A Bayesian Framework for Simultaneous Matting and 3D Reconstruction

J.-Y. Guillemaut, A. Hilton, J. Starck, J. Kilner
University of Surrey, U.K.
J.Guillemaut@surrey.ac.uk

O. Grau
BBC Research, U.K.
Oliver.Grau@bbc.co.uk

Abstract

Conventional approaches to 3D scene reconstruction often treat matting and reconstruction as two separate problems, with matting a prerequisite to reconstruction. The problem with such an approach is that it requires taking irreversible decisions at the first stage, which may translate into reconstruction errors at the second stage. In this paper, we propose an approach which attempts to solve both problems jointly, thereby avoiding this limitation. A general Bayesian formulation for estimating opacity and depth with respect to a reference camera is developed. In addition, it is demonstrated that in the special case of binary opacity values (background/foreground) and discrete depth values, a global solution can be obtained via a single graph-cut computation. We demonstrate the application of the method to novel view synthesis in the case of a large-scale outdoor scene. An experimental comparison with a two-stage approach based on chroma-keying and shape-from-silhouette illustrates the advantages of the new method.

1. Introduction

In this paper, we study the problem of simultaneous matting and 3D reconstruction of a scene from a set of images.

The *matting* problem [3, 18, 7] consists in assigning opacity values α between 0 and 1 to image pixels (0 for background pixels, 1 for foreground pixels, and other values for mixed pixels which simultaneously see background and foreground). Image matting approaches were initially introduced for single images [3] and more recently extended to image sequences [18, 7] although often restricted to input of a trimap (partial labelling of background/foreground) at key-frames and special camera configurations. Image matting in natural outdoor scenes remains an open problem due to visual ambiguities.

The *reconstruction* problem consists in inferring depth information from a collection of images. Earlier approaches reasoned in the image space by matching sparse or dense features across images [12]. In contrast, volumetric methods such as shape-from-silhouette [10] or voxel colouring [14, 9] reason directly in the 3D space by assessing the occupancy or emptiness of voxels in a grid. In the case of shape-from-silhouette, the decision is made by establishing whether voxels belong to the intersection of the cones back-projected from image silhouettes (foreground/background segmentation), resulting in a reconstruction called the visual hull, which provides an upper bound on scene reconstruction (in principle, it is guaranteed to enclose the true scene surface). A review of multi-view scene reconstruction algorithms can be found in [13].

Many approaches start by applying a matting algorithm independently to the images in order to compute a foreground/background segmentation, which is then used as an input to the reconstruction algorithm. The problem with this approach is that hard decisions are made at the matting stage, which may not be possible to correct at the reconstruction stage, thus affecting the final reconstruction. In shape-from-silhouette, for example, misclassification of a foreground region as background will erode the visual hull. A naive solution would be to build a conservative visual hull by allowing a tolerance on the intersection of the back-projected cones. Unfortunately, although this may be sufficient to guarantee that the real scene is contained in the visual hull in spite of matting or camera calibration errors, this produces a dilated representation of the scene, which is inaccurate. The solution proposed in this paper jointly formulates the matting and reconstruction problems, by combining information from multiple views.

1.1. Previous work

In [16], Szeliski and Golland proposed a stereo approach which estimates disparities, colours and opacities in a generalised disparity space. They formulate the problem in terms of energy minimisation, whose solution is obtained with an iterative gradient descent algorithm. Similarly, De Bonet and Viola proposed the Roxel algorithm [4] which defines an iterative multi-step procedure alternatively estimating colours, responsabilities and opacities in a voxel space. In general computing opacity values in addition to the scene geometry significantly increases the difficulty of the problem.

A simpler approach, which is plausible for scenes not

containing transparent objects or fractal surfaces, consists in restraining the problem to the segmentation of the scene into foreground and background layers; this is equivalent to restricting the previous formulations to binary opacity values. In [8], Kolmogorov *et al.* proposed two stereo algorithms, based on dynamic programming and graph-cuts respectively, and are able to achieve real-time segmentation. In [19], Zitnick *et al.* also adopted a layered representation where a colour segmentation-based stereo algorithm is used to compute a smooth disparity map for each camera, which is then refined by Bayesian matting [3] applied to 4-pixel thick boundary strips located at discontinuity jumps.

In [15] and [6], the segmentation and reconstruction problems are formulated jointly by minimising an energy function via a graph-cut algorithm. These two methods, like the method proposed in this paper, require the use of additional background images captured from the same viewpoint as the observed images. In the case of [15], the method is effectively a generalisation of shape-from-silhouette techniques. The main limitation of this approach is that silhouette intersection constraints are usually weaker than photo-consistency constraints. In the case of [6], the method enforces a more powerful photo-consistency constraint across multiple views. Although the method is similar in principle to our method, it has the following limitations: i) it is limited to special camera configurations for which a visibility constraint similar to the one used in voxel colouring [14, 9] can be defined, ii) it requires a prior estimate of the background geometry, iii) only a locally optimal solution (in a strong sense) is obtainable. A potential advantage of this approach is full 3D scene reconstruction instead of our 2.5D camera dependent representation.

