

NATURAL IMAGE MATTING FOR MULTIPLE WIDE-BASELINE VIEWS

Muhammad Sarim, Adrian Hilton, Jean-Yves Guillemaut, Takeshi Takai, Hansung Kim

Centre of Vision Speech and Signal Processing
University of Surrey, Guildford GU2 7XH, Surrey, United Kingdom.
{m.farooqui, a.hilton, j.guillemaut, t.takai, h.kim}@surrey.ac.uk

ABSTRACT

In this paper we present a novel approach to estimate the alpha mattes of a foreground object captured by a wide-baseline circular camera rig provided a single key frame trimap. Bayesian inference coupled with camera calibration information are used to propagate high confidence trimaps labels across the views. Recent techniques have been developed to estimate an alpha matte of an image using multiple views but they are limited to narrow baseline views with low foreground variation. The proposed wide-baseline trimap propagation is robust to inter-view foreground appearance changes, shadows and similarity in foreground/background appearance for cameras with opposing views enabling high quality alpha matte extraction using any state-of-the-art image matting algorithm.

Index Terms— Image matting, alpha matte, trimap, wide-baseline, multiple views.

1. INTRODUCTION

In computer vision digital image matting is considered a classical problem where a foreground object is extracted from a scene along with its opacity for compositing in a new background. The increasing demand of special effects in the media industry and the concept of virtual reality have triggered the extensive study of the matting problem. An image can be described as a combination of three layers, the foreground and the background layers blended together using an opacity layer called alpha matte. Porter and Duff [1] first put forward the mathematical relation between these layers known as the compositing equation as

$$C = \alpha F + (1 - \alpha)B, \quad (1)$$

where C represents the composite while F , B and α are the foreground, background layers and alpha matte which are identical in pixel dimensions. The alpha matte defines the pixel-wise foreground opacity and has floating point values in the range of $[0, 1]$, the extreme values represent the background and foreground colour respectively while the

intermediate values show their blending proportion due to object transparency and mixed pixels at object boundaries.

Equation (1) is under-constrained as the only known variable is the composite C . Studio environment with a known homogeneous background colour, typically blue or green, is used to constrain (1). The assumption on the background colour not to appear in the foreground leads to a simple solution of compositing (1) for alpha. Such constraints are not available in a natural image therefore a manual interaction in the form of a trimap is provided to aid the definition of the foreground and background regions. The problem is then to estimate the α value for the pixels in the unknown region given the definite trimap labels. Several solutions [2, 3, 4, 5] have been proposed to estimate a high quality alpha matte by exploiting the statistics of the definite regions. In recent techniques like [6, 7, 8, 9, 10], single view matting approaches have been extended to narrow-baseline multiple views by assuming primarily the invariant inter-view foreground appearance. Wide-baseline views of the same foreground present an extremely challenging matting problem caused by the significant appearance variations resulting due to occlusion, projective distortion, shadows and motion blur due to camera or scene movement.

In this paper we present a novel approach for wide-baseline image matting given only a single key frame trimap and set of calibrated cameras. The proposed technique reduces the manual interaction required for wide-baseline matting by limiting interaction to just one view. Quantitative evaluation demonstrate that the proposed algorithm is capable to estimate alpha matte for wide-baseline views (up to 180°) comparable to that achieved by providing manual interaction to individual views.

2. RELATED WORK

Natural image matting is an extensively studied problem in computer vision. There are no constraints on the foreground and background appearance in natural images imposing a requirement of manual interaction in the form of a trimap which defines the definite foreground, background and unknown region. The matting algorithms utilise the image statistics of the definite regions to estimate the final alpha matte. Ap-

Thanks to the financial support of the EU IST FP7 project i3DPost.

proaches like [2] fit statistical models to the local definite foreground and background region and the assumption that the region appearance is locally consistent provides the solution to (1) for α . The assumption require accurate trimaps to extract a good alpha matte. To overcome the requirement of an accurate trimap [3, 5] have used local affinities. Poisson matting [5] assumes the intensity variations are locally smooth to estimate the alpha matte. In [3] a linear model is fitted to the foreground and background colour by assuming they are locally smooth providing a closed-form solution for alpha. The propagation behavior of affinity based techniques make them prone to error accumulation. Recently [4] used a non-parametric template approach to allow the representation of local appearance structure and avoid smoothness assumptions to extract comparable high quality mattes.

