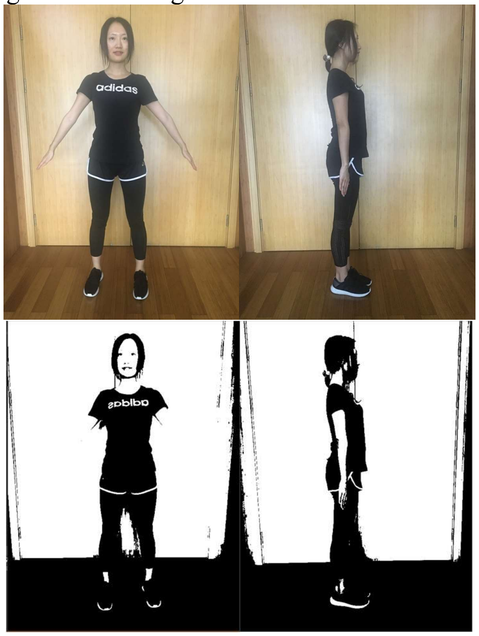

Figure 1. The result of Ostu

2.2. Canny Operator

Canny[2] edge detection operator is a multi-level edge detection algorithm developed by John F. Canny in 1986. Canny operator has three characters to detect edges: noise reduction, calculating the magnitude and direction of gradient and double thresholds.

Noise reduction: The first step is to convolute the original data with the Gauss smoothing template, and the resulting image is slightly blurred compared with the original image. Thus, single pixel noises have little effect on the image.

Gradient calculation: The edges of an image can point in different directions, so the classical Canny algorithm uses four gradient operators to calculate gradients in horizontal, vertical and diagonal directions respectively. Common edge difference operators (such as Sobel[10]) are used to calculate the difference Gx and Gy in horizontal and vertical directions. The gradient modes and directions can be calculated as follows:

$$\mathrm{G} = \sqrt{G_x^2 + G_y^2} \qquad (2)$$

$$\theta = \tan^{-1}({G_y}/{G_x}) \qquad (3)$$

Double thresholds: Canny algorithm uses two thresholds, a high threshold and a low threshold to distinguish edge pixels. If the gradient of edge pixels is greater than the high threshold value, it is considered as a strong edge point.

Today, Canny algorithm and its variants are still an excellent edge detection algorithm. But for the middle-level vision problems such as segmentation, it is still weak.

After processing the noise line by Hough transform, the effect of Canny operator is as follows, but it still has some noises:

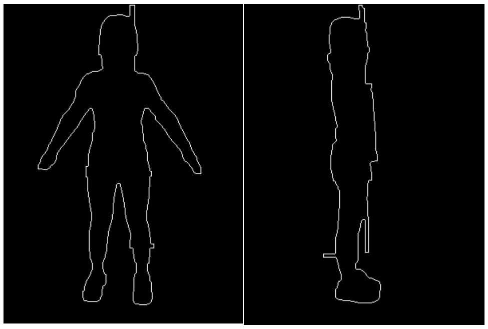

Figure 2. The result of Canny Operator

2.3. Level Set

Level Set[7] method was first proposed by Osher and Stelian in Fronts propagating with curvature dependent speed: algorithms based on Hamilton-Jacobi formulations in 1988 to solve the change process of flame shape following thermodynamic equation.

The basic idea of level Set is that the interface is regarded as the zero level set of a function (called level set function) in the one-dimensional higher space, and the evolution of the interface is extended to the one-dimensional higher space. The level set function is evolved or iterated according to the development equation it satisfies. As the level set function evolves continuously, the corresponding zero level set is also changing. When the level set evolves steadily, the evolution stops and the interface shape is obtained.

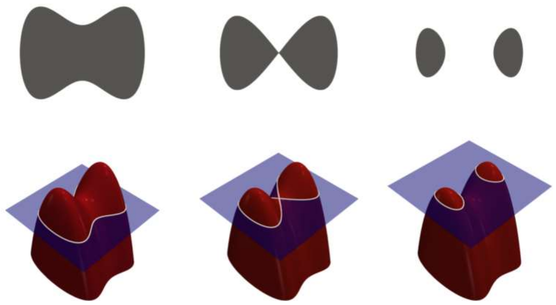

Figure 3. The first line represents a two-dimensional image, and the second line is the segmentation interface of the level set higher than the image.

In computer vision, the advantage of using level set method is that it can be used to calculate the evolution of curves and surfaces in a fixed coordinate system without knowing the parameters of curves and surfaces, so this method is also called geometric method. On the other hand, the disadvantage is also obvious. It is sensitive and easy to fall into local extremum.

It behaves differently under different iterations, and its effect is not well controlled.

