TEMPORAL TRIMAP PROPAGATION FOR VIDEO MATTING USING INFERENTIAL STATISTICS

Muhammad Sarim, Adrian Hilton and Jean-Yves Guillemaut

Centre of Vision, Speech and Signal Processing
University of Surrey
Guildford, Surrey, United Kingdom
Emails: m.sarim@surrey.ac.uk, a.hilton@surrey.ac.uk and j.guillemaut@surrey.ac.uk

ABSTRACT

This paper introduces a statistical inference framework to temporally propagate trimap labels from sparsely defined key frames to estimate trimaps for the entire video sequence. Trimap is a fundamental requirement for digital image and video matting approaches. Statistical inference is coupled with Bayesian statistics to allow robust trimap labelling in the presence of shadows, illumination variation and overlap between the foreground and background appearance. Results demonstrate that trimaps are sufficiently accurate to allow high quality video matting using existing natural image matting algorithms. Quantitative evaluation against ground-truth demonstrates that the approach achieves accurate matte estimation with less amount of user interaction compared to the state-of-the-art techniques.

Index Terms— Video matting, trimap, statistical inference

1. INTRODUCTION

Matting is a classic problem of image and video processing. It is a process of extracting foreground objects while preserving their pixel-wise opacity (alpha matte) in the scene. Once an accurate alpha matte is estimated, a foreground object can be seamlessly composited onto a new background. The matting problem is formulated as a linear interpolation [1] of distinct foreground and background images using an alpha channel as

$$C_p = \alpha_p F_p + (1 - \alpha_p) B_p, \quad (1)$$

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. The alpha value ranges from 0 to 1, where $\alpha = 0$ defines the background while $\alpha = 1$ defines the foreground. Blended pixels have intermediate alpha values. Equation (1) is clearly under-constrained as all the variables on the right hand side are unknown. In a studio environment, (1) can be constrained by using a uniform background, typically blue or green [2] providing a trivial solution to (1). For natural scenes having arbitrary background, equation (1) is constrained by manually identifying definite foreground and background regions. This interaction is referred to as a trimap where definite foreground and background pixels are represented by white and black respectively while remaining unknown pixels by gray as shown in Fig 1. Given a trimap, matting algorithms use local or global image statistics of known regions to compute the alpha values for the unknown region. Existing video matting techniques require regular manually defined key frame trimap (typically every 10-20 frames) to automatically generate the trimaps for the remaining frames by interpolation. A matting algorithm is then applied to individual frames independently.

This paper presents a framework for video matting given only $1 - 2$ manually defined key-frame trimaps for an entire video sequence. A statistical inference framework is introduced for reliable temporal propagation of trimap labels from the key-frames based on label confidence. This allows robust estimation of trimaps for all images in the sequence having illumination variation and fast non rigid object motion like hair and loose clothes. Conventional image matte estimation is then applied to estimate high-quality video mattes. The approach allows high quality video matting with considerably less manual interaction required compare to the state-of-the-art video matting techniques.

2. RELATED WORK

Techniques like [3–5] fit statistical models, typically Gaussian mixture, to the local known foreground and background pixels which are then used to estimate the foreground and background colour for the unknown pixel and finally compute its alpha value. All of these techniques assumed that the known foreground and background regions are locally smooth and have strong correlation in colour space. These algorithms tend to suffer when the local distributions overlap or the unknown region is wide. To overcome this problem

This research is supported by the EU IST FP7 project i3Dpost.

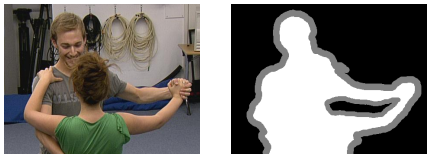


Fig. 1. Trimap: Foreground (white), background (black) and unknown (gray) regions.

techniques like [6, 7] based on local affinities have been proposed. Closed-form approach [6] fits a linear colour model to in a local window, thus defining a quadratic cost function which is then globally minimised to estimate α matte. Robust matting [7] combines affinity similar to [6] with high confidence colour samples to get a matting energy function which is then minimised to estimate alpha values.

Matting problem becomes more challenging for dynamic foreground objects in a video sequences. Previously, optical flow was successfully used in [8] to propagate trimaps from user defined key frames to the rest of the video sequence. Optical flow is often erroneous especially for large blurry motions increasing the required manual interaction. Recently techniques based on rotoscoping [9] and graph cut [10] are also been used as a semi-automated trimap generating system for videos.