In [7, 8] single view matting is formulated using triangulation between narrow-baseline views ($< 5^\circ$) with an assumption of invariant foreground and different background. A thresholded inter-view variance image is used to generate a trimap and then the local variance information is utilised to estimate the alpha matte. Planar motion of a rigid foreground across backgrounds in multiple images is assumed in [9] to extract an alpha matte by using a Bayesian framework and background mosaic. The matting problem is formulated as an estimation of a 3D boundary curve in [6], foreground and background colours are estimated using the depth information from multiple narrow-baseline views making it prone to stereo inaccuracies. In [10], rectified stereo pairs in a Gaussian pyramid are used to construct a high dimensional feature space and a local neighbor embedding is used to generate a trimap followed by α matte estimation [2]. Binary foreground/background segmentation is performed in [11] using graph-cut optimisation. A fixation constraint is used to seed the pixels for the foreground colour model. Similar foreground and background colour distributions cause unacceptable results.

Current multi-view matting algorithms generally assume similar local foreground appearance across the views and are thus limited to narrow-baseline ($< 10^\circ$) capture. In this paper we address the problem of wide-baseline views ($> 30^\circ$) having significantly different appearance without making assumptions on foreground colour and position across multiple views.

3. WIDE-BASELINE IMAGE MATTING

The problem is formally stated as follows: For a set of N wide-baseline views $\{\mathcal{I}^v\}_{v=1}^N$ and their clean background plates $\{\mathcal{B}^v\}_{v=1}^N$ estimate the alpha matte of each image, $\{\alpha^v\}_{v=1}^N$ provided a single manually generated key frame trimap, \mathcal{T}^k . We assume that the cameras are static and synchronised with know calibration parameters. Using a Bayesian inference framework and calibration information our goal is to propagate the trimap labels of \mathcal{T}^k across multiple views to obtain their trimaps \mathcal{T}^v and corresponding con-

fidence maps \mathcal{C}^v . Once a trimap is estimated, a conventional single view matting algorithm is used to estimate the alpha matte. To adaptively model the variations in foreground and background appearance across the views caused by shadows and illumination variations, the foreground and background models are updated after matte estimation for every view.

3.1. Trimap Estimation

To propagate the trimap labels, statistics of the definite know regions are utilised by constructing global foreground and background models, $\{\mathcal{M}^{GF}, \mathcal{M}^{GB}\}$, using any state-of-the-art clustering algorithm. Each of these appearance models is represented as a mixture of multivariate weighted Gaussians in colour space, $\{\mathcal{N}(\hat{\mu}_i, \hat{\Sigma}_i)\}_{i=1}^n$, with each component weighted by a confidence λ_i defined by the confidence of the member pixels. Since the wide-baseline views have non-overlapping backgrounds, the global background model is restricted to the shadow region and a separate pixel-wise background model for each view, $\{\mathcal{M}^{B,v}\}_{v=1}^N$, is constructed using the corresponding clean background plate and a constant covariance. The trimap estimation process, Fig 1, follows as: (1) region isolation, (2) initial trimap label propagation, (3) confidence map estimation and (4) trimap refinement.

3.1.1. Region Isolation

Isolating the shadow region \mathcal{R}_s in the given image \mathcal{I}^v using epipolar geometry minimises the misclassification of the pixels due to the similarity in the foreground and background appearance. Given a pair of calibrated cameras, the epipolar constraint for a point in one view defines a line, epipolar line, in the other view on which the corresponding image point must lie. Using the epipolar constraint we can isolate the shadow region \mathcal{R}_s , Fig 1(b), in the given image \mathcal{I}^v by drawing epipolar line for every pixel in the manually defined shadow region, Fig 1(a), in the key frame \mathcal{I}^k and, if available, estimated shadow region in the view \mathcal{I}^{v-1} . The final isolation is done by dilating the epipolar lines by a few pixels to account for calibration errors.

3.1.2. Initial trimap label propagation

Maximum a posterior estimates are used to initially propagate the trimap label to the pixels in \mathcal{I}^v . The posterior probability of the pixel q belonging to the i^{th} component of a model $\mathcal{M}_i(\hat{\mu}_i, \hat{\Sigma}_i)$ with mean $\hat{\mu}_i$ and covariance $\hat{\Sigma}_i$ is given by the Bayes theorem as:

$$P(\hat{\mu}_i, \hat{\Sigma}_i | x = q) = \frac{P(x=q | \hat{\mu}_i, \hat{\Sigma}_i)P(\hat{\mu}_i, \hat{\Sigma}_i)}{P(x=q)} \quad (2)$$