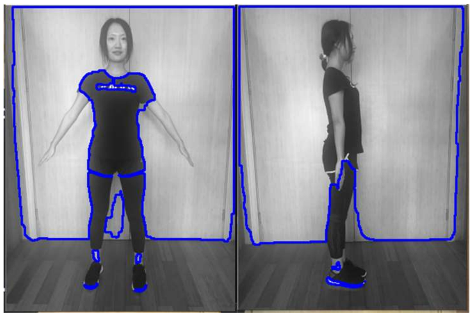

Figure 4. The result of Level Set

2.4. Grab Cut

Grab cut[4] is the best robust method in traditional image segmentation. It is a probabilistic undirected graphical model, also known as Markov random field Markov random field.

Grabcut's basic idea is to create a graph and look at the following graph. The image pixel or super-pixel is the image vertex. Then the optimization goal is to find a cut, so that the sub-images are not connected, so as to achieve segmentation. The premise is to remove the edge and weight minimum.

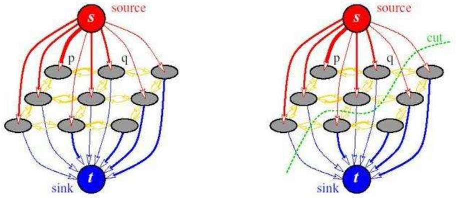

Figure 5. The Schematic diagram of Grabcut.

Grabcut method is very general, and it is also good for image segmentation with complex texture. The disadvantage is that the time complexity and space complexity are high, so super-pixel is usually used to accelerate the calculation.

The results of Grabcut under two iterations are as follows. Multiple iterations are more effective and time-consuming. Rectangular boxes are manually annotated information.

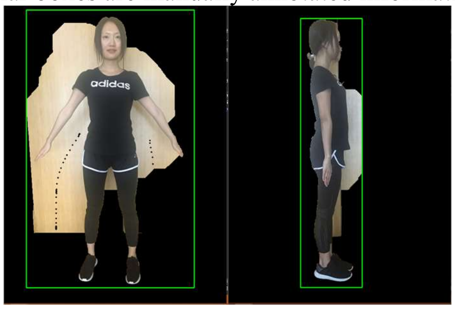

Figure 6. The result of Grab Cut

3. Matting with deep learning

For the separation of foreground and background, traditional algorithms mainly deal with segmentation, among which Grabcut and other algorithms have achieved good results.

However, when the foreground and background of the image are separated, in many cases, the image needs to be further fused to achieve the poster effect. How to integrate more naturally? With the addition of neural network, the calculation of α in Eq(2) becomes more accurate and makes matting more reliable.

In this section, Deep image matting(DIM) and Semantic human matting(SHM) will be introduced.

3.1. DIM

As a new algorithm based on deep learning, DIM[11] mainly solves the problems of low-level features and lack of high-level context in traditional methods. The depth model is divided into two stages. The first stage is the deep convolution coding-decoding network, which takes the original image and the corresponding trimaps as input and predicts the alpha matte of the image. The second stage is a small convolutional neural network, which refines the alpha matte of the first network prediction, thus having more accurate alpha values and sharpening edges. In addition, a large-scale matting data set was created, which contains 49300 training images and 1000 test images. Deep learning with large-scale datasets make the effect better.

3.1.1 Motivation

The motivation of the method comes from two problems existing in the traditional method.

First, the current method designs matting equation as a linear combination of two colors, that is, matting is regarded as a dyeing problem. This method regards color as a distinguishable feature. However, when the color spatial distribution of the current scene and background overlaps, the effect of this method is not very good. Using in-depth

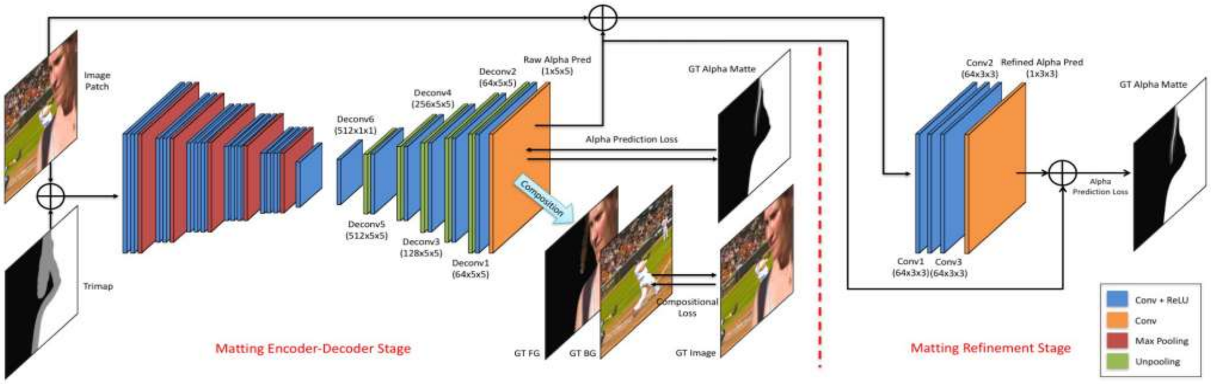

Figure 7: DIM network consists of two stages, an encoder-decoder stage (Sec. 4.1) and a refinement stage (Sec. 4.2)

learning does not rely primarily on color information, it will learn the natural structure of the image and reflect it to alpha matte.

