Int J Comput Vis manuscript No.
(will be inserted by the editor)

Joint Multi-Layer Segmentation and Reconstruction for Free-Viewpoint Video Applications

Jean-Yves Guillemaut · **Adrian Hilton**

Received: date / Accepted: date

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 environments such as sports scenes with moving cameras. 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 through-the-lens calibration compared to global approaches. We evaluate our technique in the case of challenging outdoor sports scenes captured with manually operated broadcast cameras as well as several indoor scenes with natural background. A comprehensive experimental evaluation including qualitative and quantitative results demonstrates the accuracy of the technique for high quality segmentation and reconstruction and its suitability for free-viewpoint video under these difficult conditions.

Keywords Segmentation · Reconstruction · Free-viewpoint video · Graph-cuts

Centre for Vision, Speech and Signal Processing
University of Surrey, Guildford, GU2 7XH, UK
Tel.: +44 1483 683958 Fax: +44 1483 686031
E-mail: J.Guillemaut@surrey.ac.uk

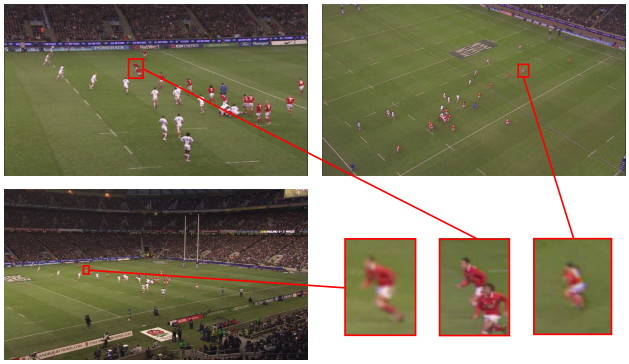


Fig. 1 Two moving broadcast camera views (top) and a locked-off camera view (bottom-left) from a rugby match.

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 (Seitz et al., 2006) 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 (Szeliski et al., 2008), which have enabled the optimisation of otherwise intractable cost functions.

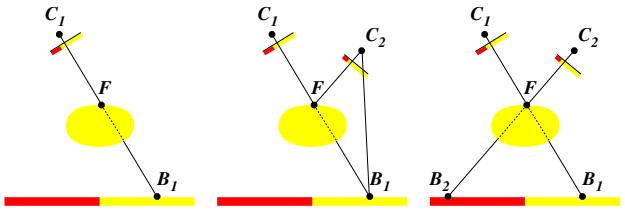
Although these algorithms are capable of high-fidelity modelling in a controlled indoor settings, their performance is often sub-optimal when applied in less controlled conditions of operation, such as those of outdoor scene capture of stadium sports (see Fig. 1) or in the case of indoor capture without use of a chroma-key background. In the

case of sports capture, reconstruction algorithms must be able to cope with a number of factors: calibration errors resulting from through-the-lens calibration of manually operated 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. Scene reconstruction is extremely challenging due to limited visual cues and relatively large errors (1–2 pixels) affecting both calibration and segmentation which prevent extraction of a globally consistent solution. It is often not possible to improve calibration accuracy as through-the-lens calibration techniques required to calibrate moving cameras are reaching their limit due to limited scene information (limited number of features, motion blur).

Indoor capture with a natural background is in theory simpler due to better framing and absence of large calibration errors, however accurate automatic segmentation of multiple cameras streams remains a challenging problem due to possible overlap in foreground and background distributions. While interactive video segmentation techniques have been proposed in recent years (Chuang et al., 2002; Li et al., 2005; Bai et al., 2009), these remain essentially limited to a single video stream and require frequent interaction (every 10–20 frames) in the form of trimaps or brush strokes to eliminate segmentation errors; these requirements make standard monocular segmentation techniques prohibitively expensive to use given the large volume of data generated during multi-view capture. In addition, none of the techniques exploit the redundancy contained in images simultaneously captured from different viewpoints to facilitate the segmentation process or improve reconstruction accuracy.

In this paper, we present a robust approach to reconstruction of scenes captured under these challenging conditions. This approach capitalises on two key ideas: (1) joint optimisation of segmentation and reconstruction can be used to disambiguate and improve each process compared to a sequential approach; (2) view-dependent reconstruction can be used to increase robustness to segmentation and calibration errors compared to global reconstruction approaches.

Many multi-view reconstruction approaches rely on a two-stage sequential pipeline where foreground/background segmentation is initially performed independently with respect to each camera, and then used as input to multi-view reconstruction. The problem with this approach is two-fold. First it requires making hard decisions at the segmentation stage, which may not be recoverable at the reconstruction stage and will deteriorate final reconstruction quality. Second, a sequential approach is sub-optimal because it does not use the multi-view information at the segmentation stage.



(a) Segmentation (b) Reconstruction (c) Joint segmentation and reconstruction

Fig. 2 Illustration of the advantage of a joint multi-view approach compared to separate segmentation and reconstruction in the case of a simple 2D example. Single view segmentation with respect to camera $\mathbf{C}_1$ is ambiguous as foreground colour $\mathbf{F}$ and background colour $\mathbf{B}_1$ are identical (a). Multi-view reconstruction is also ambiguous as the foreground hypothesis $\mathbf{F}$ and the background hypothesis $\mathbf{B}_1$ are equally photo-consistent due to lack of texture (b). Joint multi-view segmentation and reconstruction allows unambiguous selection of the foreground hypothesis $\mathbf{F}$ since the background hypothesis $\mathbf{B}_2$ is not plausible with respect to camera $\mathbf{C}_2$; this in turn allows rejection of the background hypothesis $\mathbf{B}_1$ with respect to camera $\mathbf{C}_1$ (c).

To illustrate this shortcoming, let us consider a foreground object with similar colour to the background when seen from a given viewpoint (see Fig. 2). Single view segmentation is unable to correctly classify the pixel as foreground or background due to the colour ambiguity (Fig. 2(a)). Depth estimation at this pixel is also ambiguous as both foreground and background hypotheses may be equally photo-consistent due to lack of texture (Fig. 2(b)). Introducing an additional view at the segmentation stage is however able to correctly classify this pixel as foreground and reject the incorrect hypothesis (Fig. 2(c)). The joint formulation proposed in this paper 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.

Let us illustrate the challenge in computing a globally consistent reconstruction in the presence of errors in the case of the shape-from-silhouette algorithm which computes a reconstruction by intersecting the set of visual cones defined by the foreground silhouettes (Laurentini, 1994). In the absence of segmentation and calibration errors, the shape obtained is optimal in the sense it corresponds to the maximal volume consistent with the silhouette constraints. In the presence of calibration or segmentation errors, there is no longer a guarantee that the actual scene surface is contained within the visual hull. In the case of sports data, this can produce severe truncations of players (see Fig. 17, first column). A naïve solution to compensate for these errors would be to build a conservative visual hull by introducing a tolerance on the intersection of the visual cones. Unfortunately, although this may be sufficient to guarantee that the real scene is contained in the visual hull in spite of segmentation or camera calibration errors, this would also produce a dilated

representation of the scene, which is inaccurate (see Fig. 17, second column).

The key contribution of this work is the introduction of 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 in a view-dependent manner. This overcomes limitations of previous approaches using static cameras at a similar resolution to each other and which require high-quality calibration, segmentation and matching. To achieve robust wide-baseline reconstruction our approach integrates multiple cues such as sparse affine covariant features and error tolerant photo-consistency scores into a novel energy formulation. To overcome errors in online calibration of moving cameras, which prohibit a globally consistent reconstruction, a view dependent reconstruction is proposed. This combination of view-dependent reconstruction with robust cues was essential to achieve the quality of results reported in this paper.

The paper is structured as follows. We start by reviewing related work, focusing on robust multi-view reconstruction, joint segmentation and reconstruction, and sports related research. After formulating the problem and giving an overview of the proposed approach, we give a detailed description of the joint segmentation and reconstruction technique introducing the energy minimisation framework adopted in this paper. The next section generalises the framework by adding a new energy term to iteratively enforce multi-view consistency in reconstruction between nearby cameras as well as additional shape priors. Then we address the efficient optimisation of the energy function using graph-cuts and 3D model generation from multiple view-dependent reconstructions. Finally a thorough evaluation including qualitative and quantitative results with outdoor and indoor datasets with natural background is presented before concluding.

