Robust Graph-Cut Scene Segmentation and Reconstruction for Free-Viewpoint Video of Complex Dynamic Scenes

Jean-Yves Guillemaut, Joe Kilner and Adrian Hilton
Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, GU2 7XH, UK
{j.guillemaut, j.kilner, a.hilton}@surrey.ac.uk

Abstract

Current state-of-the-art image-based scene reconstruction techniques are capable of generating high-fidelity 3D models when used under controlled capture conditions. However, they are often inadequate when used in more challenging outdoor environments with moving cameras. In this case, algorithms must be able to cope with relatively large calibration and segmentation errors as well as input images separated by a wide-baseline and possibly captured at different resolutions. In this paper, we propose a technique which, under these challenging conditions, is able to efficiently compute a high-quality scene representation via graph-cut optimisation of an energy function combining multiple image cues with strong priors. Robustness is achieved by jointly optimising scene segmentation and multiple view reconstruction in a view-dependent manner with respect to each input camera. Joint optimisation prevents propagation of errors from segmentation to reconstruction as is often the case with sequential approaches. View-dependent processing increases tolerance to errors in on-the-fly calibration compared to global approaches. We evaluate our technique in the case of challenging outdoor sports scenes captured with manually operated broadcast cameras and demonstrate its suitability for high-quality free-viewpoint video.

1. Introduction

In recent years, tremendous progress has been made in the field of image-based scene reconstruction, to such an extent that, under controlled conditions and provided a sufficient number of input images are captured, the performance of such techniques is almost on a par with that of active techniques such as laser range scanners. A good testimony of the capabilities of the current state-of-the-art is provided by the multiple-view Middlebury dataset [22] which shows top performing algorithms capable of sub-millimetre accuracy. The increase in quality can be attributed to improvements in algorithm robustness and also the emergence of more powerful optimisation techniques such as graph-cuts [25], which have enabled the optimisation of otherwise intractable cost functions.

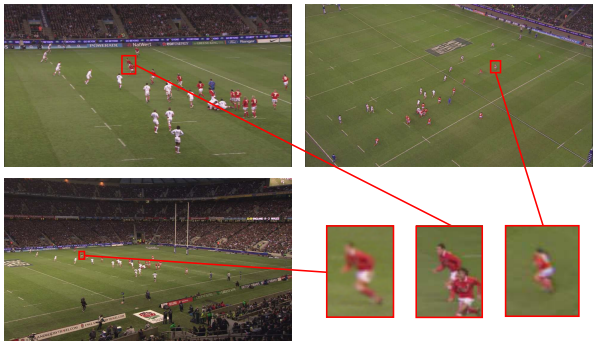


Figure 1. Two moving broadcast camera views (top) and a locked-off camera view (bottom-left) from a rugby match.

Although these algorithms are capable of high-fidelity modelling in controlled studio settings, their performance is often sub-optimal when applied in less controlled conditions of operation, such as those of outdoor scene capture of stadium sports (see Fig. 1). In this case, reconstruction algorithms must be able to cope with a number of factors: calibration errors resulting from on-the-fly calibration of moving and zooming cameras with motion blur; wide-baselines and resolution differences between camera views; non-uniform dynamic backgrounds and multiple people of a relatively small size. Scene segmentation is more difficult because of increased background complexity and the likelihood of overlapping background and foreground distributions, which make standard chroma-keying techniques unusable. In addition, camera placement and framing are manually controlled and not optimised for 3D modelling; cameras are sparsely located with manual operation of pan, tilt and zoom to frame the action.

In this paper, we present a robust approach to reconstruction of scenes captured under these challenging conditions. The approach combines multiple view visual cues such as colour, contrast and photo-consistency information,

with strong shape priors to achieve robust reconstruction for large-scale scenes with sparse wide-baseline moving cameras at different resolutions subject to errors in on-the-fly calibration. A key idea of the technique is to jointly perform segmentation and reconstruction rather than the usual sequential approach where segmentation is followed by reconstruction. We argue that a joint formulation helps disambiguate segmentation of individual views by introduction of information from multiple views and prevents propagation of errors from segmentation leading to failure in reconstruction. View-dependent optimisation with respect to each input camera also increases robustness to errors in camera calibration which prohibit a globally consistent surface reconstruction. A fully automatic system is presented for multiple view calibration, segmentation and reconstruction from moving broadcast cameras to allow high-quality free-viewpoint rendering of sports such as rugby and soccer. Evaluation of simultaneous multiple view segmentation and reconstruction in challenging outdoor scenes demonstrates considerable improvements over single view segmentation and previous multiple view reconstruction approaches. The contribution of this work is a robust framework for reconstruction from multiple moving cameras of complex dynamic scenes which combines visual cues from multiple views to simultaneously segment and reconstruct the scene overcoming limitations of previous approaches using static cameras at a similar resolution which require high-quality calibration, segmentation and matching.

2. Related work

2.1. Robust view-dependent scene reconstruction

Multiple view reconstruction from images has received considerable research interest. A recent survey [22] classifies these algorithms into four categories: (i) volumetric surface extraction, (ii) surface evolution, (iii) depth map-based and (iv) feature point-based surface extraction. Our algorithm belongs to the third category. In their seminal paper [18], Narayanan *et al*. introduced the concept of *Virtualized Reality*. They use a set of 51 cameras distributed on a hemispherical dome to compute 3D models of a dynamic scenes by first computing a depth map for each camera and then merging them into a consistent 3D model. Since then, a number of techniques based on similar two stage pipelines have been proposed. In [7], Goesele *et al*. use a window-based voting approach in order to increase depth map accuracy. The technique produces incomplete reconstructions as ambiguous or occluded areas cannot be reconstructed, but achieves high accuracy in the reconstructed areas. In [5], Campbell *et al*. introduce multiple depth hypotheses at each pixel and an unknown depth label to cope with ambiguities caused by occlusions or lack of texture.