Second, the current data set based on matting is too small. There are only 27 training pictures and 8 test pictures in alphamatting.com data set. The generalization ability of the model trained is poor. To solve this problem, the author cuts out the foreground and puts it into different backgrounds to construct a large-scale matting dataset.

3.1.2 Networks

The network structure consists of two stages, Matting encoder-decoder stage and Matting refinement stage. The structure shows as Figure 7.

Matting encoder-decoder stage

Network: Input is image block and corresponding trimap, and output is alpha prediction. In the coding stage, 14 convolution layers and 5 pooling layers are used to obtain low-resolution feature maps. In the decoding stage, a small network of 6 convolution layers is used. Five times of unpooling, alpha prediction of the original image size is obtained.

Loss: Two losses are used. The first one is alpha-prediction loss, which is the absolute difference between alpha values of prediction and alpha values of ground truth. The second loss is composite loss, which predicts the absolute difference between the RGB color value and the corresponding ground truth. Two loss are weighted by 0.5 to get the final loss.

Realization: Data processing techniques include random clipping (320 * 320), resize to 320 * 320, image flipping, etc. Before loading VGG16 model, Xavier is used to initialize the decoding phase randomly.

Matting refinement stage

Network: Four convolution layers. The input is image block and alpha prediction of prediction.

Realization: First train the coding and decoding network, and then use it to update refine network after convergence. The second network only uses alpha-prediction loss.

3.1.3 Conclusion

In order to generalize to natural images, matting algorithms must go beyond color as the main clue and take advantage of more structural and semantic features. The neural network in this paper has enough ability to capture high-order features and use them to calculate and improve the matting effect. The effects are as follows:

3.2. SHM

The input of DIM contains trimaps, so it can not get alpha mask directly from the input image. It needs to be combined with the existing segmentation algorithm or the groundtruth trimaps. When used, there will be holes. And the foreground set of DIM is not big enough.

To solve these problem, SHM comes out. It specializes in human body matting, and for the first time, it adds semantic information to get better pictures (compared with DIM) by means of chunk semantic segmentation and detail optimization.

SHM[12] uses RGB 3 channel pictures as input and directly outputs 1 channel alpha matte pictures of the same size. There is no need for additional information, such as trimap and scribbles.

3.2.1 T-net

Obtaining trimaps is a pixel-level classification problem. Each pixel belongs to one of three categories: foreground, background and unknown regions.

T-net in SHM is mainly used to solve this problem. The output of T-Net is a 3-channel feature map, which shows the probability that each pixel belongs to three categories. T-Net uses the method of semantic segmentation, and PSPNet-50 is used here. See left half of Figure 8.

3.2.2 M-net

M-Net uses the output of T-Net as the semantic input, and describes the details of human targets by generating rough alpha matte images. M-Net uses the original 3-channel RGB images and the 3-channel segmentation results of T-Net output, and combines them into the input of 6-channel. M-Net is an encoder-decoder network.

The parameters of encoder network in M-Net are the same as those of convolution layer in VGG16 classification network. See middle half of Figure 8

3.2.3 Fusion Module and Loss

The fusion module uses the probability map of unknown region obtained in T-net and the edge refinement distribution of M-net, and combines the probability distribution of foreground in T-net to get the alpha mask. The formula is as follows:

$$\alpha_p = F_s + U_s \alpha_r \tag{4}$$

See right half of Figure 8.