1.2. Our approach

We formulate the problem in terms of recovering depth and opacity values with respect to a reference camera given a set of images of a scene and a background image for each camera. Background images do not necessary need to be captured, in our implementation they are estimated directly from a sequence of images. We express the problem in terms of maximising the *a posteriori* probability given the set of input images and some strong priors on shape and alpha mattes. Although finding a general solution is difficult, we argue that a global solution is obtainable in a single graph-cut computation in the case of binary opacity values (foreground/background) and discrete depth values. The binary opacity assumption is plausible for the type of scene considered since there are no transparent objects and the scene surface is smooth. Note that in the binary case, matting is commonly referred to as foreground/background segmentation in the literature.

Our contributions are the following. Firstly, we show the advantage of using photoconsistency constraints derived from multiple views in matting, and propose a novel N-view algorithm for jointly solving the matting and reconstruction problems. The method is not restricted to pairs of cameras separated by a small baseline unlike [16, 19, 8], and there is no restriction on camera positioning unlike [6]. Secondly, we propose a novel matching score which incorporates background information in order to disambiguate conventional matching scores, and allows accurate matting in spite of possible background occlusions. The novel matching score is particularly useful when trying to establish correspondences in a scene viewed against a uniform background which tends to exacerbate the matching ambiguity. An advantage of our method compared to [6] is that it does not require a prior estimate of the background geometry.

The paper is structured as follows. After formulating the problem in mathematical terms, we introduce the general Bayesian framework. We then show how a global solution can be computed via a single graph-cut under the assumption of binary opacity values and discrete depth values. Finally we compare the method developed with a conventional reconstruction method on real images of a large-scale outdoor scene and conclude.

2. Problem formulation and notation

A scene is viewed from $N+1$ cameras indexed from 0 to N, the camera with index 0 being the reference camera. A pixel in an image is represented by a vector $\boldsymbol{p} = [u, v]$ of image coordinates; $\mathcal{P}$ denotes the set of all pixel coordinates in the reference image. A 3D point is represented by a vector $\boldsymbol{P} = [x, y, z]$. The world reference frame is defined by the reference camera, such that a 3D point with coordinates $[u, v, d]$ corresponds to the point on the ray backprojected through the image point $[u, v]$ and located at a distance d from the reference image plane. All cameras are assumed to have been geometrically calibrated so that their projection matrices M^i are known. The projection of the point $[x, y, z]$ in camera i is written $\mathsf{M}^i[x, y, z]$ for simplicity; note that to be rigorous we should have used homogeneous coordinates and written $\mathsf{M}^i[x, y, z, 1]^\top$ instead. For each camera, two images $\mathcal{C}^i$ and $\mathcal{B}^i$ are available. $\mathcal{C}^i$, called the composite image, is an image of the full scene (foreground and background), while $\mathcal{B}^i$ is an image of the background only; both images correspond to the same viewpoint and camera settings. Note that when describing a set of images, the index range is usually omitted for conciseness; for example $\{\mathcal{C}^i\}$ stands for $\{\mathcal{C}^i\}_{0 \leq i \leq N}$. The objective of the problem is to simultaneously estimate, in the reference camera, i) the opacity of each pixel, and ii) the depth of the foreground pixels. The scene depth and opacity with respect to the reference camera are represented respectively by a depth image d and an alpha matte α^0. With these notations, $d_{\boldsymbol{p}}$ and $\alpha^0_{\boldsymbol{p}}$ represent the depth and opacity of a pixel $\boldsymbol{p}$ in the reference

camera. Note that the label $d_{\boldsymbol{p}} = \infty$ is reserved for background pixels which are not assigned any physical depth. An alpha matte α^i is defined for each camera. Opacity values are constrained to be between 0 and 1, opacities of 0 and 1 representing a background and a foreground point respectively, while other values correspond to mixed pixels. The latter type of pixels occurs if the foreground is semi-transparent or when the cone defined by the backprojection of a pixel grazes the foreground surface and captures simultaneously foreground and background.

3. Bayesian framework

In a Bayesian framework, the optimum reference depth map d and alpha mattes $\{\alpha^i\}$ are estimated by maximising the posterior probability, or equivalently, in terms of log likelihoods:

$$\begin{aligned} L(d, \{\alpha^i\}|\{\mathcal{C}^i, \mathcal{B}^i\}) = L(\{\mathcal{C}^i, \mathcal{B}^i\}|d, \{\alpha^i\}) \\ +L(d) + L(\{\alpha^i\}) - L(\{\mathcal{C}^i, \mathcal{B}^i\}). \end{aligned} \quad (1)$$

The first term is the log likelihood, while the other terms are priors. The term $L(\{\mathcal{C}^i, \mathcal{B}^i\})$ being constant with respect to the optimisation variables does not contribute and can be ignored. The remaining terms are expressed in this section.

