Non-Parametric patch based video matting

Muhammad Sarim
m.farooqui@surrey.ac.uk
Adrian Hilton
a.hilton@surrey.ac.uk
Jean-Yves Guillemaut
j.guillemaut@surrey.ac.uk

Centre of Vision Speech and Signal processing,
Faculty of Engineering,
University of Surrey,
Guildford, GU2 7XH, Surrey,
United Kingdom

In computer vision, matting is the process of extracting foreground objects while preserving their pixel-wise coverage in the scene. This coverage is referred to as opacity or alpha matte. Once an accurate alpha matte is estimated, a foreground object can be seamlessly composited onto a new background. In this paper we present a novel patch based non-parametric approach for video matting. The technique provides a strong mechanism to represent local image features, colours and textures which attempts to preserve the spatial information of a natural video sequence. This overcomes the limitation of parametric algorithms [1, 2, 3, 5, 7, 8] which only rely on strong colour correlation or affinities between the nearby pixels. The matting problem was first formulated in [4] as linear interpolation of distinct foreground and background images to form a composite image as

$$C_p = \alpha_p F_p + \left(1 - \alpha_p\right) B_p. \tag{1}$$

This equation is known as the compositing equation where, C_p, F_p and B_p are the composite, foreground and background colours for the pixel p respectively while α_p is their blending proportion. Equation (1) is clearly under-constrained therefore in a studio environment it is constrained by using a uniform background [6] while in natural images and videos having arbitrary background, user interaction in the form of a trimap is required to constrain (1).

Initially we construct a clean background by utilising optical flow or inpainting approach. In a studio sequences, a background plate is generally available. Unlike other techniques, our approach constructs a trimap from the previous frame and the background plate. If it contains error, the user can interact online to define a new key frame. Initially the user defines a fine trimap T^k for the frame I^k. Background subtraction and template-wise normalised sum of square difference are used to propagate the trimap value from T^k to T^{k+1}. A square patch, ψ_p, of dimensions n is centred at an unmarked pixel p in the frame I^{k+1}. A patch set ϕ is constructed by localizing patches at all the local foreground and unknown pixels corresponding to p in the frame I^k. The most similar patch, ϕ_q, is found by

$$\phi_q = \underset{\phi_i \in \phi}{\arg\min} \; \frac{1}{n^2} \, d\left(\psi_p, \phi_i\right) \tag{2}$$

where, $d\left(\psi_p, \phi_i\right)$ is the sum of square difference between ψ_p and ϕ_i and n^2 is the number of pixels in the patch for normalization. The trimap value T_p^{k+1} of the pixel p is assigned as

$$T_p^{k+1} = \begin{cases} T_q^k & if, \;\; d\left(I_q^k, I_p^{k+1}\right) \le \varepsilon \\ unknown & otherwise \end{cases} \tag{3}$$

where, T_q^k is the trimap value of pixel q in the trimap T^k and ε is the predefined distance threshold in RGB space. A final refining step is applied to fill in the unknown holes.

A similar approach is utilized to estimate the foreground colour for every unknown pixel. A foreground patch set θ is constructed in a similar fashion to the patch set ϕ by localising a patch only at the local known foreground pixels. The normalised sum of square difference is calculated between the patch ψ_p and the set θ. The foreground colour for the pixel p is estimated as the median of the centre pixel, θ_i^c, of the N most similar patches in the set θ as

$$\tilde{f} = \mu_{1/2}\left(\theta_1^c, \theta_2^c, .., \theta_N^c\right) \tag{4}$$

The α value for pixel p in the unknown region is computed as

$$\alpha_p = \frac{c_p - \tilde{b_p}}{\tilde{f_p} - \tilde{b_p}}, \tag{5}$$

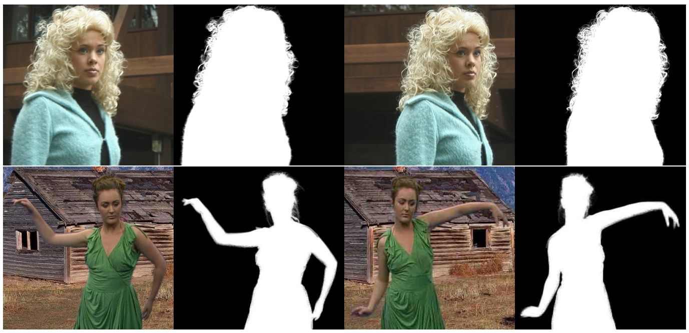

Figure 1: Temporally distant frames and their alpha mattes.

where, c_p, $\tilde{f_p}$ and $\tilde{b_p}$ are the composite, approximated foreground and background colour respectively. The process is iterated for all the unknown pixels to get the final alpha matte.

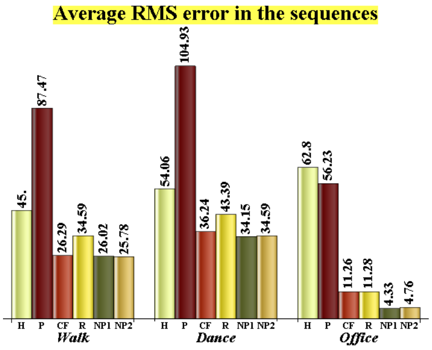

Figure 2: The initials H, P, CF, and R refer to techniques [2, 3, 7, 8] respectively while NP1 and NP2 is our technique with and without predefined key frames.

The average RMS error for different techniques applied on different video sequences is shown in the Fig 2. Quantitative evaluation shows that our technique outperforms the current state-of-the-art approaches.

[1] Y. Y. Chuang, B. Curless, D. H. Salesin, and R. Szeliski. A bayesian approach to digital matting. In *Proceedings of IEEE CVPR '01*, volume 2, pages 264–271, December 2001.

[2] P. Hillman, J. Hannah, and D. Renshaw. Alpha channel estimation in high resolution images and image sequences. In *IEEE CVPR*, pages 1063–1068, 2001.

[3] A. Levin, D. Lischinski, and Y. Weiss. A closed form solution to natural image matting. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 1:61–68, 2006. ISSN 1063-6919. `http://doi.ieeecomputersociety.org/10.1109/CVPR.2006.18`.

[4] T. Porter and T. Duff. Compositing digital images. In *ACM SIGGRAPH '84: Proceedings of the 11th annual conference on Computer graphics and interactive techniques*, pages 253–259, 1984.

[5] M. A. Ruzon and C. Tomasi. Alpha estimation in natural images. In *CVPR*, pages 18–25, June 2000.

[6] A. R. Smith and J. F. Blinn. Blue screen matting. In *ACM SIGGRAPH '96: Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 259–268, 1996.

[7] J. Sun, J. Jia, C.K. Tang, and H. Y. Shum. Poisson matting. *ACM Transactions on Graphics*, 23(3):315–321, 2004.

[8] J. Wang and M. F. Cohen. Optimized color sampling for robust matting. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 0:1–8, 2007.