2.2. Joint segmentation and reconstruction

In [23], Snow *et al*. propose a generalisation of shape-from-silhouette reconstruction by integrating the segmentation into the voxel occupancy estimation process. Goldlücke and Magnor [8] proposed a joint graph-cut reconstruction and segmentation algorithm which generalises the work of Kolmogorov and Zabih [13] by adding a background layer to the formulation. Both approaches produce global scene reconstructions and are therefore prone to calibration errors. In [12], Kolmogorov *et al*. proposed a joint segmentation and reconstruction technique which, although structurally similar to ours, targets a different application (background substitution) and is limited to two-layer segmentation from narrow baseline stereo cameras while our approach handles an arbitrary number of layers required for multi-player occlusions in sports. In [9], Guillemaut *et al*. proposed a view-dependent graph-cut approach which can be viewed as a generalisation of Roy and Cox's method [21] with the introduction of a background layer. Finally, in the context of view interpolation, Zitnick *et al*. performed view-dependent reconstruction using colour-segmentation based stereo [27]. Although similar in principle with the proposed technique, these methods assume static cameras separated by a relatively small baseline compared to those considered in this paper. Consequently, they lack robustness with respect to errors in input calibration.

2.3. Sports related research

There is a growing interest in virtual replay production in sports using free-viewpoint video techniques. Transferring studio techniques to the sports arena is a notoriously difficult problem due to the large area, limited access and cost of placing additional cameras. Ideally, solutions will work from the existing broadcast cameras which are manually operated. An initial attempt used in the Eye Vision system at the Super Bowl used camera switching between a large number of slaved cameras to perform transitions. More recent approaches used view-interpolation techniques [20] or planar billboards [19]. These techniques are usually fast, however lack of explicit geometry can result in unrealistic artefacts. Recently some more complex offline optimisation techniques based on graph-cuts [9] or deformable models [11] have been reported. These techniques were able to achieve visual quality comparable to that of the input images, however results were reported with a dedicated set of 15 closely spaced static cameras located around a quarter of the pitch. Current commercial products include the Piero system which is limited to a single camera input and the Liberovision system which is capable of photo-realistic interpolation between match cameras but remains limited to a single frame due to the requirement for manual intervention in calibration and segmentation. In contrast, the proposed

technique is able to a render full sequence from a set of sparsely located input cameras which are not required to be static.

3. View-dependent segmentation and reconstruction

3.1. Problem statement and notation

Given a reference camera and a set of n auxiliary cameras (referred to by the indices 1 to n) all synchronised and calibrated, we would like to (i) partition the reference image into its constituent background/foreground layers and (ii) estimate the depth at each pixel. This can be formulated as a labelling problem where we seek the mappings $l : \mathcal{P} \rightarrow \mathcal{L}$ and $d : \mathcal{P} \rightarrow \mathcal{D}$, which respectively assign a layer label $l_{\boldsymbol{p}}$ and a depth label $d_{\boldsymbol{p}}$ to every pixel $\boldsymbol{p}$ in the reference image. $\mathcal{P}$ denotes the set of pixels in the reference image; $\mathcal{L}$ and $\mathcal{D}$ are discrete sets of labels representing the different layer and depth hypotheses. $\mathcal{L} = \{l_1, \ldots, l_{|\mathcal{L}|}\}$ may consist of one background layer and one foreground layer (classic segmentation problem) or of multiple foreground and background layers. In this paper we assume multiple foreground layers corresponding to players at different depths and a single background layer. More precisely, foreground layers are defined by first computing an approximate background segmentation and then computing a conservative visual hull estimate [11], from which the 3D connected components are extracted to define foreground layers. The set of depth labels $\mathcal{D} = \{d_1, \ldots, d_{|\mathcal{D}|-1}, \mathcal{U}\}$ is formed of depth values d_i obtained by sampling the optical rays together with an unknown label $\mathcal{U}$ used to account for occlusions. Occlusions are common and can be severe when the number of cameras is small (Fig. 1), especially in the background where large areas are often visible only in a single camera.

3.2. Energy formulation

We formulate the problem of computing the optimum labelling (l, d) as an energy minimisation problem of the following cost function

$$E(l,d) = \lambda_{\text{colour}} E_{\text{colour}}(l) + \lambda_{\text{contrast}} E_{\text{contrast}}(l) + \lambda_{\text{match}} E_{\text{match}}(d) + \lambda_{\text{smooth}} E_{\text{smooth}}(l,d). \quad (1)$$

The different energy terms correspond to various cues derived from layer colour models, contrast, photo-consistency and smoothness priors, whose relative contribution is controlled by the parameters λ_{colour}, $\lambda_{\text{contrast}}$, λ_{match} and λ_{smooth}.

3.3. Description of the different energy terms

The colour and contrast terms are frequently used in segmentation problems, while the matching and smoothness terms are normally used in reconstruction. Here we minimise an energy functional which simultaneously involves these two types of terms. Colour and contrast terms are very similar in principle to [24] with the main distinction that we extend the formulation to an arbitrary number of layers.