3.1. Likelihood

3.1.1. Conventional model. A conventional approach would model the composite colour for any camera i in which the point $\boldsymbol{P} = [\boldsymbol{p}, d_{\boldsymbol{p}}]$ is visible as:

$$\mathcal{C}^i_{M^i\boldsymbol{P}} = \mathcal{C}^0_{M^0\boldsymbol{P}} + \boldsymbol{\eta}^i_1, \quad (2)$$

where $\{\boldsymbol{\eta}^i_1\}$ represents the image noise and view dependent appearance variations, and $\mathcal{C}^0_{M^0\boldsymbol{P}}$ is the composite colour in the reference camera (for which the point $\boldsymbol{P}$ is assumed to be visible). From this model, in a conventional stereo reconstruction, we would write:

$$L(\{\mathcal{C}^i, \mathcal{B}^i\}|d, \{\alpha^i\}) = \sum_{\boldsymbol{p}\in\mathcal{P}} E_1(\{\mathcal{C}^i\}, [\boldsymbol{p}, d_{\boldsymbol{p}}]), \quad (3)$$

$$\text{with} \quad E_1(\{\mathcal{C}^i\}, \boldsymbol{P}) = -\frac{1}{|\mathcal{V}(\boldsymbol{P})|} \sum_{i\in\mathcal{V}(\boldsymbol{P})} \|\mathcal{C}^i_{M^i\boldsymbol{P}} - \mathcal{C}^0_{M^0\boldsymbol{P}}\|^2. \quad (4)$$

$\mathcal{V}(\boldsymbol{P})$ represents the set of camera indices for which the point $\boldsymbol{P}$ is visible, and $|\mathcal{V}(\boldsymbol{P})|$ denotes the cardinality (number of elements) of this set. The method used to assess the visibility will be described in Section 4. A point which is not visible in any camera, would be assigned for example a fixed penalty score or the score obtained without considering visibility (this score would be expected to be low in that case). For robustness, the intensity difference computed in Eq. (4) can be replaced by the sum of squared difference (SSD) or the Normalised Cross Correlation (NCC) computed over a window. In the two camera case, this formulation is equivalent to the one used in stereo matching (see *e.g.* [12]), while the N-camera generalisation is similar to a colour-consistency measure (see *e.g.* [14, 9]). Such a formulation assumes an opaque scene, and neglects mixed pixels at object boundaries. We illustrate below two other important limitations to this approach.

3.1.2. Limitation 1: background visibility. A conventional approach usually works well for reconstructing foreground points under a small baseline assumption, however this becomes ambiguous when trying to reconstruct background points because of potential foreground occluders (see Fig. 1) or even because, in the case of a larger baseline, the background seen by a camera may not be in the field of view of the other cameras. For this reason, it is unrealistic to obtain accurate matting results unless a small baseline and a large number of cameras are considered to ensure that explicit reconstruction of the background is possible.

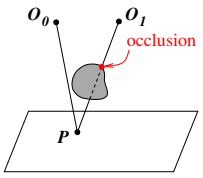


Figure 1. Example of background ambiguity. The background is visible only in camera O_0, therefore it is not possible to evaluate photoconsistency.

3.1.3. Limitation 2: matching ambiguities. The matching problem is well known to be ambiguous. The problem is illustrated on a simple example in Fig. 2. Suppose an object is placed in front of a uniform background. This is a relatively common situation (object viewed against grass, sand, blue sky...). There are many points located in the vicinity of the true surface which will produce high matching scores although these points are not part of the scene. Additional information is necessary to disambiguate the problem.

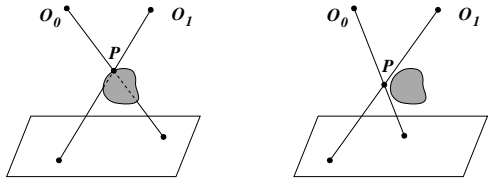


Figure 2. Example of matching ambiguity. With a uniform background, P produces consistent colours in both cases, although the second case (right) corresponds to an incorrect depth assumption.

3.1.4. Novel model incorporating background information. The key idea to address the first limitation is to incorporate opacities in the formulation so as to express background visibility. In this new formulation, background is no longer treated as a conventional 3D layer with standard depth assignments, but is modelled by a set of images. As such, the colour of background points should be consistent, for each camera, with the colour predicted by the background images. Note that the background images can be estimated from sequences of images containing the full scene. The appearance of foreground points, on the other hand, should be consistent with the foreground colour $\mathcal{F}^0_{\boldsymbol{p}}$ seen by the reference camera. $\mathcal{F}^0_{\boldsymbol{p}}$ is related to $\mathcal{C}^0_{\boldsymbol{p}}$, $\mathcal{B}^0_{\boldsymbol{p}}$ and $\alpha^0_{\boldsymbol{p}}$ by the compositing equation [3, 18, 7]

$$\mathcal{C}^0_{\boldsymbol{p}} = \alpha^0_{\boldsymbol{p}}\mathcal{F}^0_{\boldsymbol{p}} + (1-\alpha^0_{\boldsymbol{p}})\mathcal{B}^0_{\boldsymbol{p}}, \tag{5}$$

or equivalently can be expressed as

$$\mathcal{F}^0_{\boldsymbol{p}} = \begin{cases} \frac{1}{\alpha^0_{\boldsymbol{p}}}\mathcal{C}^0_{\boldsymbol{p}} + (1-\frac{1}{\alpha^0_{\boldsymbol{p}}})\mathcal{B}^0_{\boldsymbol{p}} & \text{if} \quad \alpha^0_{\boldsymbol{p}} \neq 0, \\ \mathcal{B}^0_{\boldsymbol{p}} & \text{if} \quad \alpha^0_{\boldsymbol{p}} = 0. \end{cases} \tag{6}$$