2 Related work

2.1 Robust multi-view scene reconstruction

Image-based rendering techniques are often represented along an axis known as the image-geometry continuum which measures the amount of geometric modelling required (Kang et al., 2006). In this representation, Kang et al. identify three main categories of algorithms depending on whether they use no geometry, implicit geometry or explicit geometry. Techniques using no geometry such as the light field (Levoy and Hanrahan, 1996) or the lumigraph (Gortler et al., 1996) require a large number of images to approximate the continuous plenoptic function representing appearance variation with viewpoint and direction. These approaches are often limited to static scenes due to the difficulty of the acquisition process. Techniques using implicit geometry such as view interpolation (Chen and Williams, 1993) or image morphing (Seitz and Dyer, 1996) require dense correspondences between views to synthesise novel viewpoints. While less demanding in terms of number of views, they require a narrow-baseline and can produce artefacts due to lack of occlusion modelling. Techniques using explicit geometry (Seitz et al., 2006) are able to accurately model a scene from a small number of camera views, however they require powerful computer vision algorithms to reconstruct a reliable 3D geometric proxy. These techniques are the most suitable for the problem of free-viewpoint video rendering from wide-baseline camera configurations considered in this paper. In the remainder of this section, we will focus on dense multiple view reconstruction techniques which compute an explicit geometry model with an emphasis on robust techniques. The reader is referred to (Kang et al., 2006) for a review of techniques using no geometry or implicit geometry.

Shape-from-silhouette These techniques compute an approximation of the scene geometry known as the visual hull (Laurentini, 1994) by intersecting the set of visual cones defined by backprojecting foreground image silhouettes into the 3D space. This approach does not impose any restrictions on scene surface reflectance, texture or baseline separating cameras and as a result offers a considerable advantage in terms of robustness compared to techniques based on photo-consistency or stereo matching. A common approach to visual hull computation consists in discretising the 3D space into sub-elements called voxels and classify them into empty or occupied categories to produce a volumetric scene representation (Moezzi et al., 1997). More recent work has focused on improving the efficiency and exactness of the visual hull computation by performing the computation in the image domain (Matusik et al., 2000) or by estimating an exact polyhedral representation (Franco and Boyer, 2003). Visual hull algorithms suffer from two main limitations. Firstly, they require accurate foreground segmentation and perfect camera calibration; an exception is the probabilistic framework proposed in (Franco and Boyer, 2005). Secondly, visual-hull only gives a maximum bound on scene geometry, failing to reconstruct concavities.

Shape-from-photoconsistency These algorithms compute a more accurate approximation of the scene known as the photo-hull (Seitz and Dyer, 1999; Kutulakos and Seitz, 2000). The photo-hull is defined as the maximum volume which is photo-consistent with the set of input views. Algorithms discretise the 3D space into voxels and use an ordering constraint to iteratively carve empty voxels based on a photo-consistency score measuring colour similarity across the image projections (Seitz and Dyer, 1999; Kutulakos and Seitz, 2000). Several extensions have been proposed to add robustness to calibration errors (Kutulakos, 2000) and image noise (Broadhurst et al., 2001). A review of volumetric reconstruction techniques can be found in (Slabaugh et al.,

2001). These techniques do not require scene segmentation, however they assume sufficient texture and a diffuse surface reflectance model in order to correctly identify surface voxels; additionally they are usually very sensitive to calibration errors and often produce noisy reconstructions due to the absence of regularisation.

Global multi-view stereo reconstruction These methods seek a globally optimal reconstruction with respect to the input images. They usually combine photo-consistency information with additional cues such as silhouette constraints (Hernández Esteban and Schmitt, 2004; Sinha and Pollefeys, 2005; Starck and Hilton, 2007), sparse feature correspondences (Starck and Hilton, 2007) and shape priors to disambiguate the problem and guarantee a smooth solution. Recent advances in the field of discrete optimisation have enabled resolution of increasingly complex optimisation problems. In particular, graph-cut (Boykov et al., 2001; Kolmogorov and Zabih, 2004; Boykov and Kolmogorov, 2004) has become increasingly popular due to the strong optimality properties and has been utilised to solve a wide variety of multi-view reconstruction problems (Sinha and Pollefeys, 2005; Starck and Hilton, 2007; Vogiatzis et al., 2007). Recently surface growing approaches have been used to increase robustness of the reconstruction process (Habbecke and Kobbelt, 2007; Furukawa and Ponce, 2010). In this framework, reconstruction starts from a set of seed points with strong confidence and then a surface growing approach is employed to reliably propagate the reconstruction to neighbouring areas. The reader is referred to (Seitz et al., 2006) for a more detailed survey of multi-view stereo reconstruction algorithms. These algorithms are able to produce very high quality results even in the presence of weak texture information, however they normally require perfect calibration information to guarantee existence of a globally consistent solution. In addition, when silhouette constraints are enforced, they rely on accurately segmented foreground.

View-dependent multi-view stereo reconstruction Instead of seeking a globally consistent solution, these techniques compute a separate reconstruction for each input camera and then merge the results into a single representation. This concept was pioneered by Narayanan et al. in the Virtualized Reality^{TM} system (Narayanan et al., 1998) which used 51 cameras distributed on a 5 m hemispherical dome to reconstruct dynamic scenes and allow free-viewpoint video from an arbitrary viewpoint. Since then, many techniques based on similar two stage pipelines have been proposed. Starck and Hilton combined shape-from-silhouette with stereo matching cues to increase the accuracy of depth map extraction compared to techniques based on a single cue (Starck and Hilton, 2005). Goesele et al. proposed a window-based voting approach able to recover partial but highly accurate depth maps (Goesele et al., 2006) Bradley et al. (Bradley et al., 2008) used scaled window matching to reduce inaccuracies caused by projective distortion, combined with an adaptive depth map filtering technique to remove outliers and suppress noise. Campbell et al. introduced multiple depth hypotheses at each pixel and an unknown depth label to cope with ambiguities caused by occlusions or lack of texture (Campbell et al., 2008). Liu et al. proposed an extension to the continuous domain to eliminate quantisation errors (Liu et al., 2009). These techniques assume accurate camera calibration to extract an accurate depth map.

Error tolerant view-dependent rendering Approaches have been proposed to improve free-viewpoint video quality in presence of reconstruction and/or calibration errors. Eisemann et al. used optical flow to reduce blending artefacts due to inaccurate reconstruction or calibration (Eisemann et al., 2008). Enhanced billboard representations have also been used to generate high quality free-viewpoint video transitions (Waschbüsch et al., 2007; Germann et al., 2010; Ballan et al., 2010). Waschbüsch et al. (Waschbüsch et al., 2007) introduced the concept of 3D video billboard combining a planar geometric proxy, suited for rendering but lacking accuracy, with view-dependent displacement maps allowing accurate blending between views; they use 4D bilateral filtering to enforce spatio-temporal consistency of displacement maps. Germann et al. (Germann et al., 2010) proposed a semi-automatic algorithm to fit an articulated billboard model to soccer players and a novel rendering procedure to seamlessly blend multiple billboard components. Finally, Ballan et al. demonstrated that a billboard representation combined with a framework for generating optimised transitions can be used to produce very impressive free-viewpoint video transitions (Ballan et al., 2010).

2.2 Joint segmentation and reconstruction

Joint segmentation and reconstruction approaches extend previous reconstruction methods by incorporating segmentation or matting (opacity estimation) into a unified framework. Szeliski and Golland (Szeliski and Golland, 1999) proposed an energy minimisation approach for estimating disparities, colours and opacities in a generalised disparity space using iterative gradient descent. Similarly, De Bonet and Viola proposed the Roxel algorithm (De Bonet and Viola, 1999) which defines an iterative multi-step procedure alternatively estimating colours, responsibilities and opacities in a voxel space. Matusik et al. (Matusik et al., 2002) introduced the opacity hull generalising the view-dependent visual hull (Matusik et al., 2000) by adding view-dependent opacity; they show impressive results, but the approach requires active lighting to extract accurate mattes from each viewpoint and is not strictly speaking a joint approach as it requires prior matte estimation.

Various multi-view reconstruction frameworks have been extended to extract segmentation information. Probabilistic

estimation of voxel occupancy has been used to increase the robustness of the visual hull computation by replacing conventional binary occupancy labels by continuous values representing voxel opacity (Snow et al., 2000) or a probabilistic measure of occupancy eliminating the requirement to make hard segmentation decisions and reducing single view ambiguities (Franco and Boyer, 2005). Campbell et al. (Campbell et al., 2010) proposed a method for automatic multi-view object segmentation assuming cameras are fixating on the object of interest to initialise colour models which are then used to iteratively refine a visual hull estimate by volumetric graph-cuts.

Kolmogorov et al. (Kolmogorov et al., 2006) proposed two joint segmentation and reconstruction techniques based on dynamic programming and graph-cuts respectively capable of real-time segmentation. Although structurally similar to our approach, their techniques target a different application (background substitution) and remain limited to two-layer segmentation from narrow baseline stereo cameras. Additionally they assume accurate camera calibration. In contrast, our approach handles an arbitrary number of layers which is required for multi-player occlusions in sports and is formulated for robust reconstruction from wide-baseline views.

Zitnick et al. (Zitnick et al., 2004) computed accurate view-dependent layered representation by combining a colour segmentation-based algorithm which estimates smooth disparity maps for each camera with a Bayesian matting algorithm to refine thin boundary strips spanning depth discontinuities. Results obtained with an 8 camera array covering a 30° arc demonstrate free-viewpoint video interpolation with visual quality comparable to captured video. This technique requires accurate calibration and a narrow baseline camera configuration.