3.3.1 Colour term

The colour term exploits the fact that different layers (or groups of layers) tend to follow different colour distributions. This encourages assignment of pixels to the layer following the most similar colour model, and is defined as

$$E_{\text{colour}}(l) = \sum_{\boldsymbol{p} \in \mathcal{P}} -\log P(\boldsymbol{I}_{\boldsymbol{p}} | l_{\boldsymbol{p}}), \quad (2)$$

where $P(\boldsymbol{I}_{\boldsymbol{p}} | l_{\boldsymbol{p}} = l_i)$ denotes the probability at pixel $\boldsymbol{p}$ in the reference image of belonging to layer l_i. Similarly to [24], the model for a layer l_i is defined as a linear combination of a global colour model $P_{\text{g}}(\boldsymbol{I}_{\boldsymbol{p}} | l_{\boldsymbol{p}} = l_i)$ and a local colour model $P_{\text{l}}(\boldsymbol{I}_{\boldsymbol{p}} | l_{\boldsymbol{p}} = l_i)$:

$$P(\boldsymbol{I}_{\boldsymbol{p}} | l_{\boldsymbol{p}} = l_i) = w P_{\text{g}}(\boldsymbol{I}_{\boldsymbol{p}} | l_{\boldsymbol{p}} = l_i) + (1-w) P_{\text{l}}(\boldsymbol{I}_{\boldsymbol{p}} | l_{\boldsymbol{p}} = l_i), \quad (3)$$

where w is a real value between 0 and 1 controlling the contributions of the global and local model. A dual colour model combining global and local components allows for dynamic changes in the background. It should be noted that the local model is applicable only to static layers (this is often the case for background layers).

For a given layer l_i, the global component of the colour model is represented by a Gaussian Mixture Model (GMM)

$$P_{\text{g}}(\boldsymbol{I}_{\boldsymbol{p}} | l_{\boldsymbol{p}} = l_i) = \sum_{k=1}^{K_i} w_{ik} N(\boldsymbol{I}_{\boldsymbol{p}} | \mu_{ik}, \Sigma_{ik}), \quad (4)$$

where N is the normal distribution and the parameters w_{ik}, μ_{ik} and Σ_{ik} represent the weight, the mean and the covariance matrix of the k^{th} component for layer l_i. K_i is the number of components of the mixture model for layer l_i.

The local component of the colour model for a static layer l_i is represented by a single Gaussian distribution for each pixel $\boldsymbol{p}$:

$$P_{\text{l}}(\boldsymbol{I}_{\boldsymbol{p}} | l_{\boldsymbol{p}} = l_i) = N(\boldsymbol{I}_{\boldsymbol{p}} | \mu_{i\boldsymbol{p}}, \Sigma_{i\boldsymbol{p}}), \quad (5)$$

where the parameters $\mu_{i\boldsymbol{p}}$ and $\Sigma_{i\boldsymbol{p}}$ represent the mean and the covariance matrix of the Gaussian distribution at pixel $\boldsymbol{p}$. Learning of the colour models is described in Section 4.

3.3.2 Contrast term

The contrast term encourages layer discontinuities to occur at high contrast locations. This naturally encourages low contrast regions to coalesce into layers and favours discontinuities to follow strong edges. This term is defined as

$$E_{\text{contrast}}(l) = \sum_{(\boldsymbol{p},\boldsymbol{q}) \in \mathcal{N}} e_{\text{contrast}}(\boldsymbol{p}, \boldsymbol{q}, l_{\boldsymbol{p}}, l_{\boldsymbol{q}}), \text{ with} \quad (6)$$

$$e_{\text{contrast}}(\boldsymbol{p},\boldsymbol{q},l_{\boldsymbol{p}},l_{\boldsymbol{q}}) = \begin{cases} 0 & \text{if } l_{\boldsymbol{p}} = l_{\boldsymbol{q}}, \\ \exp\left(-\beta C(\boldsymbol{I}_{\boldsymbol{p}},\boldsymbol{I}_{\boldsymbol{q}})\right) & \text{otherwise.} \end{cases} \tag{7}$$

$\mathcal{N}$ denotes the set of interacting pairs of pixels in $\mathcal{P}$ (a 4-connected neighbourhood is assumed) and $||\cdot||$ is the L_2 norm. $C(\cdot,\cdot)$ represents the squared colour distance between neighbouring pixels, and β is a parameter weighting the distance function. Although various distance are possible for $C(\cdot,\cdot)$, we use the attenuated contrast [24]

$$C(I_{\boldsymbol{p}},I_{\boldsymbol{q}}) = \frac{||I_{\boldsymbol{p}}-I_{\boldsymbol{q}}||^2}{1+\left(\frac{||B_{\boldsymbol{p}}-B_{\boldsymbol{q}}||}{K}\right)^2 \exp\left(-\frac{z(\boldsymbol{p},\boldsymbol{q})^2}{\sigma_z}\right)}, \tag{8}$$

where $z(\boldsymbol{p},\boldsymbol{q}) = \max(||I_{\boldsymbol{p}}-B_{\boldsymbol{p}}||, ||I_{\boldsymbol{q}}-B_{\boldsymbol{q}}||)$. $B_{\boldsymbol{p}}$ is the background colour at pixel $\boldsymbol{p}$; it is provided by the local component of the colour model defined in Section 3.3.1. β, K and σ_z are parameters which are set to the standard values suggested in [24]. This formulation uses background information to adaptively normalise the contrast, thereby encouraging layer discontinuities to fall on foreground edges.