3. PROPOSED APPROACH

Let us define the basic notations used in the paper. Frame at time t is represented by I^t. The colour of a pixel p in I^t is represented by I^t_p. The corresponding trimap is represented by $\mathcal{T}^t$. The trimap label for a pixel p is assigned as foreground $\mathcal{F}$, background $\mathcal{B}$ or unknown $\mathcal{U}$ and denoted by $\mathcal{T}^t_p$. A confidence map is also assigned to each trimap and its pixel-wise value is given as $\mathcal{C}^t_p$. A clean background plate is denoted by B. A detailed description is as follows.

3.1. Manual interaction

The user manually defines a trimap $\mathcal{T}^t$ for a frame I^t. A confidence map $\mathcal{C}^t$ is also associated to $\mathcal{T}^t$, where definite foreground and background pixels are assigned the highest confidence of 1 while the unknown pixels are given a confidence of zero. Additional key-frame trimaps can be added to allow for temporal changes in the foreground appearance.

3.2. Local pixel-wise background model $\mathcal{M}^{LB}$

A clean background plate B is either captured explicitly or learnt from the available video sequence. We use Gaussian mixtures to model the local background appearance variations at pixel level caused due to movement, changes in illumination and camera noise.

3.3. Global models $(\mathcal{M}^{GF}, \mathcal{M}^{GB})$

Global appearance models generalise common appearance and temporal variations reducing the number of required key-frames. A global foreground GMM $\mathcal{M}^{GF}$ in RGB colour space is estimated from the definite foreground pixels in $\mathcal{T}^t$.

Although a more precise pixel-wise local background model is already estimated, it can not model the variations observed due to the presence of foreground and temporal changes compared to the clean background plate B. The global background GMM $\mathcal{M}^{GB}$ is used to model these variations. All the background pixels in the trimap $\mathcal{T}^t$ that are not modelled by the local background model $\mathcal{M}^{LB}$ are used to learnt the global background model $\mathcal{M}^{GB}$.

3.4. Initial trimap label propagation

As the foreground scene is a person or an object who may move between successive frames it is not possible to model the local foreground statistics on a per-pixel basis a priori without knowledge of the foreground scene motion. If the global foreground and local background appearance at a particular pixel are similar a local foreground model of pixel statistics with a narrow per-pixel distribution is required to accurately label foreground pixels. The discussion on the local foreground model is delayed until step 3.5.

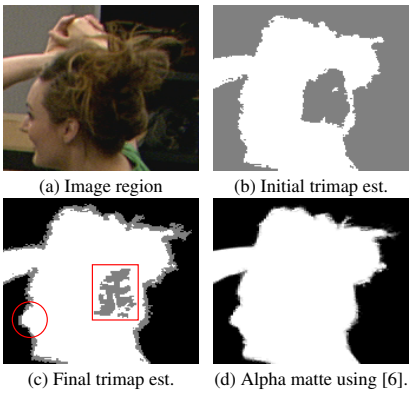
In the case of pixels where there is an ambiguity between foreground and background model membership it is preferable to label them as unknown rather than incorrectly label pixels. This is referred as an initial temporal trimap propagation and results in a binary segmentation represented as $\mathcal{T}^{t+1}_{init}$. The foreground trimap label $\mathcal{F}$ is propagated to the pixels in the frame I^{t+1} using maximum a posterior, MAP, estimation of labels based on appearance models. The MAP estimates are obtained by maximising the Baye's theorem over the entire set of model components $\mathcal{M}_i$ as

$$(\mu_{ml}, \Sigma_{ml})_{\mathcal{M}_{ml}} = \arg\max_{\mathcal{M}_i} \frac{p(x = q \mid \mu_i, \Sigma_i) p(\mu_i, \Sigma_i)}{p(x = q)}, \tag{2}$$

where $p(x = q \mid \mu_i, \Sigma_i)$ is the conditional probability of pixel q given the model component $\mathcal{M}_i$ with mean μ_i and covariance Σ_i. The term $p(\mu_i, \Sigma_i)$ is the prior for the i^{th} cluster while $p(x = q)$ is the prior for pixel q and is independent of the cluster parameters therefore ignored in optimisation. Separate MAP estimates are obtained using equation (2) over $\mathcal{M}^{GF}$ and $\mathcal{M}^{GB}$ for pixel q to find the most likely global foreground and background components, $\mathcal{M}^{GF}_{ml}$, $\mathcal{M}^{GB}_{ml}$ respectively. Let us refer the local background model as $\mathcal{M}^{LB}_{ml}$.