Goldlücke and Magnor (Goldlücke and Magnor, 2003) proposed a joint multi-view graph-cut reconstruction and segmentation algorithm based on global optimisation of an energy function combining photo-consistency information and colour probability models extending aprevious multi-view reconstruction (Kolmogorov and Zabih, 2002) by adding a background layer to the formulation. This approach produces a global scene reconstruction and is therefore prone to calibration errors. In addition, its applicability is limited to narrow baseline camera configurations.

Guillemaut et al. (Guillemaut et al., 2007, 2009) introduced a view-dependent optimisation framework which estimates a layered-depth representation with respect to each input camera. Joint segmentation and reconstruction is achieved by construction of a view-dependent graph similar to that used in (Roy and Cox, 1998) with an additional background layer (Guillemaut et al., 2007) or by optimisation of a more general view-dependent energy function incorporating additional energy terms and a discontinuity preserving smoothness term (Guillemaut et al., 2009). In all cases, optimisation is performed using graph-cut. This work forms the basis for the approach presented in this paper which extends the framework by generalising the energy function to enforce multiple-view consistency between cameras and introducing a method for robustly merging wide-baseline estimates when calibration accuracy is sufficiently high.

With the exception of (Guillemaut et al., 2007, 2009), all methods assume accurately calibrated 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 and segmentation.

2.3 Sports related research

There is a growing interest in virtual replay production in sports using free-viewpoint video techniques. Transferring studio techniques for free-viewpoint video to the sports arena is a notoriously difficult problem due to the much larger reconstruction area, use of moving and zooming cameras which must be calibrated on-the-fly, uncontrolled illumination conditions and presence of a natural background. Ideally, to avoid additional cost solutions will work from the existing broadcast cameras which are independently manually operated to follow play of interest. A key aspect to transfer studio-based reconstruction algorithm is robustness.

An initial attempt at transferring such algorithms was the Eye Vision system[1] based on the Virtualized Reality^{TM} technology introduced in (Narayanan et al., 1998). The system was used at Super Bowl XXXV (2001) to generate transitions by switching between cameras distributed around the stadium. To generate convincing transitions the system required use of a large number of cameras (> 30) and a motorised pan-tilt camera tracking system to guarantee that all cameras are simultaneously looking at the same point of interest. Transitions produced using this technology were very impressive but showed noticeable jumps between cameras.

A number of approaches have used view-interpolation (Connor and Reid, 2003; Kimura and Saito, 2005; Inamoto and Saito, 2007) or planar billboard representations (Ohta et al., 2007) to produce smooth transitions between viewpoints. Connor and Reid introduced a MAP framework for estimating a multi-view layered representation of background and foreground players used to synthesise novel views by linear interpolation. Inamoto and Saito (Inamoto and Saito, 2007) decompose soccer scenes into multiple layers representing static background objects and dynamic foreground players. Static background objects are approximated using a piecewise planar representation, allowing new view synthesis via homographic transformation of each planar com-

[1] Eye Vision, `http://www.ri.cmu.edu/events/sb35/tksuperbowl.html`

ponent. Foreground dynamic layers require a more complex modelling process to establish dense correspondence for view morphing. A similar approach decomposing a sport scene into layers has been applied in the context of tennis in (Kimura and Saito, 2005). Interpolation techniques are very appealing as they do not require explicit scene modelling or full calibration, however they tend to be limited to synthesising views in the vicinity of input cameras. Ohta et al. (Ohta et al., 2007) proposed a real-time technique based on a planer billboards which allows real-time free-viewpoint video. Use of a simple geometric proxy presents some advantages in terms of data representation and processing time, however lack of explicit geometry modelling may cause blending artefacts and distortions in a wide baseline camera configuration.

Volumetric reconstruction techniques such as shape-from-silhouette have been used for free-viewpoint video of soccer scenes in (Grau et al., 2005). The approach can be used to render views from arbitrary viewpoints on the pitch, however it requires a narrow baseline camera configuration and accurate calibration and segmentation to avoid player truncation artefacts. Recently more complex offline optimisation techniques based on graph-cuts (Guillemaut et al., 2007) or deformable models (Kilner et al., 2007) have been introduced to increase reconstruction accuracy. These techniques were able to achieve a visual quality comparable to that of the input images, however as in (Grau et al., 2005), results were only 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[2] which is limited to a single camera input and the Liberovision system[3] 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. Germann et al. (Germann et al., 2010) have proposed an interactive technique for fitting an articulated billboard model to a pair of images allowing novel view synthesis by piecewise interpolation of each billboard component. Impressive results were demonstrated from a single pair of cameras, however extension of the technique to full videos is non-trivial due to the requirement for manual interaction.

This paper extends previous research reported in (Guillemaut et al., 2009) by introducing a generalised framework to enforce multi-view consistency between cameras and additional shape priors (see Section 6). This increases the quality of the model generated compared to previous results reported in (Guillemaut et al., 2009). In addition, we describe a technique based on Poisson surface reconstruction to merge the view-dependent estimates into a single mesh model (see Section 8); this eliminates noise and produces a simpler mesh representation in the case of indoor data. A comprehensive qualitative and quantitative evaluation based on 13 datasets and containing several thousand frames of multi-view video sequences demonstrates the improvement over existing state-of-the-art techniques.

[2] Piero, `http://www.bbc.co.uk/rd/projects/virtual/piero`

[3] Liberovision, `http://www.liberovision.com`

3 Problem statement

The problem consists in estimating a layered depth representation, incorporating both segmentation of the image into its constituent layers and depth at each pixel in these layers, at each time instant and for each camera. Cameras are assumed to be synchronised. In addition we do not require a narrow baseline set-up as is often the case in stereo reconstruction; in fact we only assume a small number of cameras which are separated by a relatively large baseline. For simplicity, let us consider a single input camera, referred to thereafter as the reference camera, and its n neighbouring cameras indexed from 1 to n and referred to thereafter as auxiliary cameras; layered-depth reconstruction for all cameras is obtained by considering each input camera in turn as the reference camera. Our aim is to (i) partition the reference image into its constituent background/foreground layers and (ii) estimate the depth at each pixel for the foreground layers.

Formally, the problem can be formulated as a labelling problem where we seek the mappings $l : \mathscr{P} \rightarrow \mathscr{L}$ and $d : \mathscr{P} \rightarrow \mathscr{D}$, which respectively assign a layer label $l_{\mathbf{p}}$ and a depth label $d_{\mathbf{p}}$ to every pixel $\mathbf{p}$ in the reference image. $\mathscr{P}$ denotes the set of pixels in the reference image; $\mathscr{L}$ and $\mathscr{D}$ are discrete sets of labels representing the layer and depth hypotheses. $\mathscr{L} = \{l_1, \ldots, l_{|\mathscr{L}|}\}$ may consist of one background layer and one foreground layer (classic segmentation problem) or of multiple foreground and background layers. The set of depth labels $\mathscr{D} = \{d_1, \ldots, d_{|\mathscr{D}|-1}, \mathscr{U}\}$ is formed of depth values d_i obtained by sampling the optical rays together with an unknown label $\mathscr{U}$ accounting 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 only visible in a single camera.

In this paper, the focus is on foreground extraction and reconstruction. As a result, we assume multiple foreground layers corresponding to people/objects located at different depths and a single background layer. Depth is estimated only for foreground layers (background layers will received an unknown depth label). There are several motivations for not explicitly reconstructing the background. Firstly, this cannot be done reliably from the small number of cameras considered here (due to severe occlusions and framing limitations which prevent large portions of the background from being visible in more than a single view). Secondly, background reconstruction is often not required as post-production pipelines often only require foreground objects for composition with a virtual background set or a separately captured

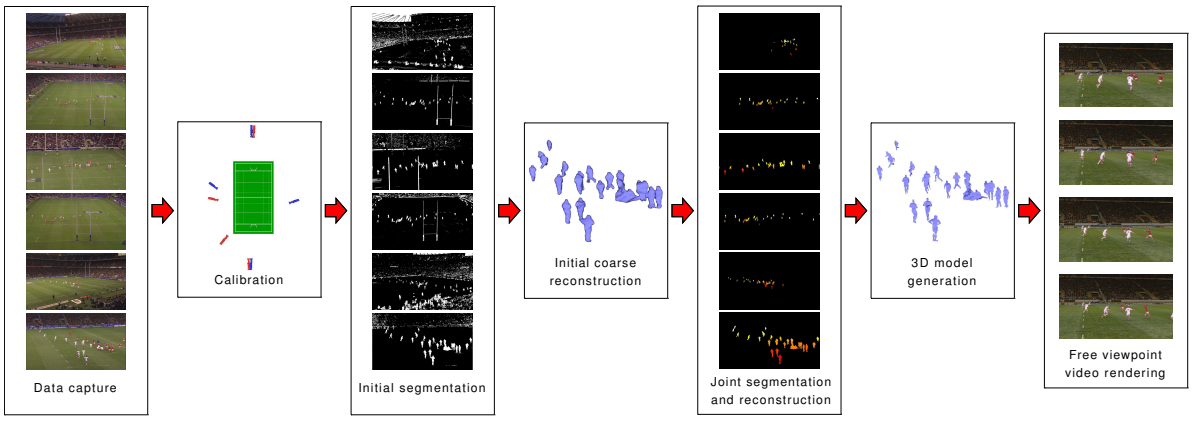


Fig. 3 Workflow of the proposed method.

photo-realistic model. Nonetheless our formulation remains general and could in principle estimate background depth.