3.3.3 Matching term

The matching term is based on the idea that a correct depth estimate must have similar appearances in the images in which the point is visible. Similarity can be measured at each pixel based on photo-consistency or robust feature constraints when available. This term encourages depth assignments to maximise these multi-view similarity measures. Our formulation combines dense (photo-consistency based) and sparse (SIFT feature based) constraints:

$$E_{\text{match}}(d) = E_{\text{dense}}(d) + E_{\text{sparse}}(d). \tag{9}$$

The dense matching score is defined as

$$E_{\text{dense}}(d) = \sum_{\boldsymbol{p}\in\mathcal{P}} e_{\text{dense}}(\boldsymbol{p},d_{\boldsymbol{p}}), \text{ with} \tag{10}$$

$$e_{\text{dense}}(\boldsymbol{p},d_{\boldsymbol{p}}) = \begin{cases} S(\boldsymbol{P}(\boldsymbol{p},d_{\boldsymbol{p}})) & \text{if } d_{\boldsymbol{p}} \neq \mathcal{U}, \\ S_{\mathcal{U}} & \text{if } d_{\boldsymbol{p}} = \mathcal{U}. \end{cases} \tag{11}$$

$\boldsymbol{P}(\boldsymbol{p},d_{\boldsymbol{p}})$ denotes the coordinates of the 3D point along the optical ray passing through pixel $\boldsymbol{p}$ and located at a distance $d_{\boldsymbol{p}}$ from the reference camera. The function $S(\cdot)$ measures the similarity of the reference camera with the auxiliary cameras in which the hypothesised point $\boldsymbol{P}(\boldsymbol{p},d_{\boldsymbol{p}})$ is visible. For weakly textured scenes such as the ones considered in this paper, standard normalised cross correlation similarity measures are inadequate. A more appropriate choice in this case is an error tolerant photo-consistency measure similar to [15]. This computes photo-consistency over extended regions of radius r_{tol} rather than single pixels, and thereby compensates for calibration errors or non-uniform image sampling. The photo-consistency score between the reference image and the auxiliary camera i is defined as

$$\text{photo}_i(\boldsymbol{X}) = \max_{(\boldsymbol{q}-\pi_i(\boldsymbol{X}))^2 < r_{\text{tol}}} \frac{(\boldsymbol{I}_{\boldsymbol{p}}-\boldsymbol{I}^i_{\boldsymbol{q}})^2}{\sigma_i^2}, \tag{12}$$

where σ_i^2 normalises the photo-consistency measure for each auxiliary camera i and the function $\pi_i(\boldsymbol{X})$ projects the hypothesised 3D point $\boldsymbol{X}$ into the image plane of camera i. A robust combination rule is defined as the sum of the k most photo-consistent pairs denoted by $\mathcal{B}_k$

$$S(\boldsymbol{X}) = \sum_{i\in\mathcal{B}_k} \text{photo}_i(\boldsymbol{X}). \tag{13}$$

The sparse matching score is defined as

$$E_{\text{sparse}}(d) = \sum_{\boldsymbol{p}\in\mathcal{P}} e_{\text{sparse}}(\boldsymbol{p},d_{\boldsymbol{p}}), \text{ with} \tag{14}$$

$$e_{\text{sparse}}(\boldsymbol{p},d_{\boldsymbol{p}}) = \begin{cases} 0 & \text{if } \mathcal{F}(\boldsymbol{p}) = \emptyset \text{ or } d_{\boldsymbol{p}} \in \mathcal{F}(\boldsymbol{p}), \\ \infty & \text{otherwise.} \end{cases} \tag{15}$$

$\mathcal{F}(\boldsymbol{p})$ denotes the set of depth labels located within a distance T from a sparse constraints at pixel $\boldsymbol{p}$. This forces the reconstructed surface to pass nearby existing sparse 3D correspondences. Because of calibration errors, we do not require the reconstruction to match exactly the sparse constraints, but allow a tolerance controlled by the parameter T. We use affine-covariant features [17, 16] which are known to be robust to changes in viewpoint and illumination. In this paper, we used the Hessian-affine feature detector. Image features which appear located at the foreground/background junction cannot be used to establish reliable correspondences; we use the learnt colour models defined in Section 3.3.1 to discard such features based on their ratio of foreground and background pixels. Features are represented using the SIFT descriptor and matched based on a nearest neighbour strategy. Robust matching is ensured by restricting the search to areas within a tolerance distance from the epipolar lines. The left-right spatial consistency (reciprocity) constraint is enforced together with temporal consistency which requires corresponding features between camera views to be in correspondence temporally with the previous or the next frame.