We can thus define a foreground image $\mathcal{F}^0$ seen by the reference camera. In our general formulation which allows non-binary alpha values, for mixed pixels, the contribution of the two models is weighed according to the alpha values. Mathematically, this is modelled as follows for a camera i in which the point $\boldsymbol{P} = [\boldsymbol{p}, d_{\boldsymbol{p}}]$ is visible:

$$\mathcal{C}^i_{M^i\boldsymbol{P}} = \alpha^i_{M^i\boldsymbol{P}}\mathcal{F}^0_{M^0\boldsymbol{P}} + (1-\alpha^i_{M^i\boldsymbol{P}})\mathcal{B}^i_{M^i\boldsymbol{P}} + \boldsymbol{\eta}^i_2, \tag{7}$$

where $\{\boldsymbol{\eta}^i_2\}$ represent the image noise. Coming back to the example shown in Fig. 1, the background point $\boldsymbol{P}$ is no longer ambiguous, although occluded in the second camera, because a score can be computed by comparing composite and foreground colours in the reference camera.

The solution to the second limitation is based on the assumption that foreground points must be dissimilar to the background from at least one view. The measure of the likelihood is therefore penalised according to the similarity between background and composite colour such that

$$L(\{\mathcal{C}^i, \mathcal{B}^i\}|d, \{\alpha^i\}) = \sum_{\boldsymbol{p}\in\mathcal{P}} E_2(\{\mathcal{C}^i, \mathcal{B}^i, \alpha^i\}, [\boldsymbol{p}, d_{\boldsymbol{p}}]), \tag{8}$$

$$\text{with} \quad E_2(\{\mathcal{C}^i, \mathcal{B}^i, \alpha^i\}, \boldsymbol{P}) = -\frac{1}{|\mathcal{V}(\boldsymbol{P})|} \tag{9}$$

$$\sum_{i\in\mathcal{V}(\boldsymbol{P})} \left[T_{k_l}(\|\mathcal{C}^i_{M^i\boldsymbol{P}} - \mathcal{B}^i_{M^i\boldsymbol{P}}\| < t_l \wedge \alpha^i_{M^i\boldsymbol{P}} > \alpha_l)\right.$$
$$\left. \|\mathcal{C}^i_{M^i\boldsymbol{P}} - \alpha^i_{M^i\boldsymbol{P}}\mathcal{F}^0_{M^0\boldsymbol{P}} - (1-\alpha^i_{M^i\boldsymbol{P}})\mathcal{B}^i_{M^i\boldsymbol{P}}\|^2\right],$$

$$\text{and} \quad T_k(b) = \begin{cases} k & \text{if} \quad b = \text{true}, \\ 1 & \text{if} \quad b = \text{false}. \end{cases} \tag{10}$$

$\wedge$ represents the AND logical operator. The term $T_{k_l}(\|\mathcal{C}^i_{M^i\boldsymbol{P}} - \mathcal{B}^i_{M^i\boldsymbol{P}}\| < t_l \wedge \alpha^i_{M^i\boldsymbol{P}} > \alpha_l)$ is a penalty term. t_l is a threshold measuring the similarity of a colour with the background. The function penalises errors by a multiplicative factor k_l when a pixel with a high chance of being foreground ($\alpha^i_{M^i\boldsymbol{P}} > \alpha_l$) is similar to the background ($\|\mathcal{C}^i_{M^i\boldsymbol{P}} - \mathcal{B}^i_{M^i\boldsymbol{P}}\| < t_l$). The similarity is expressed here in terms of a threshold, however more sophisticated classification methods could be considered. In practice, the thresholding method was sufficient to obtain accurate results in our application. Coming back to Fig. 2(right), the ambiguity has been removed because although $\boldsymbol{P}$ is still consistent, this assignment has been penalised and is now less likely than the background hypothesis along the same ray.

3.2. Priors on shape

We write $L(d)$ as a sum of the terms $L_1(d)$ and $L_2(d)$ expressing different priors on the scene geometry.

3.2.1. Visual hull prior. In many situations, it is possible to compute approximate silhouettes of the object to reconstruct. Given a set of image silhouettes, the visual hull provides in principle a volume which is guaranteed to contain the surface to be reconstructed. However, if the silhouettes are inaccurate or calibration is inexact, the resulting volume will be a truncated reconstruction of the scene and the previous assertion will not hold. A solution in this case is to compute a conservative estimate of the visual hull. A point is identified as part of the conservative visual hull if its projection in the images are located within a distance r from the silhouettes. We denote by $\mathcal{H}$ such a conservative visual hull estimate; for r appropriately chosen, $\mathcal{H}$ gives an upper bound on foreground scene geometry in spite of initial segmentation or calibration errors. Regarding the background, no depth estimate is required. Such points are represented by a layer which is not assigned any particular position in space and is identified by the depth label ∞. The prior on depth is modelled as follows by assuming a uniform distribution within the visual hull:

$$L_1(d) = \textstyle\sum_{\boldsymbol{p}\in\mathcal{P}} F(d_{\boldsymbol{p}} \neq \infty \wedge [\boldsymbol{p}, d_{\boldsymbol{p}}] \notin \mathcal{H}), \tag{11}$$

$$\text{with} \quad F(b) = \begin{cases} -\infty & \text{if} \quad b = \text{true}, \\ 0 & \text{if} \quad b = \text{false}. \end{cases} \tag{12}$$