4 Method overview

The proposed method consists of the following stages illustrated in Fig. 3.

Data acquisition: Data is captured using multiple synchronised video cameras surrounding the scene and separated by a wide-baseline. The cameras can be static or moving.

Calibration: In the case of indoor sequences, static cameras are pre-calibrated; for sports sequences, pre-calibration is not possible and all cameras are calibration on-the-fly using through-the-lens calibration techniques.

Initial segmentation: This is performed by applying a simple segmentation technique such as chroma-keying or difference keying to produce a coarse segmentation which is used to initialise the initial coarse reconstruction stage.

Initial coarse reconstruction: This stage uses a conservative silhouette intersection technique to extract a coarse foreground scene reconstruction based on visual hull (Laurentini, 1994). An error tolerant visual hull algorithm such as the conservative visual hull algorithm proposed in (Kilner et al., 2007) is necessary to account for errors in calibration and initial segmentation. This is used to define the possible layer and depth hypotheses at each pixel. Each connected mesh component of the visual hull defines a layer and a bound on its possible depth value.

Joint segmentation and reconstruction: The approach is based on joint estimation of layer and depth labels through optimisation of an energy function combining colour, contrast, matching, smoothness and multi-view consistency cues based on graph-cuts. The optimisation is performed in a view-dependent manner with respect to each input camera and is robust to errors in calibration and initial segmentation compared to global optimisation techniques. A novel approach to enforce multi-view consistency is introduced.

3D model generation: A final model is obtained by merging the layered-depth representations. For indoor sequences, a single surface representation is extracted through a Poisson surface reconstruction algorithm; for sports sequences which contain large calibration errors a simpler merging approach is preferred.

Free-viewpoint video rendering: A view-dependent texture mapping technique is used to synthesise novel views. The viewpoint is controlled interactively to render the free-viewpoint camera trajectory required for match analysis.

A critical part of the pipeline is the joint segmentation and reconstruction stage. By performing the layered-depth estimation in a view-dependent manner, we allow accurate segmentation and reconstruction in spite of calibration errors. This contrasts with global approaches which seek a unique scene representation but require high calibration accuracy for correct operation. Calibration inaccuracies produce inconsistencies limiting the applicability of global reconstruction techniques which simultaneously consider all views; view-dependent techniques are more tolerant to such inaccuracies because they only use a subset of the views for reconstruction of a layered depth from each camera view. Another important aspect is the joint optimisation of segmentation and reconstruction. Joint processing presents an advantage compared to sequential segmentation and reconstruction as it avoids propagation of errors between the two stages. Joint processing is also natural in the sense that many of the cues used for segmentation are useful for reconstruction and vice versa. For example, high contrast provides valuable information to identify a layer discontinuity but it is also an indicator of a possible depth discontinuity. Finally, a novel method to improve the multi-view consistency of the estimates corresponding to each camera view will be introduced. This approach will be described in details in Sections 5–7.

5 Joint segmentation and reconstruction

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

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

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

Colour and contrast terms are frequently used in segmentation problems, while matching and smoothness terms are normally used to solve reconstruction problems. Here we minimise an energy functional which simultaneously involves these two types of terms. The colour and contrast terms are similar to those considered in (Sun et al., 2006) with the main distinction that we extend the formulation to an arbitrary number of layers. Although structurally similar to other energy minimisation techniques, to achieve robust wide-baseline reconstruction our approach integrates multiple cues such as sparse affine covariant features and error tolerant photo-consistency scores. Individual cues are not novel in computer vision, but yield a novel non trivial formulation when combined. In the remainder of this section, we describe in detail how each term is formulated, the actual optimisation of the energy function being deferred to Section 7.

5.1 Colour term

The colour term uses learnt colour models for each layer or group of layers following a similar distribution in order to assign the most likely layer label at each pixel in the reference image. The term is defined as

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

where $P(\mathbf{I}_{\mathbf{p}}|l_{\mathbf{p}} = l_i)$ denotes the probability at pixel $\mathbf{p}$ in the reference image of belonging to layer l_i. Similarly to (Sun et al., 2006), the model for a layer l_i is defined as a linear combination of a global colour model $P_{\text{g}}(\mathbf{I}_{\mathbf{p}}|l_{\mathbf{p}} = l_i)$ and a local colour model $P_{\text{l}}(\mathbf{I}_{\mathbf{p}}|l_{\mathbf{p}} = l_i)$ such that

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

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 is more discriminative than a model considering either a purely local or global colour model and allows for dynamic changes in the background. It should be noted that the local model is only applicable 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)

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

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 (K_i was set to 5 in the case of sports data and 10 in the case of indoor data).

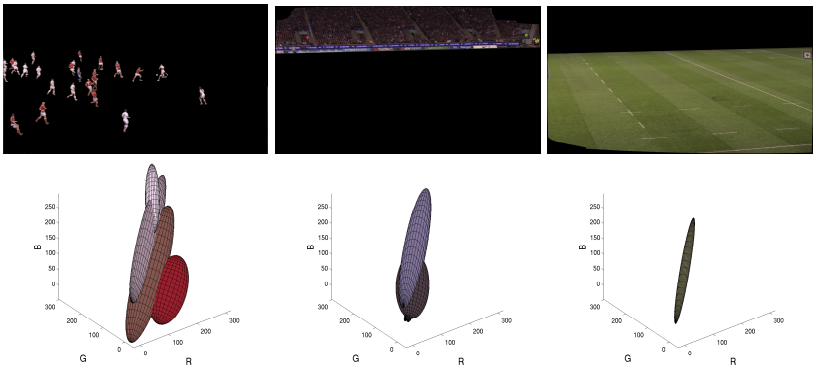
The local component of the colour model for a static layer l_i is represented by a single Gaussian distribution for each pixel $\mathbf{p}$:

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

where the parameters $\mu_{i\mathbf{p}}$ and $\Sigma_{i\mathbf{p}}$ represent the mean and the covariance matrix of the Gaussian distribution at pixel $\mathbf{p}$.

Learning the global colour model requires a single keyframe per camera in which each layer has been manually labelled. A GMM is then constructed for each layer from the annotated colour samples using the expectation maximisation algorithm (Dempster et al., 1977) initialised with a k-means clustering estimate. It should be noted that the models are not updated over time as our algorithm is only applied to relatively short sequences (500 frames). For the processing of longer sequences, more sophisticated algorithms could be used to model temporal variations in illuminations.

Learning the local colour models for static background layers requires multiple images in order to extract reliable statistics at each pixel. This is done for each camera either by explicitly capturing a background sequence where foreground people and objects have been removed from the scene or by estimating local models from the full scene image using a robust mosaicking technique to automatically remove the dynamic foreground objects. In the latter case, images are projectively warped to a reference image using calibration information followed by robust temporal filtering at each pixel to identify the background samples to be used for estimation of the local model. In either cases, a necessary assumption to be able to build the local model is that cameras are static or nodal (i.e. undergoing a combination of rotation and zooming); the absence of translation guarantees the existence of homographic relations between images necessary for mosaicking. Some examples of global and local colour models constructed in the case of a sport sequence can be seen in Fig. 4 and Fig. 5.



(a) Foreground model (b) Crowd model (c) Pitch model

Fig. 4 Example of global colour models in the case of a rugby sequence. The bottom row shows the GMM obtained from the training image (top row) for different layers; each component is represented by a 3σ-ellipsoid centred at the mean value.

Fig. 5 Example of local colour model. At each pixel, the local background model is represented by its mean value (top) and its variance (bottom); the variance image has been histogram equalised for display purposes.