3.3.4 Smoothness term

The smoothness term encourages the depth labels to vary smoothly within each layer. This is useful in situations where matching constraints are weak (poor photoconsistency or a low number of sparse constraints) and insufficient to produce an accurate reconstruction without the support from neighbouring pixels. It is defined as

$$E_{\text{smooth}}(l,d) = \sum_{(\boldsymbol{p},\boldsymbol{q})\in\mathcal{N}} e_{\text{smooth}}(l_{\boldsymbol{p}},d_{\boldsymbol{p}},l_{\boldsymbol{q}},d_{\boldsymbol{q}}), \text{ with} \tag{16}$$

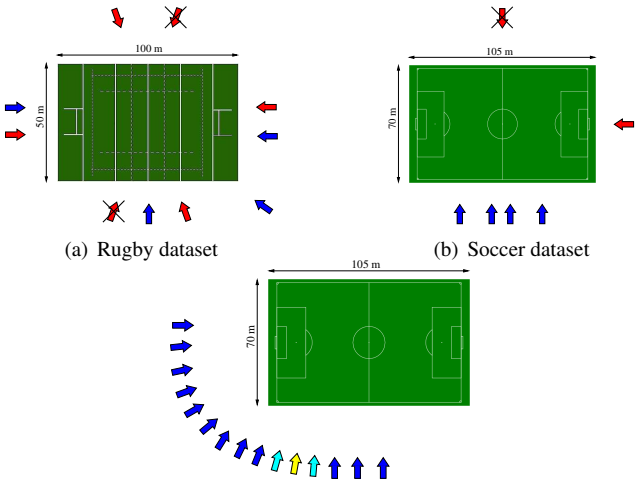
$$e_{\text{smooth}}(l_{\boldsymbol{p}}, d_{\boldsymbol{p}}, l_{\boldsymbol{q}}, d_{\boldsymbol{q}}) = \quad (17)$$
$$\begin{cases} \min(|d_{\boldsymbol{p}} - d_{\boldsymbol{q}}|, d_{\max}) & \text{if } l_{\boldsymbol{p}} = l_{\boldsymbol{q}} \text{ and } d_{\boldsymbol{p}}, d_{\boldsymbol{q}} \neq \mathcal{U}, \\ 0 & \text{if } l_{\boldsymbol{p}} = l_{\boldsymbol{q}} \text{ and } d_{\boldsymbol{p}}, d_{\boldsymbol{q}} = \mathcal{U}, \\ d_{\max} & \text{otherwise.} \end{cases}$$

Discontinuities between layers are assigned a constant smoothness penalty $d_{\max}$, while within each layer the penalty is defined as a truncated linear distance. Such a distance is discontinuity preserving as it does not over-penalise large discontinuities within a layer; this is known to be superior to simpler non-discontinuity functions (see [4, 14]). This term also encourages unknown features to coalesce within each layer.

The choice of shape prior is crucial. A commonly used prior is to assume locally constant depth (fronto-parallel assumption). In this case, a label $(l_{\boldsymbol{p}}, d_{\boldsymbol{p}})$ corresponds to the point from layer $l_{\boldsymbol{p}}$ and located at a distance $d_{\boldsymbol{p}}$ from the reference camera centre along the ray emanating from pixel $\boldsymbol{p}$. Although this yields good quality results when supported by strong matching cues, this results in bias towards flat figure models which do not give good alignment between views. An alternative approach which we use here is to place samples along the iso-surfaces of the visual hull, which results in a reconstructed surface biased towards the visual hull iso-surfaces. We call this prior the iso-surface prior. In this case, a label $(l_{\boldsymbol{p}}, d_{\boldsymbol{p}})$ corresponds to the first point of intersection between the ray emanating from pixel $\boldsymbol{p}$ and the $d_{\boldsymbol{p}}$-th iso-surface in the interior of the visual hull's connected component corresponding to layer $l_{\boldsymbol{p}}$. To account for calibration and matting error, we use the error tolerant visual hull proposed in [11]. Unlike the fronto-parallel prior, the iso-surface prior is view-independent and results in reconstructions more realistic and likely to coincide in the absence of strong matching cues. It can be noted that the choice of a fronto-parallel or an iso-surface prior affects the correspondence between labels and the 3D points they represent, however it does not change the formulation in Eq. (17) since the set of depth values remain an ordered set of discrete values.

3.4. Graph-cut optimisation

Optimisation of the energy defined by Eq. (1) is known to be NP-hard. However, an approximate solution can be computed using the expansion move algorithm based on graph-cuts [4]. Proof of the regularity of the energy function, required for alpha expansion optimisation, is provided in [1]. The expansion move proceeds by cycling through the set of labels $\alpha = (l_\alpha, d_\alpha)$ in $\mathcal{L} \times \mathcal{D}$ and performing an α-expansion iteration for each label until the energy cannot be decreased (see [4]). An α-expansion iteration is a change of labelling such that each pixel $\boldsymbol{p}$ either retains its current value or takes the new label α. Each α-expansion iteration can be solved exactly by performing a single graph-cut using the min-cut/max-flow algorithm [3]. After convergence of the algorithm, the result obtained is guaranteed to be a strong local optimum [4]. The α-expansion algorithm was initialised with the visual hull estimate; convergence has been found to be insensitive to the choice of initialisation. In practice, convergence is usually achieved in 3 or 4 cycles of iterations over the label set. We improve computation and memory efficiency by dynamically reusing the flow at each iteration of the min-cut/max-flow algorithm [2]. This results in a speed-up of an order of two.



(a) Rugby dataset (b) Soccer dataset

(c) Evaluation dataset

Figure 2. Camera layout for the different datasets. Broadcast cameras are represented as blue arrows, locked-off cameras as red arrows. Crossed-out cameras were not used due to inadequate framing. The ground truth camera is indicated in yellow; additional left-out cameras are shown in light blue colour.

4. Results

Testing of the algorithm was performed on two wide-baseline datasets from an international 6-Nations rugby game (400 frames) and a European Championship soccer game (100 frames). In addition, a quantitative evaluation of the free-viewpoint capabilities of the different techniques was performed on a soccer game (100 frames) captured using a special camera rig providing additional ground-truth views. Due to space limitations, examples shown in the paper are restricted to single frames, however results on full sequences can be seen in the attached video. The reconstruction pipeline is the same for all datasets and can be broken into four stages: (i) camera calibration, (ii) initial segmentation using difference keying, (iii) initial scene reconstruction using conservative visual hull, and (iv) joint refinement of segmentation and reconstruction based on the proposed approach.