The term $p(\hat{\mu}_i, \hat{\Sigma}_i)$ is the prior for the cluster and is given by the cluster confidence, λ_i . $P(x = q)$ is the prior for pixel q and is independent of the models and therefore is ignored in maximising (2) to obtain MAP estimates $(\hat{\mu}_{ml}, \hat{\Sigma}_{ml})$ as

$$(\hat{\mu}_{ml}, \hat{\Sigma}_{ml})_{\mathcal{M}_{ml}} = \arg \max_{\mathcal{M}} p(x = q | \hat{\mu}_i, \hat{\Sigma}_i) \lambda_i. \quad (3)$$

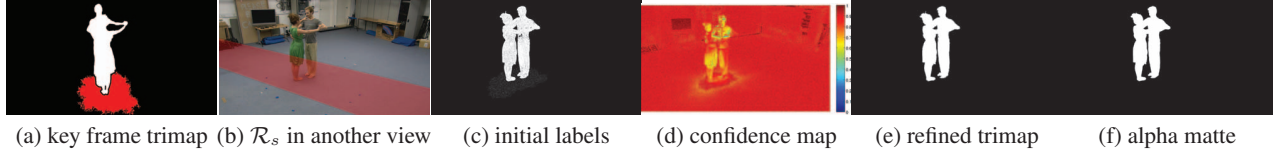


Fig. 1. Steps in wide-baseline image matting, (a) shadow region marked as red and (b) isolated \mathcal{R}_s in a different view masked with red.

Separate MAP estimates are obtained using (3) to find the most likely global foreground and background models, $\{\mathcal{M}_{ml}^{GF}, \mathcal{M}_{ml}^{GB}\}$ corresponding to minimum squared Mahalanobis distance $\{\mathcal{D}_{min}^{GF}, \mathcal{D}_{min}^{GB}\}$. The squared Mahalanobis distance of pixel q from pixel-wise background model $\mathcal{M}^{LB,v}$ is given by \mathcal{D}^{LB} . Mahalanobis distance is chi-square distributed, $\mathcal{D} \sim \chi^2(d)$ over $d = 3$ degree of freedom for RGB colour space, we use inferential statistics based on the χ^2 test to infer the trimap label for the pixel q . Three separate null hypothesis are defined under a significance test using a critical value of $\chi_{\gamma,d}^2$ at a significance level of γ as

$$\begin{aligned} \mathcal{H}_0^{GF} &: q \in \mathcal{M}_{ml}^{GF} & | & \mathcal{D}_{min}^{GF} \leq \chi_{\gamma,d}^2 \\ \mathcal{H}_0^{GB} &: q \in \mathcal{M}_{ml}^{GB} & | & \mathcal{D}_{min}^{GB} \leq \chi_{\gamma,d}^2 \\ \mathcal{H}_0^{LB} &: q \in \mathcal{M}^{LB,v} & | & \mathcal{D}^{LB} \leq \chi_{\gamma,d}^2 \end{aligned} \quad (4)$$

The trimap label $\mathcal{T}^v(q)$ is propagated to pixel q in the trimap \mathcal{T}^v corresponding to \mathcal{I}^v according to Tab. 1. The initial propagated trimap is shown in Fig 1(c).

3.1.3. Confidence Map Estimation

We can associate a confidence level to each pixel in the trimap \mathcal{T}^v to construct the confidence map \mathcal{C}^v , Fig 1(d), which is used to refine the initial trimap. The confidence levels of the most likely global foreground and global or local background clusters, $\{\lambda_{ml}^f, \lambda_{ml}^b\}$, and the corresponding minimum squared Mahalanobis distances, $\{\mathcal{D}_{min}^f, \mathcal{D}_{min}^b\}$, are used to formulate the confidence function. If pixel q corresponds to the foreground hypothesis its confidence is assigned as ,

$$\mathcal{C}^v(q) = \lambda_{ml}^f (1 - e^{-\mathcal{D}_{min}^b / \chi_{\gamma,d}^2}) e^{-\mathcal{D}_{min}^f / \chi_{\gamma,d}^2}, \quad (5)$$

where, $\chi_{\gamma,d}^2$ is the critical value at significance level of γ over d degrees of freedom. The confidence for a background pixel is estimated in a similar way.