3.2.2. Smoothness prior. A smoothness constraint is enforced between pairs of neighbouring image points $\boldsymbol{p}$ and $\boldsymbol{q}$ by defining:

$$L_2(d) = -\sum_{\{\boldsymbol{p},\boldsymbol{q}\}\in\mathcal{N}} k_{s_1} D_d(d_{\boldsymbol{p}}, d_{\boldsymbol{q}}), \tag{13}$$

where D_d penalises depth assignments for neighbours defined by a four-connected neighbourhood $\mathcal{N}$ according to their relative values. Using simple differencing between depth values may be problematic as it penalises large jumps at discontinuities. To eliminate this problem, research has

focused on using discontinuity preserving measures such as the Potts model or the truncated linear distance [2]. The problem with considering discontinuity preserving functions is that it increases significantly the complexity of the algorithm and makes it necessary to compromise by computing only a local solution using for example the α-expansion algorithm proposed in [2]. In our approach we use a trade-off between these two types of measures which consists in measuring the linear distance through the visual hull only, *i.e.* points located outside of the visual hull do not contribute. We note this distance $D_{\text{VH}}(d_{\boldsymbol{p}}, d_{\boldsymbol{q}})$. This distance does not overpenalise jumps between components of the scene which are located far apart, while it is still allowing computation of a global optimum.

Similarly to previous work [2, 6], we incorporate context information to encourage depth discontinuities at regions of high intensity gradient. We use the Deriche filter [5] to extract a set of edges $\mathcal{E}$ from the reference image and weigh the distance accordingly. This introduces robustness to noise as edges are computed over a smoothed area rather than pixel differences as in [2, 6]. This is written as:

$$D_d(d_{\boldsymbol{p}}, d_{\boldsymbol{q}}) = T_{k_d}(\boldsymbol{p} \in \mathcal{E} \ \vee \ \boldsymbol{q} \in \mathcal{E}) D_{\text{VH}}(d_{\boldsymbol{p}}, d_{\boldsymbol{q}}), \quad (14)$$

with $\vee$ denoting the OR logical operator and T_{k_d} already defined in Eq. (10) with $k_d < 1$.

3.3. Priors on alpha mattes

As in the previous section, we express $L(\{\alpha^i\})$ as a sum of different priors on the alpha mattes.

3.3.1. Trimap prior. We assume that a trimap is available. A trimap is a partition of the input image in three sub-sets $\{\mathcal{P}_{\text{BG}}, \mathcal{P}_{\text{FG}}, \mathcal{P}_{\text{X}}\}$ which defines respectively background, foreground, and ambiguous regions. In practice, this is easily obtained from the initial approximate image segmentation.

$$L_1(\{\alpha^i\}) = \sum_i \sum_{\boldsymbol{p} \in \mathcal{P}} F\left[(\boldsymbol{p} \in \mathcal{P}_{\text{BG}} \ \wedge \ \alpha^1_{\boldsymbol{p}} = 1) \vee (\boldsymbol{p} \in \mathcal{P}_{\text{FG}} \ \wedge \ \alpha^1_{\boldsymbol{p}} = 0)\right] \quad (15)$$

3.3.2. Smoothness prior. Similarly to the case of depth values, we define a smoothness prior to encourage segmentation of connected regions by defining (for $k_\alpha < 1$):

$$L_2(\{\alpha^i\}) = -\sum_i \sum_{\{\boldsymbol{p},\boldsymbol{q}\} \in \mathcal{N}} k_{s_2} D_\alpha(\alpha^i_{\boldsymbol{p}}, \alpha^i_{\boldsymbol{q}}), \quad (16)$$

$$\text{with} \quad D_\alpha(\alpha^i_{\boldsymbol{p}}, \alpha^i_{\boldsymbol{q}}) = T_{k_\alpha}(\boldsymbol{p} \in \mathcal{E} \ \vee \ \boldsymbol{q} \in \mathcal{E}) |\alpha^i_{\boldsymbol{p}} - \alpha^i_{\boldsymbol{q}}|. \quad (17)$$

3.3.3. Consistency with shape. We assume that a background point (represented by an infinite depth) must have a zero opacity, while a foreground point (represented by a finite depth) must have a non-zero opacity. This is enforced by adding the following term:

$$L(d, \{\alpha^i\}) = \sum_{\boldsymbol{p} \in \mathcal{P}} \sum_i F\Big[(\alpha^i_{M^i[\boldsymbol{p}, d_{\boldsymbol{p}}]} = 0 \wedge d_{\boldsymbol{p}} = \infty) \vee (\alpha^i_{M^i[\boldsymbol{p}, d_{\boldsymbol{p}}]} \neq 0 \wedge d_{\boldsymbol{p}} \neq \infty)\Big]. \quad (18)$$

4. Implementation

Our formulation was proposed in the general case of continuous depths and opacities. Unfortunately, there is no simple solution in this case, however we show that a global solution to the problem can be computed very efficiently via a single graph-cut computation under the assumption of binary opacity values and discrete depth values. This is a reasonable assumption for the sports scene considered here, because in the case of opaque objects with smooth geometries, the number of mixed pixels is relatively small.

