Automatic Motion-blurred Hand Matting for Human Soft Segmentation in Videos

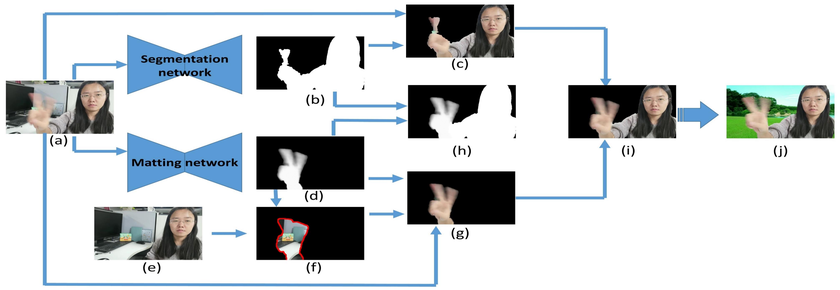
Xiaomei Zhao and Yihong Wu (yhwu@nlpr.ia.ac.cn)

1. NLPR, Institute of Automation, Chinese Academy of Sciences, Beijing, China.
2. University of Chinese Academy of Sciences, Beijing, China

Introduction

Human segmentation technology is very useful in many video or image editing applications. It helps users to change background and brings users a lot of fun. Many excellent image segmentation networks, such as FCN, SegNet, U-Net, and Deeplab can be used for human segmentation. However, these segmentation methods can't deal with motion-blurred hands, which are very common in practical applications. In this paper, we propose a method to deal with human frames with motion-blurred hands. Two short GIF demos of the proposed method are shown in the following.

The flow chart of our method



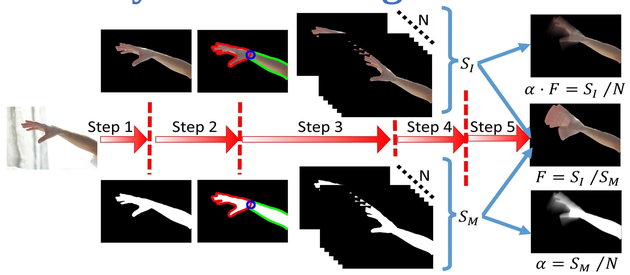
Firstly we segment human by Deeplab v3+ [1],

Then we employ matting network to extract the motion-blurred hand.

Finally, we combine these two kinds of results to generate our final human soft segmentations.

[1] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *ECCV*, 2018, pp. 833–851.

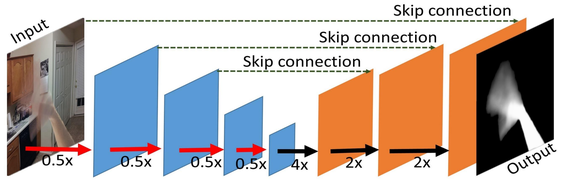
Synthetic data generation



Because it is very difficult to obtain the groundtruth alpha mattes of real motion-blurred hand images, we employ synthetic dataset for training.

We generate synthetic motion-blurred hand images by rotating wrist, elbows, and shoulders. In the above flow chart, we take rotating wrist as an example to introduce each step of our method. N denotes the number of new generated hand images; S_I and S_M denote the sum of new generated hand images and hand masks; α and F denote synthetic alpha matte and hand foreground; motion-blurred hand is calculated by $\alpha \cdot F$.

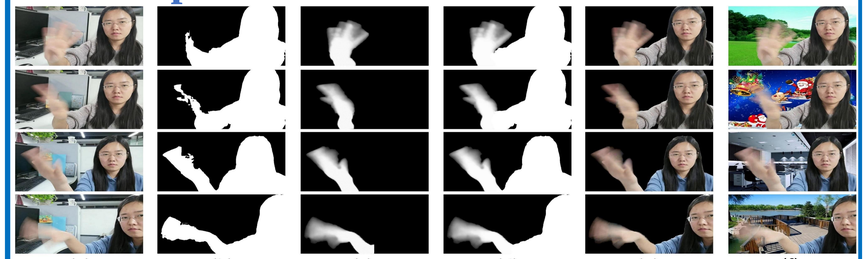
Motion-blurred hand matting network



We employ Xception[2]-based encoder-decoder network for motion-blurred hand matting. The encoder contains 4 downsampling steps. The decoder contains 3 upsampling steps. In each upsampling steps, skip connection is used to recover spatial information.

[2] F. Chollet. "Xception: deep learning with depthwise separable convolutions," in CVPR, 2017, pp. 1800-1807.

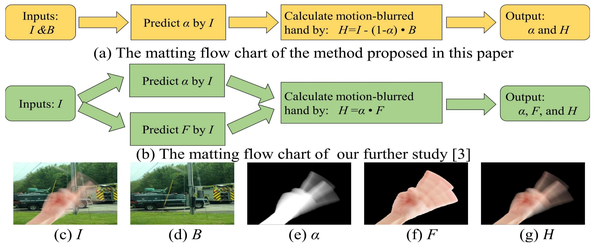
Experimental results on real videos



We can see that human segmentation results in (b) are bad around motion-blurred hands. This problem is successfully solved by our method.

(a) Original frames. (b) Human segmentation results of Deeplab v3+ .
(c) Alpha mattes of motion-blurred hands predicted by matting network.
(d) Human soft segmentation results of the proposed method.
(e) The extracted humans with semi-transparent motion-blurred hands.
(f) New frames with new backgrounds.

Further study



The method proposed in this paper needs background images, which are inconvenient to obtain. In our further study, we propose a method which can extract motion-blurred hand from a single image [3]. In this figure, I denotes original image, B denotes background image, α denotes alpha matte, F denotes foreground image, H denotes the extracted motion-blurred hand.

[3] X. Zhao and Y. Wu, "Automatically Extract Semi-transparent Motion-blurred Hand from a Single Image," in IEEE Signal Processing Letters. doi: 10.1109/LSP.2019.2939754