Perceptually motivated benchmark for video matting

Mikhail Erofeev[1]
merofeev@graphics.cs.msu.ru
Yury Gitman[1]
ygitman@graphics.cs.msu.ru
Dmitriy Vatolin[1]
dmitriy@graphics.cs.msu.ru
Alexey Fedorov[1]
afedorov@graphics.cs.msu.ru
Jue Wang[2]
juewang@adobe.com

[1] Lomonosov Moscow State University
Moscow, Russia
[2] Adobe Systems
Seattle, WA, United States

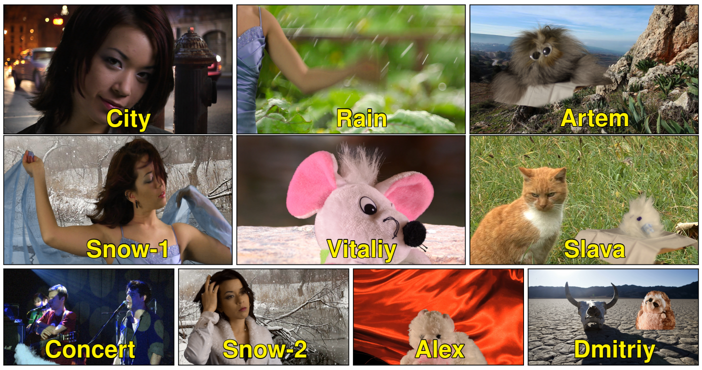

Figure 1: Preview frames of our test sequences.

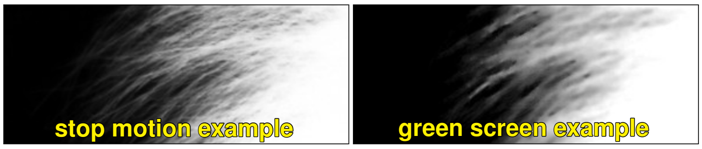

Figure 2: Alpha mattes from chroma keying and stop-motion capture for the same image region. The stop-motion result is significantly better at preserving details.

Despite recent progress in the field of video matting, neither public data sets nor even a generally accepted method of measuring quality has yet emerged. In this paper we present an online benchmark for video-matting methods. Using chroma keying and a reflection-aware stop-motion capturing procedure, we prepared 12 test sequences. Then, using subjective data, we performed extensive comparative analysis of different quality metrics. The goal of our benchmark is to enable better understanding of current progress in the field of video matting and to aid in developing new methods.

Formally, matting is an inverse alpha-compositing problem: i.e., given pixel I, we want to find transparency value $\alpha \in [0;1]$, foreground pixel F, and background pixel B so that

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

The problem is ill posed yet solvable by considering the affinity of pixels in natural images. Matting of natural images is well studied, and according to [3], natural-image matting algorithms are continuously improving.

Video matting is a relatively new research direction that arose recently as available processing power increased. Applied to video, matting has two special requirements: tolerance of sparse user input and temporal coherence of the resulting transparency values.

Despite the rising interest, research in the field of video matting is still weakly organized. In fact, many developers estimate the quality of their methods by visual comparison [1, 2, 4].

The two main challenges facing an effective comparison are preparation of the data set and choice of a quality metric. In this paper we address both challenges and describe a benchmark, available at http://videomatting.com, that provides a comparison, two training sequences with ground-truth transparency, and multiple visualizations for convenient analysis of the comparison results.

To prepare the data set (see Figure 1), we imposed four requirements on our test sequences: high quality for the ground-truth transparency, natural appearance, complexity, and diversity. To satisfy the first two requirements, we used two different techniques of foreground-object capture: namely, capture in front of a green screen and sequential photography (stop motion) against different backgrounds. Chroma keying enabled us to obtain alpha mattes of natural-looking objects with arbitrary motion. Nevertheless, this technique cannot guarantee that the alpha maps are natural, because it assumes the screen color is absent from the foreground object (see Figure 2). To get alpha maps that have a very natural appearance, we used the stop-motion method.

We designed the following procedure to perform stop-motion capture: an object with a fuzzy edge sits on the platform in front of an LCD monitor. The object rotates in small, discrete steps along a predefined 3D trajectory, controlled by two servomotors connected to a computer. After each step, the digital camera in front of the setup captures the motionless object against a set of background images. At the end of the process, we remove the object, and the camera again captures all of the background images.

Following [3], we can solve for transparency values in a system of alpha-compositing equations (see Equation 1). Instead, however, we add reflectance term to Equation 1 to allow for lighting variations caused by background-image changes thus ensuring correct transparency value for object's parts that reflect light emitted by the monitor.

Finally, we composed the extracted objects with ground-truth transparency map over a set of challenging backgrounds and prepared several trimap-width gradations.

Having obtained our test sequences, we conducted extensive subjective comparison of 12 matting methods. Using the collected data, we then compared different quality metrics. The results showed that alpha temporal coherence is significantly more important to human perception of video-matting quality than accuracy and that despite the imperfection of optical-flow techniques, use of these techniques can improve estimation of matte temporal consistency.

To simplify access to the benchmark and enable addition of new methods, we created the http://videomatting.com website, which contains scatterplots and rating tables for different quality metrics. At the time of publication benchmark includes comparison of 12 methods by accuracy and temporal coherence. Besides the comparison, the website includes instructions for new participants.

[1] Xue Bai, Jue Wang, and David Simons. Towards temporally-coherent video matting. In *International Conference on Computer Vision (ICCV)*, pages 63–74, 2011. doi: 10.1007/978-3-642-24136-9_6.

[2] Inchang Choi, Minhaeng Lee, and Yu-Wing Tai. Video matting using multi-frame nonlocal matting laplacian. In *European Conference on Computer Vision (ECCV)*, pages 540–553, 2012. doi: 10.1007/978-3-642-33783-3_39.

[3] Christoph Rhemann, Carsten Rother, Jue Wang, Margrit Gelautz, Pushmeet Kohli, and Pamela Rott. A perceptually motivated online benchmark for image matting. In *Computer Vision Pattern Recognition (CVPR)*, pages 1826–1833, 2009. URL www.alphamatting.com.

[4] Mikhail Sindeev, Anton Konushin, and Carsten Rother. Alpha-flow for video matting. In *Asian Conference on Computer Vision (ACCV)*, pages 438–452, 2013. doi: 10.1007/978-3-642-37431-9_34.