Calibration is performed on-the-fly from pitch line features [26]. This typically produces calibration errors of

the order of 1-3 pixels which are rather large given player resolution (a pixel roughly corresponds to 5cm on widely framed views) and pitch dimension (100×50 m). Initial segmentation is obtained by computing a background image plate for each input view based on standard mosaicing techniques (this does not require separate background capture) and thresholding the colour difference between the background plate and original image. Initial segmentation is inaccurate due to overlapping foreground and background colour distributions, which is exacerbated by the presence of compression artifacts and motion blur. For this reason, segmentation thresholds are set to low values to prevent foreground erosion. Initial reconstruction is obtained by error tolerant shape-from-silhouette which prevents scene truncation in the presence of calibration and segmentation errors [11].

Joint refinement of segmentation and reconstruction is performed in parallel for each input camera, using the set of neighbouring cameras as auxiliary cameras (usually no more than 1 or 2 auxiliary cameras are usable due to the wide-baseline). We define two background layers corresponding to pitch and crowd area, and as many foreground layers as there are components in the conservative visual hull. The local colour models used for the background is learnt automatically from the background plate and its variance. The global colour model for each layer is learnt using samples from the background plate in the case of the background and from a single manually annotated key-frame for each camera in the case of the foreground; a mixture of 5 components was used for each layer. The parameters weighting the contribution of the different energy terms were set to: $\lambda_{\text{colour}} = 0.5$, $\lambda_{\text{match}} = 0.5$, $\lambda_{\text{contrast}} = 1.0$ and $\lambda_{\text{smooth}} = 0.1$. w balancing local and global colour models was set to 0.5. Isosurfaces were sampled every 5 mm and the maximum penalty d_{max} was set to 20. The same parameter settings were used for all datasets. Refinement run-time varies according to the number of foreground labels in the scene and is of the order of a minute per reference frame on an Intel Core 2 Duo 2.4GHz CPU.

4.1. Rugby and soccer datasets

The rugby dataset consists of six locked-off cameras and four moving broadcast cameras, while the soccer dataset consists of only two locked off cameras and four broadcast cameras (see Fig. 2). All cameras are high-definition 1920x1080 acquired at 25Hz. The broadcast cameras are manually operated zooming and rotating cameras. The locked-off cameras have a wider framing to give coverage of different sections of the pitch and are therefore not all simultaneously usable.

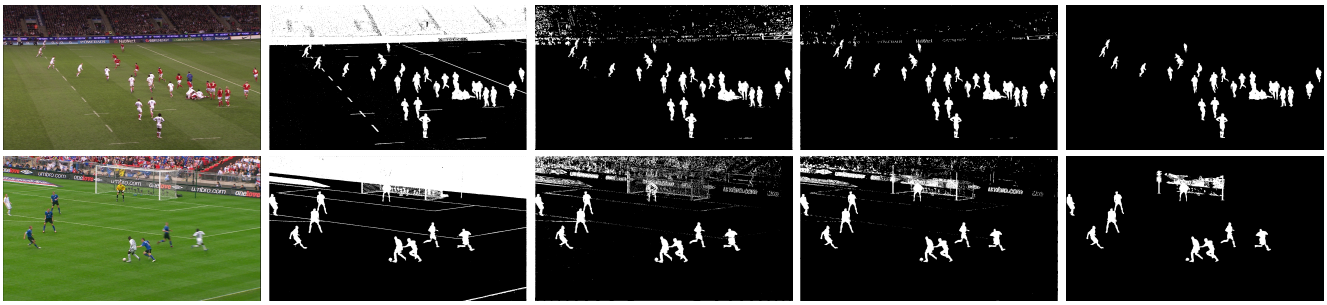
In terms of segmentation, we compared our approach against three other techniques: (i) user assisted chroma-keying, (ii) difference-keying, (iii) background-cut [24]. Results for the four techniques are presented in Fig. 3 and the accompanying video. Clearly a purely global technique such as chroma-keying produces poor results as it is applicable only to grass areas. Difference keying produces better results due to the locality of the colour model it is based on, however it fails in the crowd area which is non-static and in areas where foreground and background colour are similar. Background cut is able to improve further the results by combining local and global models thereby increasing robustness and adding tolerance to non-static background elements, however it yields inaccurate results in ambiguous areas. The proposed approach, combining multiple view information to disambiguate the problem, produces a cleaner segmentation than all other methods.

To evaluate reconstruction, our technique was compared to three standard techniques: (i) conventional visual hull, (ii)conservative visual hull (with 2 pixel tolerance), and (iii) stereo refinement of the conservative visual hull with no colour, contrast or smoothness term. Results are shown in Fig. 4. As expected the visual hull produces large truncations in the presence of calibration and segmentation errors. These truncations have been eliminated in the conservative visual hull but have been replaced by some protrusions and phantom volumes. Stereo refinement of the conservative visual hull results in a very noisy reconstruction; this illustrates the weakness of the available photo-consistency information. In contrast, the proposed technique yields a smooth reconstruction with accurate player boundaries and the elimination of phantom volumes.