Our graph construction is similar to the construction proposed in [11], with the difference that: i) nodes are placed only in the occupied volume defined by a conservative visual hull, thus resulting in a sparse graph which does not require a large amount of memory and for which a min-cut can be computed efficiently, ii) our graph incorporates additional nodes representing the background of the scene, in order to enable simultaneous matting and reconstruction. The global solution to our optimisation problem is computed with a single graph-cut using the min-cut/max-flow algorithm [1]. This guarantees optimality of the solution (a global optimum is obtained, contrary to [6] which computes a local optimum) and a time efficient implementation which does not require multiple graph-cut computation.

We build a directed capacitated graph as illustrated in Fig. 3. The general structure of the graph is dictated by the geometry of the reference camera considered. Rays are backprojected from each image pixel and sampled with a fixed depth increment Δd. Two types of nodes (represented by white-filled circles in Fig. 3) can be distinguished: foreground nodes and background nodes. The foreground nodes are located at the grid points inside the visual hull, while the background points are placed at the end of the rays. In fact, background nodes can be placed at any location after the foreground nodes since no assumption is made about their depth. In our implementation they are placed in an arbitrary plane located behind the visual hull. Trimap information is incorporated in the graph by removing background nodes on rays where the trimap indicates it should be foreground, and *vice versa* in the case of rays seeing background points. The first node and the last node along each ray are connected respectively to the source $\boldsymbol{s}$ and the sink $\boldsymbol{t}$ of the graph by an edge with infinite capacity. Such a construction guarantees that the visual hull prior and the trimap prior defined in the previous setion are enforced.

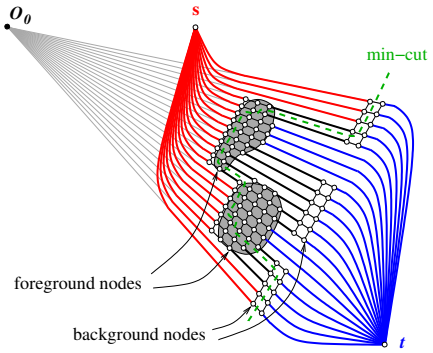


Figure 3. Graph structure.

Edges located along rays are assigned costs corresponding to the likelihood terms $E_2(\{\mathcal{C}^i, \mathcal{B}^i, \alpha^i\}, [\boldsymbol{p}, d_{\boldsymbol{p}}])$ defined in Eq. (9), with $\alpha^i = 1$ for foreground nodes and $\alpha^i = 0$ for background nodes. The corresponding cost for a 3D point $[\boldsymbol{p}, d_{\boldsymbol{p}}]$ is placed between the node at this location and the previous node $[\boldsymbol{p}, d_{\boldsymbol{p}} - \Delta d]$. Note that this requires the introduction of auxiliary nodes for points located on the camera side of the visual hull boundary. To increase robustness in the case of noisy images, the matching score is computed over a 5×5 window using the SSD instead of single pixel differences. The remaining edges located along rays are introduced to ensure continuity of the path between source and sink along each ray; they are assigned an infinite capacity (represented by a thick edge in Fig. 3) so that a cut through these locations is not possible. Edges located across rays are assigned costs corresponding to the smoothness priors. Such an edge connecting two foreground nodes $[\boldsymbol{p}, d_{\boldsymbol{p}}]$ and $[\boldsymbol{q}, d_{\boldsymbol{q}}]$ is assigned a cost with value $k_{s_1} D_d(d_{\boldsymbol{p}} d_{\boldsymbol{q}})$ defined in Eq. (14). The Deriche filter [5] was used to locate edges in the reference observed image. Edge connecting background nodes form the interface between background and foreground layers, where discontinuity in α may occur, and are used to impose the smoothness cost $k_{s_2} D_\alpha(\alpha^i_{\boldsymbol{p}}, \alpha^i_{\boldsymbol{q}})$ for the opacity, as defined in Eq. (17).

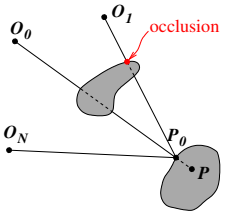


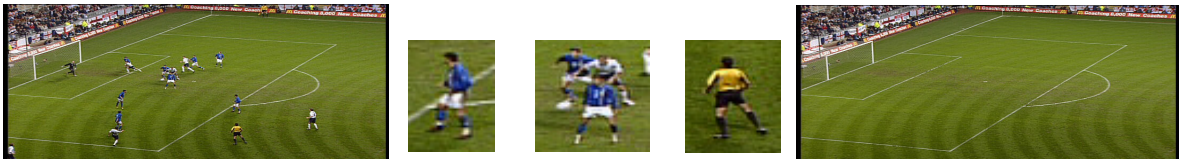
Figure 4. Visibility computation.