5.2 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

$$E_{\text{contrast}}(l) = \sum_{(\mathbf{p},\mathbf{q})\in\mathscr{N}} e_{\text{contrast}}(\mathbf{p},\mathbf{q},l_\mathbf{p},l_\mathbf{q}), \text{ with} \tag{6}$$

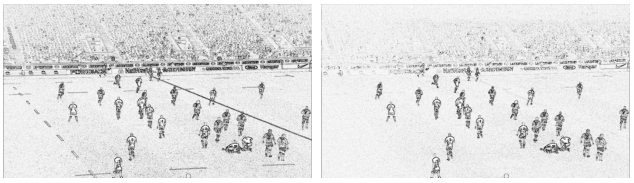
$$e_{\text{contrast}}(\mathbf{p},\mathbf{q},l_\mathbf{p},l_\mathbf{q}) = \begin{cases} 0 & \text{if } l_\mathbf{p} = l_\mathbf{q}, \\ \exp\left(-\beta C(\mathbf{I}_\mathbf{p},\mathbf{I}_\mathbf{q})\right) & \text{otherwise.} \end{cases} \tag{7}$$

$\mathscr{N}$ denotes the set of interacting pairs of pixels in $\mathscr{P}$ (a 4-connected neighbourhood is assumed) and $||\cdot||$ is the L_2 norm. $C(\cdot,\cdot)$ represents the squared colour distance between neighbouring pixels. β is a parameter weighting the distance function; it is set as $\beta = (2\langle||I_\mathbf{p} - I_\mathbf{q}||^2\rangle)^{-1}$ (Rother et al., 2004) where $\langle\cdot\rangle$ denotes the expectation operator.

The simplest choice for $C(\cdot,\cdot)$ would be a squared Euclidean colour distance $||I_\mathbf{p} - I_\mathbf{q}||$. This would yield a standard contrast term encouraging layer discontinuities to follow any edge in the image regardless of whether it is foreground or background (see Fig. 6(a)). Better performance can be obtained instead by using the attenuated contrast (Sun et al., 2006) defined as

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

where $z(\mathbf{p},\mathbf{q}) = \max(||I_\mathbf{p} - B_\mathbf{p}||, ||I_\mathbf{q} - B_\mathbf{q}||)$. $B_\mathbf{p}$ is the background colour at pixel $\mathbf{p}$ and is provided by the local component of the colour model defined in Section 5.1. The parameters K and σ_z are chosen as $K = \langle||B_\mathbf{p} - B_\mathbf{q}||\rangle$ and $\sigma_z = 2\langle z(\mathbf{p},\mathbf{q})^2\rangle$ respectively. The attenuated contrast uses background colour information to attenuate background edges.



(a) Standard (b) Adaptively attenuated

Fig. 6 Visualisation of different forms of contrast terms.

This produces a normalised contrast term where only foreground edges remain, thereby encouraging layer discontinuities to fall on foreground edges only (see Fig. 6(b)). In practice, some background edges may remain due to changes in background hoardings and crowd movement.

5.3 Matching term

The matching term encourages the reconstructed surface to be photo-consistent across the multiple views. This is based on the idea that correct depth hypotheses yield similar appearances across the images in which they are visible, while incorrect depth hypotheses usually result in inconsistent projected appearances. Our formulation combines two classes of similarity measures: a dense measure based on normalised cross-correlation (NCC) or photo-consistency enforced at each pixel in the reference image; and sparse feature constraints enforced at distinguishable image regions defined by affine covariant features that are usually more robust to changes in view-point and illumination. The matching term is defined as

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

Note that because the term $E_{\text{sparse}}(d)$ will be defined to impose a set of hard constraints, it is not necessary to weigh this term with respect to the term $E_{\text{dense}}(d)$.

Dense matching term This term evaluates similarity at each pixel in the reference camera and is defined as

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

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

$\mathbf{P}(\mathbf{p}, d_{\mathbf{p}})$ denotes the coordinates of the 3D point along the optical ray passing through pixel $\mathbf{p}$ which is located at a distance $d_{\mathbf{p}}$ from the reference camera. The function $S(\cdot,\cdot)$ measures the similarity between the reference camera and the auxiliary cameras in which the hypothesised point $\mathbf{P}(\mathbf{p}, d_{\mathbf{p}})$ is visible. $S_{\mathscr{U}}$ is a fixed penalty for labelling a pixel as unknown. Use of an unknown label adds robustness to the algorithm by providing a way to avoid labelling pixels that cannot be accurately reconstructed such as self-occluded pixels or pixels located at discontinuities.

Different metrics can be used to compute the similarity score S at a hypothesised surface point $\mathbf{P}$ between the reference camera and an auxiliary camera i. Let us denote by $\mathbf{p}$ and $\mathbf{q} = \pi_i(\mathbf{P})$ the projections of the hypothesised point $\mathbf{P}$ into the reference camera and the auxiliary camera i respectively. For sufficiently textured scenes, such as the ones considered in the indoor result section, a common choice in the multi-view reconstruction literature is the *NCC* computed over a square window centred on the pixels $\mathbf{p}$ and $\mathbf{q}$ (Seitz et al., 2006). The NCC value is between -1 and 1 and is then mapped to the interval $[0,1]$ as follows

$$s_1(\mathbf{p}, \mathbf{q}) = \exp(-\mu NCC(\mathbf{p}, \mathbf{q})), \tag{12}$$

where μ is the decay parameter (set to 1 in this paper). For less textured scenes, such as the ones considered in the sports result section, a more suitable metric is the photo-consistency (Seitz and Dyer, 1999) which we define as

$$s_2(\mathbf{p}, \mathbf{q}) = \frac{(\mathbf{I}_{\mathbf{p}} - \mathbf{I}^i_{\mathbf{q}})^2}{\sigma_i^2}, \tag{13}$$

where σ_i^2 normalises the photo-consistency measure for each auxiliary camera i; this parameter was fixed to 100 (images values being assumed to be in the range $[0,255]$) in all the experiments reported in this paper. To account for variations in scale and orientation, matching is performed in the space of rectified images; this guarantees similarity in scale and orientation during matching in spite of changes in camera zoom and framing.

To add robustness to calibration errors, the previous measures are computed over extended regions of radius r_{tol} rather than single pixels in a similar fashion to (Kutulakos, 2000). This defines error tolerant dense matching scores defined as

$$s'_1(\mathbf{p}, \mathbf{q}) = \max_{(\mathbf{q}-\pi_i(\mathbf{X}))^2 < r_{\text{tol}}} \exp(-\mu NCC(\mathbf{p}, \mathbf{q})), \tag{14}$$

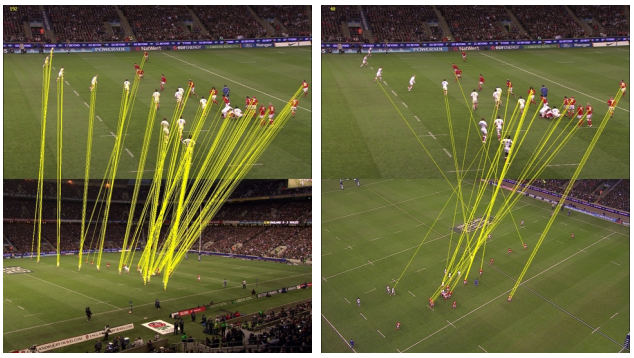


Fig. 7 Matched sparse features for two camera pairs.

and

$$s'_2(\mathbf{p}, \mathbf{q}) = \max_{(\mathbf{q}-\pi_i(\mathbf{X}))^2 < r_{\text{tol}}} \frac{(\mathbf{I}_{\mathbf{p}} - \mathbf{I}^i_{\mathbf{q}})^2}{\sigma_i^2}, \tag{15}$$

respectively in the case of NCC and photo-consistency. The parameter r_{tol} is set to 2 for sports data and 0 for indoor data.

Having defined the similarity score between the reference camera and an auxiliary camera, the final score $S(\cdot,\cdot)$ is obtained by combining the pairwise scores defined by the reference camera and all neighbouring auxiliary cameras. To make the measure tolerant to occlusions in some of the auxiliary cameras, a robust combination rule is defined as the sum of the k most photo-consistent pairs denoted by $\mathscr{B}_k$:

$$S(\mathbf{p}, \mathbf{X}) = \sum_{i\in\mathscr{B}_k} s'(\mathbf{p}, \pi_i(\mathbf{X})), \tag{16}$$

where s' is either s'_1 or s'_2 depending whether NCC or photo-consistency is used. Other forms of occlusions robust matching scores have been previously used in (Vogiatzis et al., 2007; Campbell et al., 2008) and have been shown to increase robustness to occlusions.

Sparse matching term This term is only active at the centre of distinguishable foreground regions at which a sparse feature correspondence has been established. The sparse matching score is defined as

$$E_{\text{sparse}}(d) = \sum_{\mathbf{p}\in\mathscr{P}} e_{\text{sparse}}(\mathbf{p}, d_{\mathbf{p}}), \text{ with} \tag{17}$$

$$e_{\text{sparse}}(\mathbf{p}, d_{\mathbf{p}}) = \begin{cases} 0 & \text{if } \mathscr{F}(\mathbf{p}) = \emptyset \text{ or } d_{\mathbf{p}} \in \mathscr{F}(\mathbf{p}), \\ \infty & \text{otherwise.} \end{cases} \tag{18}$$

$\mathscr{F}(\mathbf{p})$ denotes the set of depth labels located within a distance T from a sparse constraint at pixel $\mathbf{p}$. This forces the reconstructed surface to pass nearby 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. This parameter was set to 1 m for the sports scenes considered in this paper.