The technique can be used to synthesise views from novel viewpoints where real cameras cannot be physically located such as views from inside the pitch. Some examples of free-viewpoint video sequences can be seen in the attached video. Each depth-map is converted to a mesh and then rendered using view-dependent texture mapping techniques [6]. High quality rendering from view-dependent geometry and images is made possible by the use of the proposed surface reconstruction technique. This accurately aligns wide baseline views based on stereo matches and gives a smooth surface approximation based on iso-surfaces of the visual-hull shape prior in regions of uniform appearance which commonly occur on player shirts or due to views sampled at significantly different resolutions. Further results can be found in the submitted video.

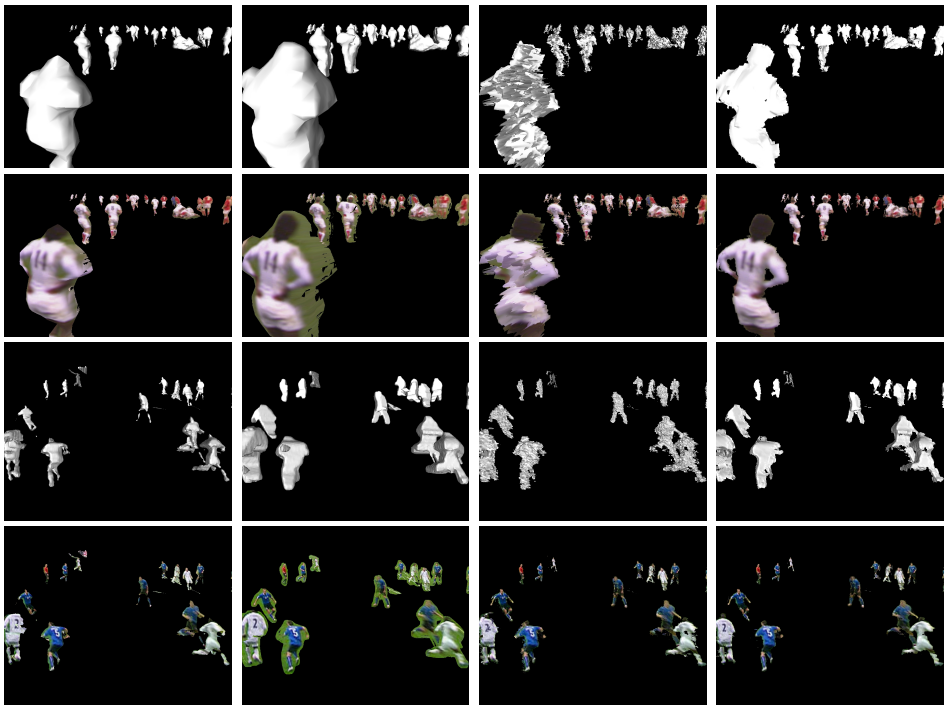
4.2. Evaluation dataset

A quantitative evaluation of free-viewpoint rendering accuracy was performed on a dataset of 15 closely spaced cameras. One camera (shown in yellow in Fig. 2) was excluded from reconstruction and used to provide ground truth. Three sets of reconstruction cameras where defined by excluding one, two or three neighbouring cameras (see Fig. 2), thus defining sets with increasing baseline. All four



Original Chroma-key Difference key Background cut Proposed

Figure 3. Example of segmentation results on rugby (top) and soccer (bottom) data (see attached video for full sequence).



Visual hull Cons. visual hull Stereo refinement Proposed

Figure 4. Example of reconstruction results on rugby (top) and soccer (bottom) data (see attached video for full sequence).

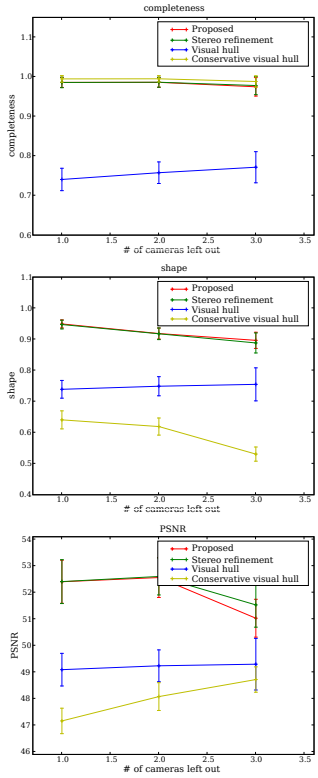


Figure 5. Quantitative evaluation of free-viewpoint video performance.

reconstruction techniques were then used to synthesise the sequences of images from the same viewpoint as the ground truth camera. The quality of the synthesised images was assessed by measuring their similarity with the ground truth images. Evaluation was performed over 100 frames based on the completeness, shape and PSNR scores defined in [10]. Results are shown in Fig. 5. The visual hull produces the worst results. The conservative visual hull produces a complete reconstruction but a poor estimate of shape and appearance. The stereo refinement of the conservative visual hull and the proposed technique both score high on all scores with marginal differences. Although scores are comparably high for these two techniques, the visual quality is usually greater for the proposed technique (see attached video). This is not reflected in the evaluation scores which are not able to capture essential visual features such as temporal consistency and image smoothness. The visual quality improvements resulting from the proposed technique can be best seen in the accompanying video.

