Multiview-Consistent Color Sampling for Alpha Matting

Markus Kettern[1] and Peter Eisert[1,2]

[1]Fraunhofer Heinrich Hertz Institute, Berlin, Germany†
[2]Humboldt Universität zu Berlin, Germany

Abstract

We propose a method for incorporating multiview information into color sampling methodologies for alpha matting. We extend the typical trimap input for matting algorithms to 3D space, enabling the association of stripes of pixels in semi-transparent regions like hair between different views without knowing or inferring their spatial structure. We also formulate a new cost function penalizing deviations of foreground color estimates between associated pixel stripes and its possible integration into a state-of-the-art color sampling approach.

1. Introduction

Alpha matting describes the process of generating a transparency map α and a foreground color map F for an object in image I, enabling realistic composition of the object over new background images. Matting is not an automatic segmentation process and requires input indicating some foreground and some background regions in I. This input is typically specified by a *trimap* consisting of three disjoint regions M_f, M_b, M_u representing the foreground, background and unknown regions of I, respectively.

Most approaches to compute α and F involve estimating the foreground and background color for each pixel $\mathbf{p} \in M_u$ depending on the color sets $I(M_f)$ and $I(M_b)$, either by first creating some model for these colors or by sampling from very small subsets $C_f(\mathbf{p}) \subset M_f$, $C_b(\mathbf{p}) \subset\in M_b$ (e.g. 20 points each in [WC07]) which typically depend on the position of $\mathbf{p}$. In the 2D search space spanned by these subsets, the combination $\mathbf{f} \in C_f(\mathbf{p}), \mathbf{b} \in C_b(\mathbf{p})$ minimizing

$$\mathcal{E}_c(\mathbf{p},\mathbf{f},\mathbf{b}) = \|I(\mathbf{p}) - (\hat{\alpha}(\mathbf{p})I(\mathbf{f}) + (1-\hat{\alpha}(\mathbf{p}))I(\mathbf{b}))\| \quad (1)$$

with

$$\hat{\alpha}(\mathbf{p}) = \frac{(I(\mathbf{p}) - I(\mathbf{b}))\,(I(\mathbf{f}) - I(\mathbf{b}))}{\|I(\mathbf{f}) - I(\mathbf{b})\|^2} \quad (2)$$

is inferred, often by brute-force search.

Only recently, [HRR*11] proposed a stochastic approach for this sampling process to be efficiently carried out in a bigger search space, using the complete boundary of M_f, M_u as C_f and the boundary of M_b, M_u as C_b, independent of the position of $\mathbf{p}$. They use the cost function

$$\mathcal{E}(\mathbf{p},\mathbf{f},\mathbf{b}) = \omega\mathcal{E}_c(\mathbf{p},\mathbf{f},\mathbf{b}) + \mathcal{E}_s(\mathbf{p},\mathbf{f}) + \mathcal{E}_s(\mathbf{p},\mathbf{b}) \quad (3)$$

$$\mathcal{E}_s(\mathbf{p},\mathbf{f}) = \frac{\|\mathbf{f}-\mathbf{p}\|}{\min_{\mathbf{g}\in M_f}\|\mathbf{g}-\mathbf{p}\|} \quad (4)$$

with $\mathcal{E}_c(\mathbf{p},\mathbf{f},\mathbf{b})$ as defined in (1), $\mathcal{E}_s(\mathbf{p},\mathbf{b})$ defined analogue to $\mathcal{E}_s(\mathbf{p},\mathbf{f})$ and ω being a weighting constant compensating for the different scales of measurement. This function is minimized for each pixel in an iterative process alternating between a propagation step and a random search step. In the propagation step the current results $\mathbf{f_p}, \mathbf{b_p}$ for every pixel $\mathbf{p} \in M_u$ are replaced by $\mathbf{f_q}, \mathbf{b_q}$ for pixels $\mathbf{q}$ adjacent to $\mathbf{p}$ if $\mathcal{E}(\mathbf{p},\mathbf{f_q},\mathbf{b_q}) < \mathcal{E}(\mathbf{p},\mathbf{f_p},\mathbf{b_p})$. In the random search step, a new pair $\hat{\mathbf{f}}, \hat{\mathbf{b}}$ is chosen randomly and replaces $\mathbf{f_p}, \mathbf{b_p}$ if producing a smaller cost for $\mathbf{p}$. Here, we propose a method for applying this approach to multiview image sets in order to obtain mattes and foreground maps which are more consistent between the individual views. Often, mattes obtained from color sampling are smoothed by using them as a prior to the closed-form global matting method proposed in [LLW08].

2. Constructing a 3D trimap

In a multiview setup, trimaps can either be obtained from manual annotation, automatically by volumetric closure of an initial binary segmentation, or by a combination of interactive and automatic methods like trimap propagation [SHG09]. From the trimaps and the calibration information,

† This work was supported by the European FP7 project REACT (288369)

DOI: 10.2312/conf/EG2012/posters/031-032

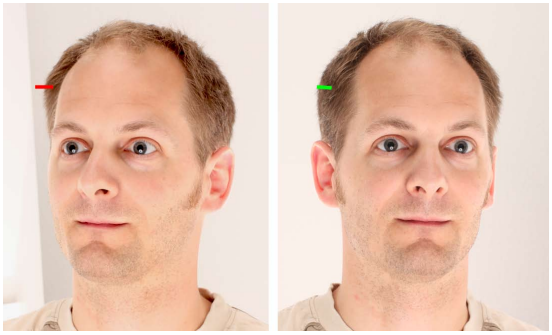

Figure 1: *Corresponding point sets. The red line indicates a stretch of an epipolar line passing through region $M_u^{(i)}$ in view i, the green line represents the corresponding stretch in a different view j which does not lie entirely in $M_u^{(j)}$.*

we can create two visual hulls [Lau94], one for the foreground layers $M_f^{(i)}$ for all views i, representing the minimum volume of the object and one for the union of foreground and unknown layers $M_f^{(i)} \cup M_u^{(i)}$ representing its maximum extent in space. This pair of hulls forms a 3D trimap which will enable us to promote color consistency between the foreground color estimations for the individual views.