The sparse constraints considered are defined by the centre of affine-covariant features (Mikolajczyk et al., 2005; Mikolajczyk and Schmid, 2005) 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 overlap the foreground/background boundary cannot be used to establish reliable correspondences; we use the learnt colour models defined in Section 5.1 to discard such features based on their ratio of foreground and background pixels. Having detected likely candidate for sparse foreground matches, a SIFT descriptor is computed for each candidate region. Matching is then performed based on a nearest neighbour strategy. Robust matching is ensured by restricting the search to areas within a tolerance distance from the epipolar lines and making use of calibration information to cross validate the correspondences by verifying that they have similar rectified heights. The left-right spatial consistency (reciprocity) constraint is also enforced together with temporal consistency which requires corresponding features between camera views to be in correspondence temporally with the previous or the next frame. Examples of correspondences in the case of rugby can be seen in Fig. 7.

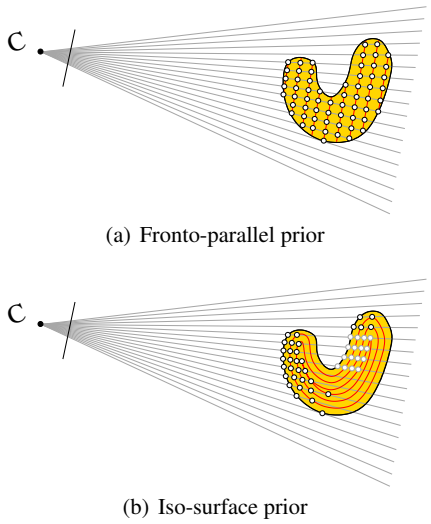
5.4 Smoothness term

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

$$E_{\text{smooth}}(l,d) = \sum_{(\mathbf{p},\mathbf{q})\in\mathcal{N}} e_{\text{smooth}}(l_\mathbf{p},d_\mathbf{p},l_\mathbf{q},d_\mathbf{q}), \text{ with} \tag{19}$$

$$e_{\text{smooth}}(l_\mathbf{p},d_\mathbf{p},l_\mathbf{q},d_\mathbf{q}) = \tag{20}$$
$$\begin{cases} \min(|d_\mathbf{p}-d_\mathbf{q}|,d_{\max}) & \text{if } l_\mathbf{p}=l_\mathbf{q} \text{ and } d_\mathbf{p},d_\mathbf{q}\neq\mathscr{U}, \\ 0 & \text{if } l_\mathbf{p}=l_\mathbf{q} \text{ and } d_\mathbf{p}=d_\mathbf{q}=\mathscr{U}, \\ d_{\max} & \text{otherwise.} \end{cases}$$

Discontinuities between layers are assigned a constant smoothness penalty $d_{\max}$ (set to 50 times the depth sampling step size for all datasets), 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 (Boykov et al., 2001; Kolmogorov and Zabih, 2004). This term also encourages unknown features to coalesce within each layer. The choice of depth prior adopted has an important influence on the smoothing effect and the reconstruction obtained. Two types of priors are considered in this paper.



(a) Fronto-parallel prior

(b) Iso-surface prior

Fig. 8 Comparison of different types of depth priors. Depth labels are represented as circles; grey circles represent second sets of iso-level labels ignored during graph-cut optimisation.

Fronto-parallel depth prior This is the most commonly used form of prior. It assumes that the depth is locally constant. In this case, a label $(l_\mathbf{p},d_\mathbf{p})$ corresponds to the point from layer $l_\mathbf{p}$ and located at a distance $d_\mathbf{p}$ from the reference camera centre along the ray emanating from pixel $\mathbf{p}$ (see Fig. 8(a)). The depth values $d_\mathbf{p}$ are uniformly sampled. Depth hypotheses are restricted to the interior of a visual hull estimate of the scene so as to restrict the number of depth hypotheses and improve efficiency. To account for calibration and matting error, we use the error tolerant visual hull proposed in (Kilner et al., 2007) which we call conservative visual hull. Although this yields good quality results when supported by strong matching cues such as in the case of the indoor scenes considered, this results in bias towards flat figure models which do not give good alignment between views in the case of sports data where matching cues are weak (see Fig. 17, 4$^{\text{th}}$ column).

Iso-surface depth prior An alternative approach which was introduced in (Guillemaut et al., 2009) in the context of sports scenes consists in placing samples along the iso-surfaces of the visual hull. This 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 $(l_\mathbf{p},d_\mathbf{p})$ corresponds to the first point of intersection between the ray emanating from pixel $\mathbf{p}$ and the $d_\mathbf{p}$-th iso-surface in the interior of the visual hull's connected component corresponding to layer $l_\mathbf{p}$ (see Fig. 8(b)). Unlike the fronto-parallel prior, the iso-surface prior is view-independent and results in more realistic reconstructions which are more likely to be consistent between views in the absence of strong matching cues (see Fig. 17). A drawback of the iso-surface prior is the overhead associated with iso-surfaces computation. Iso-surfaces

are defined by first detecting ray intersections with the mesh which gives the 0-th iso-surface; the remaining iso-surfaces are obtained by iterative search along the ray moving away from the camera centre until the distance stops increasing. An axis-aligned bounding box (AABB) tree structure (Cohen et al., 1995) is used to efficiently query the distance of any point to the mesh surface. Another limitation of the iso-surface approach is that it does not scale well to surfaces containing holes or folds as this results in multiple points bearing the same iso-surface value along a same ray within a layer; such points cannot be easily incorporated into our graph-cut optimisation framework which requires each depth hypothesis to be represented by a unique label. Incorporating multiple intersection within a layer would require introducing auxiliary labels representing the different intersections and ensuring that identical labels are spatially contiguous which is clearly non-trivial. We circumvent this problem by considering only the first set of iso-levels within each connected mesh component. This is a reasonable assumption for sports data which usually contains smooth meshes with genus 0. A simpler approach for introducing a visual hull bias in the reconstruction which does not suffer from these limitations will be introduced in the next section.

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 this does not change the formulation in Eq. (20) since the set of depth values remain an ordered set of discrete values.

6 Improving multi-view consistency

The layered depth extraction framework presented so far is completely view-dependent in the sense that each camera is considered in turn as a reference camera and processed independently from the remaining cameras. While this presents an advantage in robustness to calibration error compared to a purely global approach, it is desirable nonetheless to incorporate some elements of dependency between the processing of each camera in order to improve multi-view consistency. In this section, we extend the previous framework by locally enforcing some multi-view constraints. The main idea is to use the view-dependent layered depth estimates obtained separately for each camera as depth priors to constrain the layered depth estimation of neighbouring cameras. Also, for sports data this framework is used to bias the reconstruction towards the visual hull by treating the visual hull surface as an additional depth prior. This additional prior is view-independent and can be used to disambiguate the reconstruction problem in situations where photo-consistency information is poor such as in the case of sports data. This approach is simpler than the approach introduced in the previous section which required resampling the 3D space using the conservative visual hull iso-surfaces and it will be demonstrated to yield better result in Section 9.

The problem is formulated as the energy optimisation of the following cost function

$$E'(l,d) = E(l,d) + \lambda_{\text{aux}} E_{\text{aux}}(d) + \lambda_{\text{VH}} E_{\text{VH}}(d), \tag{21}$$

where $E_{\text{aux}}(d)$ and $E_{\text{VH}}(d)$ represents priors respectively defined by the auxiliary cameras' depth maps and the visual hull whose contributions are weighted by λ_{aux} and λ_{VH}. Enforcement of the multi-view constraints is performed in an iterative manner. At the first iteration, no layered depth reconstruction has been computed and so the term $E_{\text{aux}}(d)$ enforcing consistency between the different depth maps is ignored. At subsequent iterations, the depth map estimates obtained for each camera provide additional constraints defined in the term $E_{\text{aux}}(d)$, the process being iterated until convergence of the algorithm each time re-using the result of the previous iteration to constrain the current iteration. Additional constraints are provided by the term $E_{\text{VH}}(d)$ when a visual hull prior is enforced. In this case, the visual hull is updated after each iteration to benefit from the improvement in segmentation. In practice, two or three iterations are sufficient to obtain a noticeable improvement in multi-view consistency when only the term $E_{\text{aux}}(d)$ is used such as for indoor sequences; when using the additional term, such as for sports sequences, $E_{\text{VH}}(d)$ a single iteration is sufficient to enforce strong multi-view consistency.