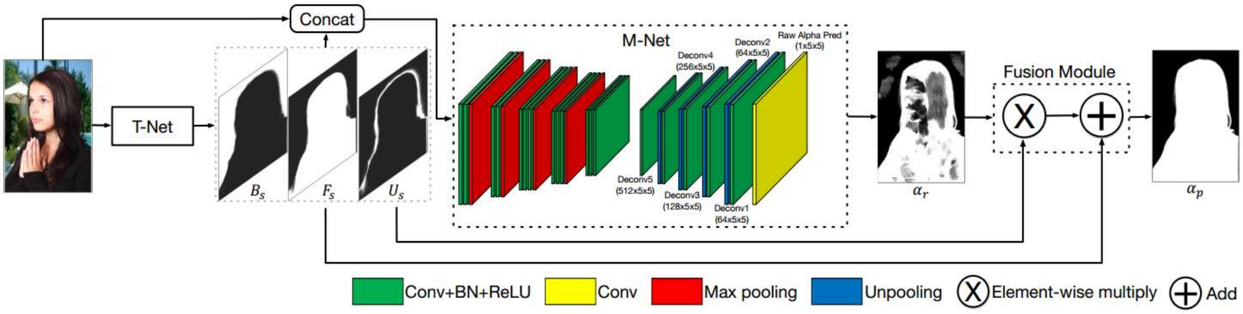

Figure 8: Overview of semantic human matting method. Given an input image, a *T-Net*, which is implemented as PSPNet-50, is used to predict the 3-channel trimap. The predicted trimap is then concatenated with the original image and fed into the *M-Net* to predict the raw alpha matte. Finally, both the predicted trimap and raw alpha matte are fed into the *Fusion Module* to generate the final alpha matte according to Eq. 4. The entire network is trained in an end-to-end fashion.

It uses cross-entropy loss and L1 loss, calculates the loss function of trimap and alpha at the same time, and weights the total loss.

3.2.4 Conclusion

This article aims at the single direction of matting, specializes in image segmentation, and adds semantic information for the first time to get better pictures by means of block semantic segmentation and detail optimization.

The results are as follows, obviously the effect is better:

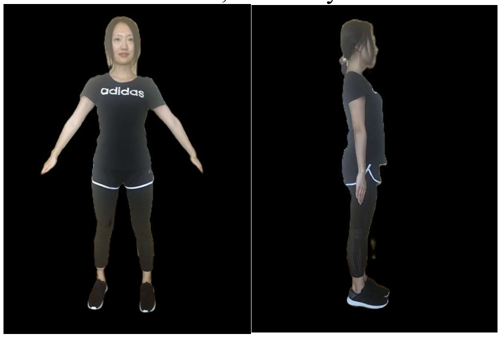

Figure 9. The result of SHM

4. Summary

Due to the limited space, the bottom-up clustering-based segmentation methods, such as k-means[13], meanshift[14] and so on, as well as SVM[15], watershed[16] and other algorithms are not covered in k-means this paper. From segmentation to matting, from geometric methods to neural networks, foreground extraction methods are becoming more and more refined. More and more methods, such as *Fast Deep Matting for Portrait Animation on Mobile Phone*[17], are moving towards faster and more efficient directions.

References

[1] Otsu N. A threshold selection method from gray-level histograms[J]. IEEE transactions on systems, man, and cybernetics, 1979, 9(1): 62-66..

[2] Canny J. A computational approach to edge detection[M]//Readings in computer vision. Morgan Kaufmann, 1987: 184-203.

[3] Evans, L. (1998), Partial Differential Equations, American Mathematical Society, ISBN 978-0-8218-0772-9

[4] Rother C, Kolmogorov V, Blake A. Grabcut: Interactive foreground extraction using iterated graph cuts[C]//ACM transactions on graphics (TOG). ACM, 2004, 23(3): 309-314.

[5] Shi J, Malik J. Normalized cuts and image segmentation[J]. Departmental Papers (CIS), 2000: 107.

[6] Boykov Y, Veksler O, Zabih R. Fast approximate energy minimization via graph cuts[C]//Proceedings of the Seventh IEEE International Conference on Computer Vision. IEEE, 1999, 1: 377-384.

[7] Vese L A, Chan T F. A multiphase level set framework for image segmentation using the Mumford and Shah model[J]. International journal of computer vision, 2002, 50(3): 271-293.

[8] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 3431-3440.

[9] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.

[10] Sobel I, Feldman G. A 3x3 isotropic gradient operator for image processing[J]. a talk at the Stanford Artificial Project in, 1968: 271-272.

[11] Xu N, Price B, Cohen S, et al. Deep image matting[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 2970-2979.

[12] Chen Q, Ge T, Xu Y, et al. Semantic Human Matting[C]//2018 ACM Multimedia Conference on Multimedia Conference. ACM, 2018: 618-626.

[13] Hartigan J A, Wong M A. Algorithm AS 136: A k-means clustering algorithm[J]. Journal of the Royal Statistical Society. Series C (Applied Statistics), 1979, 28(1): 100-108.

[14] Comaniciu D, Meer P. Mean shift: A robust approach toward feature space analysis[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2002 (5): 603-619.

[15] Ricci E, Perfetti R. Retinal blood vessel segmentation using line operators and support vector classification[J]. IEEE transactions on medical imaging, 2007, 26(10): 1357-1365.

[16] Strahler A N. Quantitative analysis of watershed geomorphology[J]. Eos, Transactions American Geophysical Union, 1957, 38(6): 913-920.

[17] Zhu B, Chen Y, Wang J, et al. Fast deep matting for portrait animation on mobile phone[C]//Proceedings of the 25th ACM international conference on Multimedia. ACM, 2017: 297-305.