For accurate computation of the likelihood costs, only the cameras in which the hypothesised point is visible must be taken into account. We define a simple visibility test based on the conservative visual hull. For a point $\boldsymbol{P}$, we compute the closest visual hull intersection $\boldsymbol{P}_0$ of the reference camera ray passing through it (see Fig. 4). Visibility of $\boldsymbol{P}$ is then assessed as the visibility of $\boldsymbol{P}_0$. The reference camera is always included in the set of cameras regardless of whether the point $\boldsymbol{P}$ is visible or not, as we seek to extract the visible surface that generates the observed image in the reference camera. In the case of a background node, visibility cannot be assessed simply as no depth values are available for these points, however, considering only the reference camera proved to be sufficient in practice. Visibility can be pre-computed for each node in the graph.

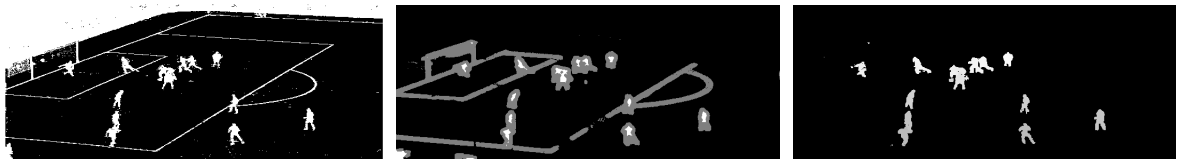
5. Results

Results are shown on images of a football match. 15 static cameras are arranged on one side of stadium, resulting in a total baseline of approximately 90° and a coverage of half the football pitch. Each camera was calibrated by using the lines present on the football pitch as described in [17]. The resolution of the images is 720 × 288 (see Fig. 5(a) for an example). No background images were available, however an estimate of the background was computed by median filtering for each camera view. An example background image is shown in Fig. 5(c). Solving the matting and reconstruction problem is very challenging in this scenario because: i) the image resolution is low (players arms and legs are typically less that 5 pixels wide), ii) image quality is poor due to compression artefacts, iii) calibration errors (an image error of one pixel corresponds to a 3D error of approximately 5 cm), iv) uncontrolled lighting conditions. Fig. 5(b) shows a magnification of the image shown in Fig. 5(a). Note the similarity of the background with foreground in certain areas: skin colour is close to grass colour, while the lines are the same colour as the shorts. Also note the multiple occlusions that occur in such a scene. An example of foreground/background segmentation obtained by chroma-keying is shown in Fig. 5(d). This technique suffers from two main limitations. Firstly it cannot eliminate the pitch lines which have a similar colour to the players, secondly it sometimes fails on players skin which is similar to the background, often resulting in arms and legs being disconnected or removed.

A conservative visual hull with a 3-pixel tolerance is obtained by dilating the image silhouettes by 3 pixels prior to silhouette intersection. Trimaps are defined for each camera by marking image points automatically as: i) background (represented in black in Fig. 5(e)) if located outside of the projection of the conservative visual hull, ii) foreground (white in Fig. 5(e)) if located inside the projection of the conventional visual eroded by 2 pixels, iii) unknown otherwise (grey in Fig. 5(e)). Our algorithm is applied to consecutive triplets of images, the reference camera being chosen as the most central, with the following settings: $\Delta d = 5$ cm,



(a) Composite image (b) Magnification of some players in (a) (c) Background image



(d) Silhouette image (chroma-keying) (e) Trimap (f) Depth map

Figure 5. Different images corresponding to a same camera viewpoint.

$k_l = 3$, t_l is such that approximately 99% of the background points in the trimap would fall below the threshold when compared against the composite colours, $k_d = k_\alpha = 0.5$, $k_{s_1} = 0.008$, and $k_{s_2} = 0.002$. The choice of the parameters is not critical. In principle the optimum setting for t_l could be learnt automatically for each image by analysing the distribution of errors computed on background and foreground points defined by the trimap, however this was not considered in this implementation. Fig. 5(f) shows the depth map obtained as well as the background/foreground segmentation. This shows a significant improvement over the input chroma-key segmentation shown in Fig. 5(d). Lines are eliminated and the matting follows more accurately the players contours, up to small ambiguities which correspond to the possible presence of mixed pixels.

Finally we show how the method can be applied to novel view synthesis. The method is applied to each camera, thus producing a depth map for each view, for which a mesh of the foreground surface is built. Each mesh is textured with the original image from the corresponding reference camera. Discontinuity jumps in meshes are handled by splitting triangles connecting vertices located on different foreground layers based on thresholding of the depth change between adjacent pixels. For efficiency, only the two meshes nearest to the virtual view point are rendered, in a far-to-near order. An example of a novel synthesised view is shown in Fig. 6(c). This is a dramatically more accurate and realistic representation of the scene than the ones obtained with either a standard visual hull (Fig. 6(a)) or a conservative visual hull with a 2-pixel tolerance (Fig. 6(b)). Note for example the missing arms and legs which occur with the standard visual hull because of calibration and matting errors. Although the 2-pixel tolerance introduced for the conservative visual hull is able to prevent missing arms and legs, it results in a dilated reconstruction which is inaccurate and appears unrealistic because players are surrounded by background. The method has been applied to a sequence of 100 images, for which a video clip showing the scene from a virtual view-point, with inclusion of an artificial background, has been generated (see Fig. 7).