We start by defining the term $E_{\text{aux}}(d)$ enforcing consistency between multiple depth maps. Let us denote by $d_i(d_{\mathbf{p}})$ the shortest Euclidean distance between the 3D point with depth $d_{\mathbf{p}}$ at pixel $\mathbf{p}$ in the reference camera and the dense set of 3D points corresponding to the depth map of the i-th auxiliary camera. Naïve computation involving evaluation of the distance for every possible pair would be prohibitively expensive due to the dense nature of the depth maps. Instead, an AABB tree (Cohen et al., 1995) is used to efficiently compute this distance. The term $E_{\text{aux}}(d)$ is defined as

$$E_{\text{aux}}(d) = \frac{1}{\sum_i \cos\gamma_i} \sum_i \sum_{\mathbf{p}} \cos\gamma_i [e_{\text{aux}}(d_i(d_{\mathbf{p}}))]^2, \text{ with} \tag{22}$$

$$e_{\text{aux}}(d_i(d_{\mathbf{p}})) = \begin{cases} \min(|d_i(d_{\mathbf{p}})|/d_{\max}, 1) & \text{if } d_{\mathbf{p}} \neq \mathscr{U}, \\ 0 & \text{if } d_{\mathbf{p}} = \mathscr{U}. \end{cases} \tag{23}$$

The energy is defined as the sum of squared distances weighted by the cosine of the angle between the reference camera and each auxiliary camera in order to favour more closely located cameras. To avoid over-penalising discrepancies occurring at discontinuities or due to calibration errors, a truncated linear distance is used. This prevents penalising large distances exceeding the pre-defined value $d_{\max}$ introduced in Section 5.4.

The optional term $E_{\text{VH}}(d)$ enforcing consistency with the visual hull is similarly defined. Let us denote by $d_{\text{VH}}(d_{\mathbf{p}})$

the shortest Euclidean distance between the 3D point with depth $d_{\mathbf{p}}$ at pixel $\mathbf{p}$ in the reference camera and the visual hull surface. The term is defined as

$$E_{\mathrm{VH}}(d) = \sum_{\mathbf{p}} [e_{\mathrm{VH}}(d_{\mathrm{VH}}(d_{\mathbf{p}}))]^2, \text{ with} \tag{24}$$

$$e_{\mathrm{VH}}(d_{\mathrm{VH}}(d_{\mathbf{p}})) = \begin{cases} \min(|d_{\mathrm{VH}}(d_{\mathbf{p}})|/d_{\max}, 1) & \text{if } d_{\mathbf{p}} \neq \mathscr{U}, \\ 0 & \text{if } d_{\mathbf{p}} = \mathscr{U}. \end{cases} \tag{25}$$

Note that unlike the iso-surface prior introduced in the previous section, this approach does not require the reconstructed surface to lie inside the volume defined by the prior. This presents some advantage compared to the iso-surface approach as this does not require use of a conservative visual hull which by nature tends to produce a bias towards a dilated reconstruction. Use of a standard visual hull instead tends to produce a more photo-consistent reconstruction in non-truncated areas such as the trunk of the players while possibly truncated visual hull areas such as the arms and legs will rely on the standard smoothness prior to be correctly recovered. The effects of these energy terms on the final reconstruction will be illustrated in Section 9.2.

7 Energy optimisation

7.1 Optimisation using expansion move algorithm

Optimisation of the energy defined by Eq. (21) is known to be NP-hard. However, an approximate solution can be computed using the expansion move algorithm based on graph-cuts (Boykov et al., 2001). Proof of the regularity of the energy function, required for α-expansion optimisation, is provided in the Appendix. The expansion move proceeds by cycling through the set of labels $\alpha = (l_\alpha, d_\alpha)$ in $\mathscr{L} \times \mathscr{D}$ and performing an α-expansion iteration for each label until the energy cannot be decreased (see (Boykov et al., 2001)). An α-expansion iteration is a change of labelling such that each pixel $\mathbf{p}$ 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 (Boykov and Kolmogorov, 2004). After convergence of the algorithm, the result obtained is guaranteed to be a strong local optimum (Boykov et al., 2001). For all experiments reported in this paper, the α-expansion algorithm was initialised with the conservative 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.

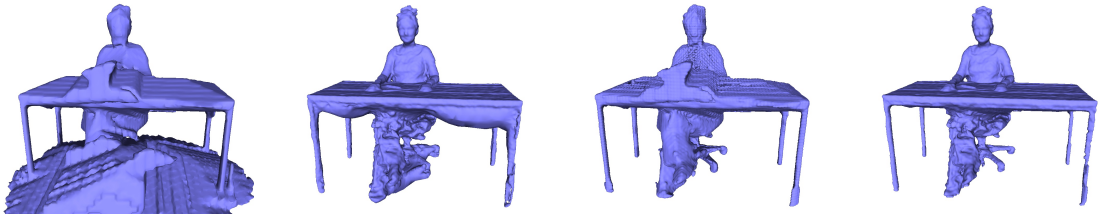
7.2 Improving efficiency

Reusing the flow We improve computation and memory efficiency by dynamically reusing the flow at each iteration of the min-cut/max-flow algorithm (Alahari et al., 2008). The idea of recycling the dual (flow) solutions is based on the concept of re-parametrisation introduced in (Kohli and Torr, 2007) in the case of dynamic MRFs with sub-modular energy function and extended to non-sub-modular energy functions in (Alahari et al., 2008). During the first cycle of iterations, the flow is computed in a normal manner for each α-expansion and is stored. During subsequent iterations, the flow is obtained by reusing the flow computed at the previous iteration using re-parametrisation. This resulted in a speed-up of an order of two for the scenes considered.

Multi-scale implementation Solving a labelling problem involving all pixels in the reference image can be problematic with large dimensional images. In practice, the problem is restricted to a subset of the pixels through use of the visual hull, however even after that the total number of pixel to label may be prohibitively large especially in the case of indoor images where foreground is usually more tightly framed. A considerable memory saving is obtained by running the first iteration at a coarse resolution (sub-sampled by a factor of 3 in the case of the indoor scenes, full resolution in the case of sports scenes), and subsequent iterations at full resolution after having re-computed the conservative visual hull after each iteration. Subsequent iterations are usually faster to process as most background pixels are usually removed after the first iteration.

8 Model generation and rendering

A model is generated from the obtained layered depth reconstruction and used for final rendering. Two approaches are considered for generating the model depending on the size of calibration errors. In the case of large calibration errors (1–2 pixels) such as with the sports sequences, each layered depth representation defining a view-dependent 2.5D foreground scene representation is converted into a regular mesh with vertices defined by image pixel locations. Vertex connectivity is decided based on the layer segmentation and thresholding of the angle separating the line segment connecting 3D surface points defined by pairs of neighbouring pixels and the optical ray passing through the midpoint of the pixel pair (a threshold of 80° is used). This allows pixel belonging to different layers or located at a depth discontinuity to be correctly converted into separate mesh components. For sports scenes, rendering is performed using only a subset of the meshes obtained (those closest to the virtual camera path) rather than the full set of meshes in order to alleviate artefacts due to combining multiple inaccurately



(a) Visual hull (b) Poisson (c) Poisson+visual hull (d) Proposed

Fig. 9 Meshes obtained through application of different merging techniques to the depth maps shown in the bottom row of Fig. 14.

calibrated reconstructions. From a practical point of view, the virtual sequence is divided into a set of shots, each using a different subset of cameras. Artefacts due to switching from one shot to the other are avoided by ensuring that successive shots have a sufficient number of shared cameras. Rendering is performed using view-dependent texture mapping (Debevec et al., 1996). The colour at a given pixel in the virtual camera is obtained by linear blending of the nearest two cameras weighted by their angular distance to the virtual camera. To avoid hard edges, the binary segmentation is refined into an alpha matte by defining a narrow band (2–4 pixels) uncertainty regions along the segmentation edges and an alpha matte is estimated using the closed-form matting algorithm proposed in (Levin et al., 2008).

If calibration errors are low ($<$1 pixel), improved results can be obtained by merging the different representations into a unique global representation using Poisson surface reconstruction (Kazhdan et al., 2006). The algorithm produces a watertight reconstruction and eliminates the redundancy contained in the view-dependent meshes. Direct application of the algorithm is usually not sufficient to obtain an accurate reconstruction as there can be large portions of the foreground which are not visible in any camera which tends to produce large protrusions (see Fig. 9(b)). In (Vlasic et al., 2009), Vlasic et al. compensate for the missing data by including the visual hull mesh in the set of oriented meshes used for Poisson surface reconstruction. For the data considered here, the visual hull is usually not accurate enough to be combined with the view-dependent reconstructions without introducing significant artefacts (see Fig. 9(c)). Instead we perform a two-pass Poisson surface reconstruction. During the first pass, only the points defined by the different view-dependent depth maps are used, which yield a reconstruction containing protrusions at occluded regions (Fig. 9(b)). In the second pass, protrusions are removed by adding visual hull sample points located inside the reconstruction obtained at the first pass to the view-dependent sample points. This effectively reduces the addition of points to badly reconstructed areas, thus removing protrusions without deteriorating correctly reconstructed regions (see Fig. 9(d)). It should be noted that the Poisson surface reconstruction takes as input an oriented set of points while the depth recovered is not orientated. Orientation is estimated based on neighbouring pixels.