5. Conclusions and future work

This paper introduced a novel approach for simultaneous multi-layer segmentation and reconstruction of complex dynamic scenes captured with multiple moving cameras at

different resolutions. View-dependent graph-cut optimisation is proposed combining visual colour and contrast cues, previously used in 2D image segmentation, with multiple view dense photo-metric correspondence and sparse spatio-temporal feature matching, previously used in reconstruction. Strong priors on shape and appearance are combined with multiple view visual cues to achieve accurate segmentation and reconstruction which is robust to relatively large errors (1-3 pixels RMS) in on-the-fly camera calibration. A fully automatic system is presented for multiple view calibration, segmentation and reconstruction from moving broadcast cameras to allow high-quality free-viewpoint rendering of sports such as rugby and soccer. Evaluation of simultaneous multiple view segmentation and reconstruction in challenging outdoor scenes demonstrates considerable improvements over independent single view segmentation and conventional visual-hull or stereo reconstruction. Sports reconstruction presents a challenging problem with a small number (6-12) of independently manually operated panning and zooming broadcast cameras, sparsely located around the stadium to cover a large area with multiple players. This results in multiple view wide-baseline capture at different resolutions with motion blur due to player and camera motion. The framework enables high-quality rendering of novel views from wide-baseline moving camera acquisition of sports, overcoming limitation of previous multiple view reconstruction algorithms.

Acknowledgements This work was supported by TSB/EPSRC projects "i3Dlive: interactive 3D methods for live-action media" (TP/11/CII/6/I/AJ307D, TS/G002800/1) and "iview: Free-viewpoint Video for Interactive Entertainment Production" (TP/3/DSM/6/I/15515, EP/D033926/1).

References

[1] `http://www.guillemaut.org/publications/09/GuillemautICCV09_supplemental.pdf`.

[2] K. Alahari, P. Kohli, and P. Torr. Reduce, reuse & recycle: Efficiently solving multi-label MRFs. In *CVPR*, 2008.

[3] Y. Boykov and V. Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *PAMI*, 26(9):1124–1137, 2004.

[4] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *PAMI*, 23(11):1222–1239, 2001.

[5] N. Campbell, G. Vogiatzis, C. Hernández, and R. Cipolla. Using multiple hypotheses to improve depth-maps for multi-view stereo. In *ECCV*, volume I, pages 766–779, 2008.

[6] P. Debevec, C. Taylor, and J. Malik. Modeling and rendering architecture from photographs: a hybrid geometry- and image-based approach. In *SIGGRAPH*, pages 11–20, 1996.

[7] M. Goesele, B. Curless, and S. Seitz. Multi-view stereo revisited. In *CVPR*, pages 2402–2409, 2006.

[8] B. Goldlücke and M. Magnor. Joint 3D-reconstruction and background separation in multiple views using graph cuts. In *CVPR*, volume 1, pages 683–688, 2003.

[9] J.-Y. Guillemaut, A. Hilton, J. Starck, J. Kilner, and O. Grau. A Bayesian framework for simultaneous matting and 3D reconstruction. In *3DIM*, pages 167–174, 2007.

[10] J. Kilner, J. Starck, J.-Y. Guillemaut, and A. Hilton. Objective quality assessment in free-viewpoint video production. *SPIC*, 24:3–16, 2008.

[11] J. Kilner, J. Starck, A. Hilton, and O. Grau. Dual-mode deformable models for free-viewpoint video of sports events. In *3DIM*, pages 177–184, 2007.

[12] V. Kolmogorov, A. Criminisi, A. Blake, G. Cross, and C. Pother. Probabilistic fusion of stereo with color and contrast for bilayer segmentation. *PAMI*, 28(9):1480–1492, 2006.

[13] V. Kolmogorov and R. Zabih. Multi-camera scene reconstruction via graph cuts. In *ECCV*, volume III, pages 82–96, 2002.

[14] V. Kolmogorov and R. Zabih. What energy function can be minimized via graph cuts? *PAMI*, 26(2):147–159, 2004.

[15] K. Kutulakos. Approximate N-view stereo. In *ECCV*, volume I, pages 67–83, 2000.

[16] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *PAMI*, 27(10):1615–1630, 2005.

[17] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool. A comparison of affine region detectors. *IJCV*, 65(1-2):43–72, 2005.

[18] P. Narayanan, P. Rander, and T. Kanade. Constructing virtual worlds using dense stereo. In *ICCV*, pages 3–10, 1998.

[19] Y. Ohta, I. Kitahara, Y. Kameda, H. Ishikawa, and T. Koyama. Live 3D video in soccer stadium. *IJCV*, 75(1):173–187, 2007.

[20] I. Reid and K. Connor. Multiview segmentation and tracking of dynamic occluding layers. In *BMVC*, volume 2, pages 919–928, 2005.

[21] S. Roy and I. Cox. A maximum-flow formulation of the N-camera stereo correspondence problem. In *ICCV*, pages 492–499, 1998.

[22] S. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *CVPR*, pages 519–528, 2006.

[23] D. Snow, P. Viola, and R. Zabih. Exact voxel occupancy with graph cuts. In *CVPR*, volume 1, pages 345–352, 2000.

[24] J. Sun, W. Zhang, X. Tang, and H.-Y. Shum. Background cut. In *ECCV*, volume 3954, pages 628–641, 2006.

[25] R. Szeliski, R. Zabih, D. Scharstein, O. Veksler, V. Kolmogorov, A. Agarwala, M. Tappen, and C. Rother. A comparative study of energy minimization methods for Markov random fields with smoothness-based priors. *PAMI*, 30(6):1068–1080, 2008.

[26] G. Thomas. Real-time camera tracking using sports pitch markings. *J. Real-Time Image Proc.*, 2:117–132, 2007.

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