3.1.4. Trimap Refinement

To refine the trimap, a local foreground model \mathcal{M}^{LF} is constructed, for every low confidence background pixel, using the high confidence foreground pixels within a circular window of dimension r . A null hypothesis that pixel $q \in \mathcal{M}^{LF}$ is defined as $\mathcal{H}_0^{lf} : q \in \mathcal{M}_{ml}^{LF} \mid \mathcal{D}_{min}^{lf} \leq \chi_{\gamma,d}^2$ based on its minimum squared Mahalanobis distance \mathcal{D}_{min}^{lf} to the most likely cluster $\mathcal{M}_{ml}^{LF} \in \mathcal{M}^{LF}$ for a critical value of $\chi_{\gamma,d}^2$ at a significance level of γ . The trimap label for all pixels satisfying

| Region | \mathcal{H}_0^{GF} | \mathcal{H}_0^{GB} | \mathcal{H}_0^{LB} | $\mathcal{T}^v(q)$ |
|--|----------------------|----------------------|----------------------|--------------------|
| $q \in \mathcal{R}_s$ | <i>true</i> | <i>false</i> | <i>false</i> | <i>foregrd.</i> |
| (shadow) | <i>false</i> | <i>true</i> | <i>false</i> | <i>backgrd.</i> |
| $q \in \neg\mathcal{R}_s$ | <i>true</i> | <i>true / false</i> | <i>false</i> | <i>foregrd.</i> |
| (not shadow) | <i>false</i> | <i>true / false</i> | <i>true</i> | <i>backgrd.</i> |
| $q \in (\mathcal{R}_s \cup \neg\mathcal{R}_s)$ | <i>otherwise</i> | | | <i>unknown</i> |

Table 1. Trimap label assignment for pixel q .

the null hypothesis \mathcal{H}_0^{lf} is reassigned to unknown U and their confidence is reevaluated using (5) with \mathcal{D}_{min}^{lf} . The refined trimap is shown in Fig 1(e).

3.2. Matte Estimation

Once the trimap \mathcal{T}^v for the view \mathcal{I}^v is estimated, the alpha matte α^v can be recovered using a state-of-the-art matting algorithm such as [3, 4], shown in Fig 1(f).

3.3. Global Model Update

To model the variations in the foreground and background appearance due to the shadows and intensity variations the global models are updated prior to processing the image \mathcal{I}^{v+1} . After estimating the final alpha matte of the view \mathcal{I}^v , the background pixels in the shadow region and the foreground pixels which were marked as unknown U in the trimap propagation step are separately modeled as mixture of Gaussians. The global models are updated by appending them with these new lower confidence observations.

4. RESULTS AND EVALUATION

We have used high definition static, synchronised and calibrated cameras in a circular rig with an angular separation of $\approx 45^\circ$ to capture an indoor and outdoor scene. Key frames for a single view and manually drawn trimaps are shown in Fig 2 along with the estimated alpha mattes for other views using different methods: (1) difference keying, (2) global modeling, (3) background cut [12] and (4) the proposed wide-baseline technique. Difference keying and global model comparison are not able to remove the shadow region while background cut does a better classification of shadow pixels as background but fails to define correct foreground where the foreground colour distribution is similar to the shadow region in the indoor scene and produced large artifacts in the uncontrolled outdoor scene. Visual comparison of the mattes produced by the proposed algorithm to the ground truth shows that the foreground boundary is properly defined and