Typically MAP estimate corresponds to the minimum squared Mahalanobis distances $\mathcal{Q}_{min}$. Since $\mathcal{Q}$ follows the chi-square distribution we use χ^2 statistical inference to infer a trimap label for a pixel q as either foreground $\mathcal{F}$ or as unknown $\mathcal{U}$ using three separate null hypotheses as

$$\mathcal{H}^m_0 : q \in \mathcal{M}^m_{ml} \mid \mathcal{Q}^m_{min} \leq \chi^2_{\gamma,d}, \;\; m = \{GF, GB, LB\} \tag{3}$$



(a) Image region (b) Initial trimap est. (c) Final trimap est. (d) Alpha matte using [6].

Fig. 2. Example of trimap estimation.

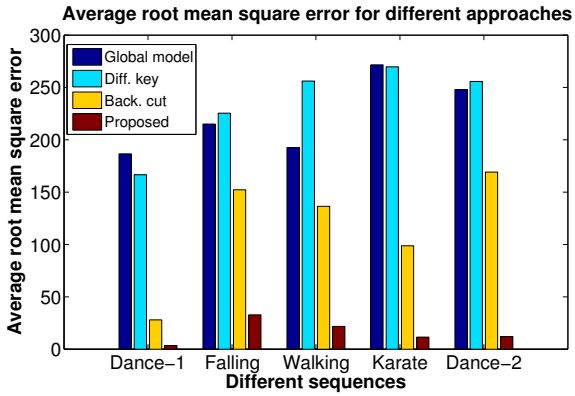


Fig. 3. Root mean square error, RMS, for different sequences.

where $\chi^2_{\gamma,d}$ is a critical value for the χ^2 distribution over d degrees of freedom at a significance level of $\gamma = 0.05$.

The initial trimap label for pixel q is propagated as

$$\mathcal{T}^{t+1}_{q,init} = \begin{cases} \mathcal{F} & \text{if, } (\mathcal{H}^{GF}_0) \wedge (\neg\mathcal{H}^{GB}_0) \wedge (\neg\mathcal{H}^{LB}_0) \\ \mathcal{U} & \text{otherwise.} \end{cases} \quad (4)$$

This initial trimap label propagation is performed independently over all the pixels in I^{t+1} as shown in Fig 2(b).

3.5. Local foreground model construction

A significant improvement in trimap labelling can be obtained by modelling the local foreground appearance. We have used the initial trimap foreground labels to estimate a local per-pixel foreground GMM $\mathcal{M}^{LF,t+1}$ by localising a circular window at each pixel.

3.6. Refinement using local foreground model

Refinement of initial trimap is performed by reassigning a trimap label to an unknown pixel q. An additional null hypothesis test based local foreground model is defined as $\mathcal{H}^{LF}_0 : q \in \mathcal{M}^{LF}_{ml} \mid \mathcal{Q}^{LF}_{min} \leq \chi^2_{\gamma,d}$ and the final trimap label assignment is performed as:

$$\mathcal{T}^{t+1}_{q} = \begin{cases} \mathcal{F} & \text{if, } (\mathcal{H}^{GF}_0 \vee \mathcal{H}^{LF}_0) \wedge (\neg\mathcal{H}^{GB}_0) \wedge (\neg\mathcal{H}^{LB}_0) \\ \mathcal{B} & \text{if, } (\mathcal{H}^{GB}_0 \vee \mathcal{H}^{LB}_0) \wedge (\neg\mathcal{H}^{GF}_0) \wedge (\neg\mathcal{H}^{LF}_0) \\ \mathcal{U} & \text{otherwise.} \end{cases} \quad (5)$$

Fig 2(c) shows the refined trimap where the foreground pixels previously labelled as ambiguous are correctly labelled.

