Paper Reading Report-02

Shreya Chawla
u7195872

Abstract

*This is my reading report for the paper titled: "***Background Matting: The World is Your Green Screen***", authored by Soumyadip Sengupta (University of Washington) et al, and published in IEEE CVPR 2020.*

All ENGN8501 submissions will be subject to ANU's TurnitIn plagiarism check, against both the original paper, internet resources, as well as all other students' submissions. So please make sure that you must write your own reports, and declare the following originality statement:

I, Shreya Chawla, hereby confirm that I am the sole author of this report and that I have compiled it in my own words.

1. Problem Statement

This paper [1] attempts to solve the underconstrained problem of background matting to extract foreground object in images and videos and composite onto a new background. It is an important task in image and video editing applications like in film production for visual effects.

The composition equation is given as: $I = \alpha F + (1 - \alpha)B$, where the image I is separated into foreground image F, a background image B and an alpha matte α. It is the inverse of image composition. Contrary to image segmentation (into foreground and background), in matting the pixels can belong to both foreground and background, represented by alpha values. [3]

2. Summary of the paper's main contributions

This paper is the first trimap-free automatic matting algorithm that utilizes a casually captured background. The paper claims that their architecture for image and video matting is novel. Other contributions include a self-supervised adversarial training to improve mattes on real images and conducted several experiments.

They claim the imaging process and the model is both easy to implement and less time consuming than the traditionally used manually created trimap (foreground/background/unknown segmentation) and green/blue screen methods. The method is stated to work even when the background is similar to foreground or has slight movements. They also claim that unlike the prior methods, their method does not fail for complex body poses and fine features like hair and fingers.

3. Method and Experiment

Their methodology to estimate foreground and alpha, makes use of an additional static natural background image without the subject B'. For image they automatically compute a soft segmentation of the person in frame other than the original image and optionally if video input is available, nearby frames are utilized to aid in matting.

They introduce a novel architecture - "Context Switching Block" (CS Block) to combine different the input cues separately. The architecture, G_{Adobe} is a fully supervised network trained on 300 labelled images from the Adobe Matting dataset followed by self-supervised adversarial training on real unlabelled videos. Further, a discriminator network D guides the training to generate realistic results for student-techer learning. It handles complex difficulties like misalignment, traces of background etc.

Comparisons to a variety of well-performing recent deep matting algorithms are made. G_{Adobe} is trained on synthetic-composite Adobe dataset and a copy of network trained on real world images called G_{Real} are both tested on real-world data. Both models are experimented on a wide range of inputs - handheld and fixed camera, indoor and outdoor conditions.

4. Critical Analysis

4.1. Significance of the paper's contributions

It is the first trimap-free automatic approach. It introduces several novel ideas like CS block. Domain gap between synthetically composed and real imagery is bridged by D, and several experiments are performed.

Prior matting techniques usually require human annotated trimaps or a green screen, both of which are inconvenient. Previously, background subtraction and segmentation did not work well for this problem as they do not solve for the partial transparencies or the foreground colours.

4.2. Validity of the authors' main claims

The limitations of the alternatives are overcome. The quantitative results like alpha matte error and user study with aggregated scores show that their method significantly outperforms the previous techniques. Qualitatively, their method is cutting edge. The range of experiments demonstrate the relative robustness of their approach to deal with a wide variety of inputs further supporting the authors' claims. They demonstrate significant improvement over the current state of the art.

4.3. Limitation and weaknesses

The methods fails when the background contains extreme movements. In that case, the image or video gets distorted especially around edges. [1]

This has been addressed in [2] paper by the same authors. They have used two deep networks - base and refinement, wherein the later one refines the former's results. Not only does the refinement network recover high resolution matting details, it also reduces redundant computations by only operating on patches selected by error prediction map rather than the whole image.

Although this method showed high-quality matting results, the architecture is limited to 512×512 resolution and runs only at 8 fps. This limitation is overcome by [2], which uses a real-time unified matting architecture that operates on 4K videos at 30fps and HD videos at 60fps, and produces higher quality results than BGM. Another limitation is the requirement of two images like for the trimap based solutions. Also, a static background with small camera motion is an ideal requirement for it work well. [1]

4.4. Extension and future work

This paper can be extended to incorporate varying resolutions and videos of different fps. The problem of movement in background in video could be further explored by using spatio-temporal information. [4] This work is human-focused and could be extended to other beings and objects as well. Another experiment is to test performance when image/video contains multiple subjects.

Some of the applications of this work are in image recognition, video and image processing like in movie CGI and Zoom virtual background. This work can be extended to these applications. [2]

4.5. Is the paper stimulating or inspiring ?

This paper achieved the state of the art results while it's technique is relatively easier to implement and faster. Several novel approaches were used or introduced. G_{Adobe} outperformes G_{real} was an unexpected result. The role of motion cues in clearer matting and the CS Block's role to effectively utilize colour difference cues, segmentation, and motion cues was an interesting study.

4.6. Conclusion and personal reflection

A state-of-the-art approach to solve image matting is introduced that captures high quality alpha and foreground subject in everyday setting. If tasked to solve this problem, I would extend the work to use spatio-temporal information from videos.

References

[1] Soumyadip Sengupta, Vivek Jayaram, Brian Curless, Steve Seitz, and Ira Kemelmacher-Shlizerman. *Deep Video Deblurring for Hand-held Cameras*. In Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2020. 1, 2

[2] Shanchuan Lin, Andrey Ryabtsev, Soumyadip Sengupta, Brian Curless, Steve Seitz, and Ira Kemelmacher-Shlizerman. *Real-Time High-Resolution Background Matting* In Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2021. 2

[3] Anat Levin, Dani Lischinski, and Yair Weiss. *A Closed Form Solution to Natural Image Matting*. In Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2006. 1

[4] Yanan Sun, Guanzhi Wang, Qiao Gu, Chi-Keung Tang, and Yu-Wing Tai. *Deep Video Matting via Spatio-Temporal Alignment and Aggregation* In CoRR. 2021. 2