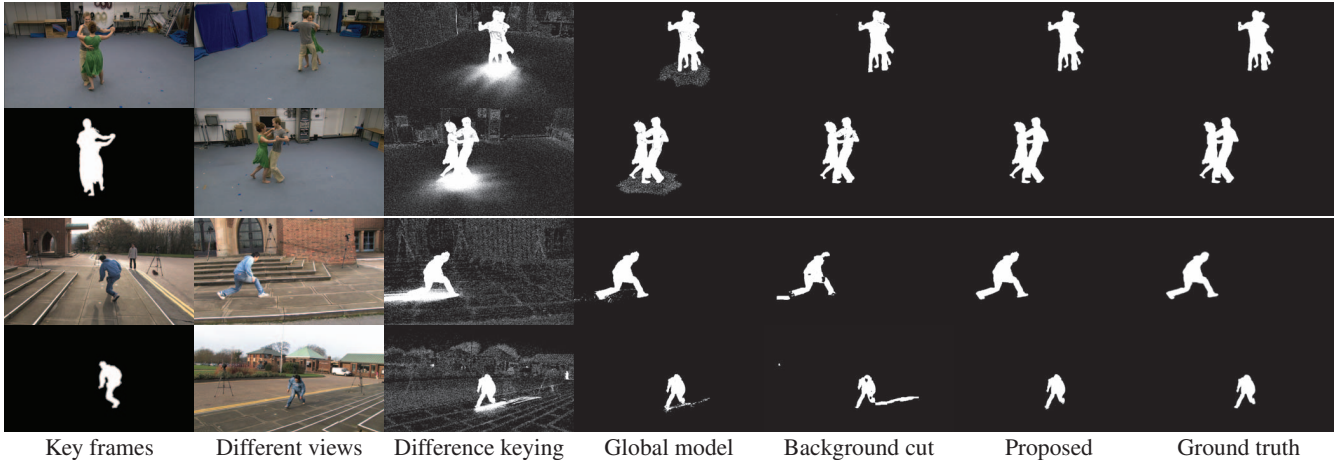


Fig. 2. Estimated alpha mattes obtained using different approaches along with the ground truths for a indoor and outdoor scene. First column shows the key frames used along with the hand drawn trimaps.

is consistent across the views and the mattes have no visible artifacts. Root mean square error RMS is used to compare the techniques quantitatively and the plot Fig 4 for all the views shows that the proposed algorithm clearly outperform other techniques.

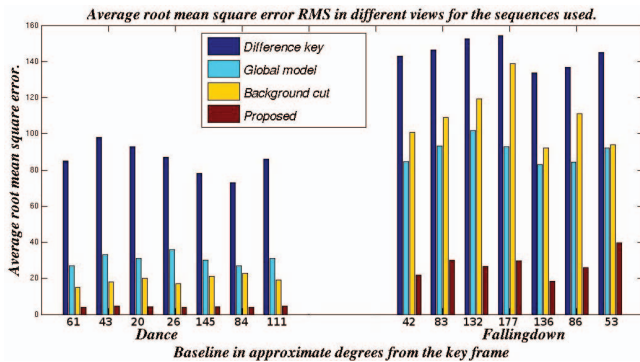


Fig. 3. RMS for all the views in the scenes used.

5. CONCLUSION

A novel wide-baseline image matting algorithm is presented using a Bayesian inference framework for propagating high quality trimap labels across multiple views using a single key frame trimap. The technique overcomes limitations of assumptions of previous narrow-baseline approaches for foreground appearance and the requirement of manual interaction for each wide-baseline view. The results presented demonstrate high quality alpha matte estimation for views having an orientation of up to 180° from the key frame. Future work will focus on the extension of the proposed technique to wide-baseline video sequences.

6. REFERENCES

- [1] T. Porter and T. Duff, “Compositing digital images,” in *ACM SIGGRAPH*, 1984, pp. 253–259.
- [2] Y. Y. Chuang, B. Curless, D. H. Salesin, and R. Szeliski, “A Bayesian approach to digital matting,” in *CVPR ’01*.
- [3] A. Levin, D. Lischinski, and Y. Weiss, “A closed form solution to natural image matting,” *CVPR ’06*.
- [4] M. Sarim, A. Hilton, J.-Y. Guillemaut, and H. kim, “Non-parametric natural image matting,” in *ICIP ’09*.
- [5] J. Sun, J. Jia, C.K. Tang, and H. Y. Shum, “Poisson matting,” *ACM TOG*, vol. 23, no. 3, pp. 315–321, 2004.
- [6] S. W. Hasinoff, S. B. Kang, and R. Szeliski, “Boundary matting for view synthesis,” in *CVPR ’04*.
- [7] M. H. Hyun, S. Y. Kim, and Y. S. Ho, “Multi-view image matting and compositing using trimap sharing for natural 3-d scene generation,” in *3DTV ’08*.
- [8] N. Joshi, W. Matusik, and S. Avidan, “Natural video matting using camera arrays,” *ACM Trans. Graph.*, vol. 25, no. 3, pp. 779–786, 2006.
- [9] Y. Wexler, A. W. Fitzgibbon, and A. Zisserman, “Bayesian estimation of layers from multiple images,” in *ECCV ’02*.
- [10] K. H. Won, S. Y. Park, and S. K. Jung, “Natural image matting based on neighbor embedding,” in *MIRAGE’07*.
- [11] N. D. F. Campbell, G. Vogiatzis, C. Hernandez, and R. Cipolla, “Automatic 3d object segmentation in multiple views using volumetric graph-cuts,” *Image and Vision Computing*, 2008.
- [12] J. Sun, W. Zhang, X. Tang, and H.-Y. Shum, “Background cut,” in *ECCV (2)*, 2006, pp. 628–641.