3.7. Confidence map estimation

Estimation of per-pixel trimap label confidence is critical to temporal trimap propagation. If a pixel q has a foreground label, the confidence is formulated as

$$\mathcal{C}^{t+1}_{q} = \frac{e^{-\mathcal{Q}^{F}_{min}/2}}{(2\pi)^{3/2}|\Sigma^{F}_{ml}|^{1/2}}\left(1 - \frac{e^{-\mathcal{Q}^{B}_{min}/2}}{(2\pi)^{3/2}|\Sigma^{B}_{ml}|^{1/2}}\right)\lambda. \quad (6)$$

The introduction of the prior confidence level $\lambda = p(\mu^{F}_{ml}, \Sigma^{F}_{ml})$ ensures that the confidence of pixel labels used to dynamically update clusters is taken into account so that the model does not drift. The confidence for a background pixel is estimated in the same way by interchanging the parameters.

3.8. Alpha matte estimation

To obtain an alpha matte any image matting algorithm can be used. In this work we have used Levin et al.'s closed form solution [6] to obtain an alpha matte α^{t+1} of the frame I^{t+1} using the propagated trimap $\mathcal{T}^{t+1}$.

3.9. Global model update

The estimated confidence of trimap label is used to update the global foreground and background models. Model update is limited to foreground and background pixels which are not represented with high-confidence in the prior global models. The new clusters obtained are appended to the prior models. This approach allows the incorporation of novel appearance information without reducing the confidence of models derived from the manually specified key-frame trimap.

4. RESULTS AND EVALUATION

Results are presented for challenging high definition natural indoor and outdoor scenes containing one or more actors performing fast non rigid motion. Two key-frames are used for the indoor sequences consists of 250 frames while a single key-frame is used for the outdoor sequences having a length of 125 frames. A comparative evaluation is performed between four different approaches: (1) global model comparison using foreground and background models only from the key frame, (2) difference keying, (3) background cut [11], and (4) proposed algorithm. The obtained alpha mattes are shown in Fig 4 with error from the ground-truth in red. To compare the techniques quantitatively we have used root mean square error, RMS. The error measure is shown in Fig. 3.

Fig. 4. Alpha mattes obtained using different approaches along with the ground truth. First column shows the different key-frames used along with the hand drawn trimaps. The matte error is shown in red.

It is clear from the results that other approaches are unable to classify pixels correctly especially in the shadow region and in regions where foreground and background appearance is similar (actor's shirt and carpet). Our technique is able to reduce the errors caused by the overlap in the models by incorporating the local foreground model. The results obtained show few visible artifacts compared to the ground truth.

5. CONCLUSION

A novel temporal trimap propagation algorithm has been introduced based on a statistical inference framework. Key-frame trimaps are sparsely specified and robustly propagated across a given video sequence to perform video matting. Results are presented which demonstrate high-quality mattes for indoor and outdoor video sequences containing complex movements. The technique only requires manual trimap input for 1-2 key-frames to process sequences of several hundred frames. This is a significant reduction in manual interaction compared with state-of-the-art video matting approaches.

6. REFERENCES

[1] T. Porter and T. Duff, "Compositing digital images," in *ACM SIGGRAPH*, 1984, pp. 253–259.

[2] A. R. Smith and J. F. Blinn, "Blue screen matting," in *ACM SIGGRAPH*, 1996, pp. 259–268.

[3] P. Hillman and J. Hannah, "Natural image matting," in *Video, Vision and Graphics*, 2005, pp. 211–18.

[4] Y. Y. Chuang, B. Curless, D. H. Salesin, and R. Szeliski, "A Bayesian approach to digital matting," in *CVPR*, 2001, vol. 2, pp. 264–271.

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

[6] A. Levin, D. Lischinski, and Y. Weiss, "A closed form solution to natural image matting," *CVPR*, pp. 61–68, 2006.

[7] J. Wang and M. F. Cohen, "Optimized color sampling for robust matting," in *CVPR*, 2007, pp. 1–8.

[8] Y. Y. Chuang, A. Agarwala, B. Curless, D. Salesin, and R. Szeliski, "Video matting of complex scenes," in *ACM SIGGRAPH*, 2002, pp. 243–248.

[9] A. Agarwala, A. Hertzmann, D. H. Salesin, and S. M. Seitz, "Keyframe-based tracking for rotoscoping and animation," *ACM Trans. Graph.*, vol. 23, no. 3, pp. 584–591, 2004.

[10] Y. Li, J. Sun, and H. Y. Shum, "Video object cut and paste," *ACM ToG.*, vol. 24, no. 3, pp. 595–600, 2005.

[11] Jian Sun, Weiwei Zhang, Xiaoou Tang, and Heung-Yeung Shum, "Background cut," in *ECCV*, 2006, pp. 628–641.