3. Color sampling over multiple views

Since transparency in digital images typically occurs at object boundaries (with hair often having multiple boundaries in one pixel), it is unreasonable to assume that corresponding pixels will have the same α value in different views since the boundary regions will be viewed from a different angle. The foreground color, however, should be equal or at least very similar in different views for most objects if strong directional lighting is avoided.

If we select a point $\mathbf{p}^{(i)} \in M_u^{(i)}$ in view i, the 3D trimap allows us to infer a segment of the corresponding epipolar line in view $j \neq i$ representing all possible locations of $\mathbf{p}^{(j)}$, the counterpart of the foreground portion of $\mathbf{p}^{(i)}$. See figure 1 for an example of such corresponding line segments. We denote the set of points in view j associated with pixel $\mathbf{p}^{(i)}$ as $S^{(j)}(\mathbf{p}^{(i)})$ with $S^{(j)}(\mathbf{p}^{(i)}) = \emptyset$ if these points are occluded in view j. The colors of the pixels in this set are expressed by a suitable color model $\Phi\left(S^{(j)}(\mathbf{p}^{(i)})\right)$.

We define the *epipolar consistency error* of a foreground hypothesis $\mathbf{f}^{(i)}$ for pixel $\mathbf{p}^{(i)}$ by the mean of the distances between $I(\mathbf{f}^{(i)})$ and $\Phi\left(S^{(j)}(\mathbf{p}^{(i)})\right)$ for all $S^{(j)}(\mathbf{p}^{(i)}) \neq \emptyset$:

$$\mathcal{E}_e(\mathbf{p},\mathbf{f}) = \frac{1}{|J|}\sum_{j\in J} \delta\left(I^{(i)}(\mathbf{f}^{(i)}), \Phi\left(S^{(j)}(\mathbf{p}^{(i)})\right)\right) \quad (5)$$

$$J = \left\{ j \,|\, S^{(j)}(\mathbf{p}^{(i)}) \neq \emptyset \right\} \quad (6)$$

where δ measures the distance of a single color to the chosen color model.

We propose to use the color line model [OW04] for this task since it is a non-parametric multi-modal color representation that can be extracted without human intervention. The 3D RGB histogram is divided into spherical slices of similar intensity in each of which points representing colors with maximum occurrence are found. A 2D-gaussian is fitted around each of these maxima and the gaussians of neighboring slices are connected if they are close to each other.

The distance function δ of an individual color to the color line model may be defined as the euclidean distance to the closest color line in RGB space. Note that computing δ is computationally inexpensive since a color set $S^{(j)}(\mathbf{p}^{(i)})$ may be expected to be represented by a very small number of color lines.

Finally we propose a modified version of the color sampling algorithm presented in [HRR*11] introducing the epipolar consistency error into cost function (3) by

$$\mathcal{E}^*(\mathbf{p},\mathbf{f},\mathbf{b}) = \omega\mathcal{E}_c(\mathbf{p},\mathbf{f},\mathbf{b}) + \omega\mathcal{E}_e(\mathbf{p},\mathbf{f}) + \mathcal{E}_s(\mathbf{p},\mathbf{f}) + \mathcal{E}_s(\mathbf{p},\mathbf{b}) \quad (7)$$

which is simple as long as $S^{(j)}(\mathbf{p}^{(i)}) \cap M_u^{(j)} = \emptyset$. If there are overlaps between these regions, $S^{(j)}(\mathbf{p}^{(i)})$ may contain colors that are blended with the background colors of view j and will severely corrupt the results of color sampling. To overcome this problem we propose to carry out each iteration step for all images before proceeding to the next step and to use the current foreground layer estimate of each view for inferring $\mathcal{E}_e(\mathbf{p},\mathbf{f})$. This introduces the need to recalculate the color model representation Φ every time a change to the foreground estimates for $S^{(j)}(\mathbf{p}^{(i)})$ is made, but since the iterative color sampling process usually converges within less than 100 iterations this effort seams feasible, especially if the process of updating the color models is accelerated, e.g. by employing an adaptive implementation of the color line model which allows to replace individual colors without recalculating the whole model.

References

[HRR*11] HE K., RHEMANN C., ROTHER C., TIANG X., SUN J.: A global sampling method for alpha matting. In *Proc. CVPR* (2011). 1, 2

[Lau94] LAURENTINI A.: The visual hull concept for silhouette-based image understanding. *PAMI 16* (1994), 150–162. 2

[LLW08] LEVIN A., LISCHINSKI D., WEISS Y.: A closed-form solution to natural image matting. *PAMI 30* (2008). 1

[OW04] OMER I., WERMAN M.: Color lines: image specific color representation. In *Proc. CVPR* (2004), vol. 2. 2

[SHG09] SARIM M., HILTON A., GUILLEMAUT J.-Y.: Wide-baseline matte propagation for indoor scenes. In *Proc. CVMP* (2009). 1

[WC07] WANG J., COHEN M.: Optimized color sampling for robust matting. In *Proc. CVPR* (2007). 1