6. Conclusions and future work

We have proposed a general framework for jointly formulating the multi-view matting and reconstruction problems. An efficient algorithm which provides a global solution to the problem via a single graph-cut computation under the assumption of binary opacity values and discrete depth values has been described. Results on a large-scale outdoor scene have demonstrated the advantages of this approach compared to a conventional two-stage approach using chroma-key and shape-from-silhouette. Future work will focus on generalising our solution to non-binary alpha values. Some potential solutions in this case could use the binary solution proposed here to initialise an iterative algorithm for optimising opacities from a discrete set of labels, or a two stage approach similar to the layered approach in [19] where opacity is refined only at depth discontinuities.

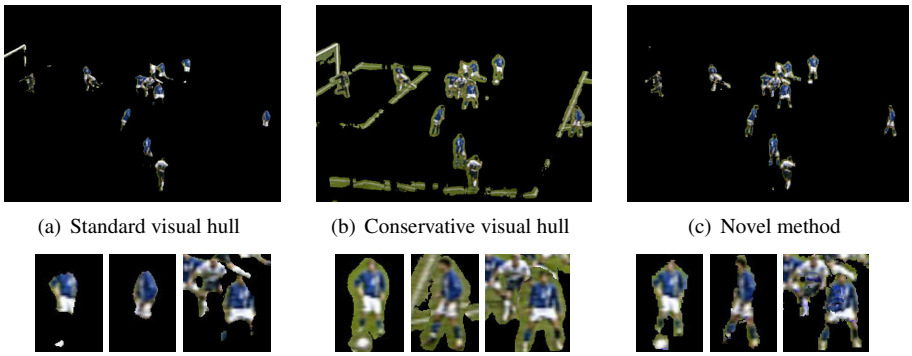
Acknowledgements

This work was supported by the DTI Technology Programme project 'iview: Free-viewpoint Video for Interactive Production' TP/3/DSM/6/I/15515 and EPSRC Grant EP/D033926/1.

References

[1] Y. Boykov and V. Kolmogorov. An experimental comparison of min-cut/max- flow algorithms for energy minimization in vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(9):1124–1137, 2004.

[2] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Transactions on*



(a) Standard visual hull (b) Conservative visual hull (c) Novel method

Figure 6. Novel views synthesised using different methods.

Figure 7. Four frames of a video showing the scene viewed from a novel view-point.

Pattern Analysis and Machine Intelligence, 23(11):1222–1239, 2001.

[3] Y.-Y. Chuang, B. Curless, D. H. Salesin, and R. Szeliski. A bayesian approach to digital matting. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 264–271, 2001.

[4] J. S. De Bonet and P. A. Viola. Roxels: Responsibility weighted 3D volume reconstruction. In *Proc. IEEE International Conference on Computer Vision*, volume 1, pages 418–425, 1999.

[5] R. Deriche. Using Canny's criteria to derive a recursively implemented optimal edge detector. *International Journal of Computer Vision*, 1(2):167–187, 1987.

[6] B. Goldlücke and M. A. Magnor. Joint 3D-reconstruction and background separation in multiple views using graph cuts. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 683–688, 2003.

[7] N. Joshi, W. Matusik, and S. Avidan. Natural video matting using camera arrays. In *Proceedings of ACM SIGGRAPH 2006*, pages 779–786, 2006.

[8] V. Kolmogorov, A. Criminisi, A. Blake, G. Cross, and C. Rother. Probabilistic fusion of stereo with color and contrast for bilayer segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(9):1480–1492, 2006.

[9] K. N. Kutulakos and S. M. Seitz. A theory of shape by space carving. *Int. J. Computer Vision*, 38(3):199–218, July 2000.

[10] A. Laurentini. The visual hull concept for silhouette-based image understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(2):150–162, 1994.

[11] S. Roy and I. J. Cox. A maximum-flow formulation of the N-camera stereo correspondenceproblem. In *Proc. IEEE International Conference on Computer Vision*, pages 492–499, 1998.

[12] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision*, 47(1-3):7–42, 2002.

[13] S. M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 519–528, 2006.

[14] S. M. Seitz and C. R. Dyer. Photorealistic scene reconstruction by voxel coloring. *International Journal of Computer Vision*, 35(2):151–173, 1999.

[15] D. Snow, P. Viola, and R. Zabih. Exact voxel occupancy with graph cuts. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 345–352, 2000.

[16] R. Szeliski and P. Golland. Stereo matching with transparency and matting. *International Journal of Computer Vision*, 32(1):45–61, 1999.

[17] G. A. Thomas. Real-time camera pose estimation for augmenting sports scenes. In *Proc. of 3rd European Conf. on Visual Media Production*, pages 10–19, 2006.

[18] Y. Wexler, A. W. Fitzgibbon, and A. Zisserman. Bayesian estimation of layers from multiple images. In *ECCV '02: Proceedings of the 7th European Conference on Computer Vision-Part III*, pages 487–501, 2002.

[19] C. L. Zitnick, S. B. Kang, M. Uyttendaele, S. Winder, and R. Szeliski. High-quality video view interpolation using a layered representation. In *Proceedings of ACM SIGGRAPH 2004*, pages 600–608, 2004.