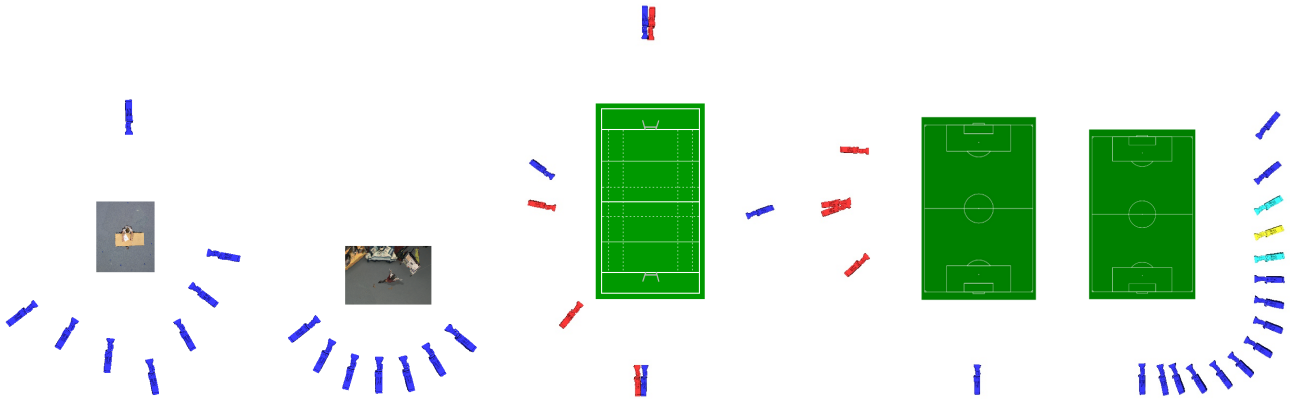
9 Results

Evaluation of the algorithm is performed using a variety of datasets with relatively wide-baseline camera placements. These can be divided into three main categories: indoor datasets (Office and Dance1 captured with high quality Thomson Viper studio cameras, Dance2 and Ball captured with lower quality Canon camcorders), outdoor sports datasets (from an international 6-Nations rugby game and a European Championship soccer game, both captured using manually operated zooming and rotating broadcast cameras and a few additional static cameras), and evaluation datasets (Middlebury multi-view datasets and narrow-baseline soccer dataset). Frames from each dataset are shown in the first column of Fig. 11 and Fig. 12. The characteristics of the these datasets are provided in Table 1; camera placement is illustrated in Fig. 10. Experiments were performed on a server with 2 quad-core 2.26 GHz Intel Xeon processors. The datasets considered cover a wide variety of capture conditions ranging from large-scale outdoor scenes with very little control on the capture conditions to smaller scale indoor scenes with different degrees of background complexity, image quality and camera configurations.

For all datasets, cameras are assumed to be calibrated. In the case of the indoor sequences, calibration is performed in a two step approach where chart based calibration (Zhang, 2000) is first used to separately estimate the intrinsic parameters for each camera and then a wand-based calibration algorithm is used to refine these parameters and estimate the extrinsic parameters in a bundle-adjustment framework (Mitchelson and Hilton, 2003). This typically produces sub-pixel accuracy calibration. In the case of the sports sequences, calibration is performed through-the-lens from pitch line features as described in (Thomas, 2007). 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).

Table 1 Characteristics of the different datasets and parameter settings.

Name	Nunmber of cameras	Number of frames	Image resolution	λ_{colour}	$\lambda_{\text{contrast}}$	λ_{match}	λ_{smooth}	λ_{aux}	λ_{VH}	w	Run-time (per frame)
Office	8 (all static)	125	1920×1080	0.6	1.0	0.4	0.001	0.5	0.0	0.3	13min56s
Dance1	8 (all static)	250	1920×1080	0.6	1.0	0.4	0.001	0.5	0.0	0.01	8min12s
Dance2	7 (all static)	400	1920×1080	0.6	5.0	0.4	0.001	0.5	0.0	0.05	4min00s
Ball	7 (all static)	400	1920×1080	0.6	5.0	0.4	0.001	0.5	0.0	0.05	4min12s
Rugby	8 (4 moving)	300	1920×1080	0.5	1.0	0.2	0.1	0.0	0.5	0.5	3min28s
Soccer	5 (4 moving)	100	1920×1080	0.5	1.0	0.2	0.1	0.0	0.5	0.5	2min25s
Temple Full	312	1	640×480	0.5	10.0	0.5	0.0001	0.2	0.0	0.0	3h26min00s
Temple Ring	47	1	640×480	0.5	10.0	0.5	0.003	0.2	0.0	0.0	47min00s
Temple Sparse	16	1	640×480	0.5	10.0	0.5	0.003	0.2	0.0	0.0	31min00s
Dino Full	363	1	640×480	0.0	10.0	1.0	0.0001	0.5	0.0	0.0	8h14min00s
Dino Ring	48	1	640×480	0.5	10.0	0.5	0.02	0.2	0.0	0.0	55min00s
Dino Sparse	16	1	640×480	0.5	10.0	0.5	0.02	0.2	0.0	0.0	34min00s
Soccer-eval	15 (all static)	100	720×288	0.5	1.0	0.2	0.1	0.0	0.5	0.5	1min03s



(a) Office/Dance1 (b) Dance2/Ball (c) Rugby (d) Soccer (e) Soccer-eval

Fig. 10 Camera layout for the different datasets. Moving cameras are represented in red, while static cameras are represented in blue. For the Soccer-eval dataset, the ground truth camera is indicated in yellow and additional left-out cameras are shown in cyan.

The same pipeline described in Fig. 3 is employed for all datasets. Initial segmentation is obtained by computing a background image plate for each input view using standard mosaicking techniques (this does not require separate background capture) and thresholding the colour difference between the background plate and original image. Initial segmentation is often inaccurate due to overlapping foreground and background colour distributions, which is exacerbated by the presence of compression artefacts and motion blur in the case of the sports data. 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 (Kilner et al., 2007). Joint refinement of segmentation and reconstruction is then performed in parallel for each input camera, using the set of neighbouring cameras as auxiliary cameras. For sports data, usually no more than 1 or 2 auxiliary cameras are usable due to the wide-baseline. The local colour models used for 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; the numbers of components in the mixture models were set to 5 in the case of the sports datasets and 10 in the case of the indoor datasets. The depth sampling step was set to 10 mm for all datasets except for the Middlebury evaluation datasets in which case a 0.5 mm step was used. The matching score used was the NCC in the case of all indoor sequences (window size of 5×5 for the Middlebury datasets and 10×10 for all other datasets) and photo-consistency in the case of all sports sequences. A summary of the parameter settings not described in the main text is given in Table 1.

The evaluation performed using these datasets covers three main aspects: segmentation, reconstruction, and application to free-viewpoint video. In each case, the algorithm is evaluated, both qualitatively and quantitatively, against state-of-the-art algorithms in the field. For the qualitative evaluation, examples shown in the paper had to be restricted to single frames, however results on full sequences showing

free-viewpoint video results together with intermediate segmentation and reconstruction results are provided in the supplementary video available from `http://www.guillemaut.org/publications/11/GuillemautIJCV11/`. Due to copyright restrictions the soccer and rugby videos cannot be released publicly.

9.1 Segmentation results

The proposed technique is compared against the following techniques:

Global colour model/Chroma keying: This approach uses a global colour model in the form of a Gaussian mixture model (GMM) to represent foreground and background layers. Each model is learnt from a manually segmented training image; the number of mixtures was manually set based on the complexity of the scene. Classification into foreground or background is made by identifying the layer which maximises the probability function derived from the colour models. This approach is used for the indoor datasets. In the case of the sports datasets, which tend to contain simpler distributions at the pitch level, a user assisted chroma keying technique was used instead (Grau et al., 2007).

Difference keying: This approach uses a background plate representing the mean colour value at each pixel. Foreground pixels are identified by first differencing the current image with the background plate and then thresholding the colour difference. For this approach, the threshold was manually adjusted for optimum results. This approach was used to initialised the proposed join optimisation technique.

Background cut (Sun et al., 2006): This approach combines a global model for foreground and background with a local model similar to that used in difference keying and additional contrast information to encourage the segmentation boundary to follow high contrast edges. The approach uses graph cut to find an optimal labelling.

9.1.1 Qualitative evaluation

Results for all the techniques are presented in Fig. 11 and Fig. 12 for selected frames; results on full sequences can be seen in the accompanying video. For sports, chroma keying tends to produce poor results as its applicability remains limited to grass areas, thus producing large errors in crowd areas and at line locations. With indoor data, the technique based on a global colour model also performs poorly, particularly for scenes with complex overlapping background and foreground colour distributions such as those of the Dance2 and Ball sequences. Difference keying usually produces better results due to the locality of the colour model it uses, however it fails in non static background areas, such as the crowd in the case of the sports sequences, or in areas where foreground and background colours are similar. This approach also tends to fail in shadowed areas as can be seen in the case of the Office and Dance1 sequences. Background cut produces better results than the previous two method 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. In Fig. 11 the only observable artefacts with the proposed method compared to ground truth are in areas of strong shadows.

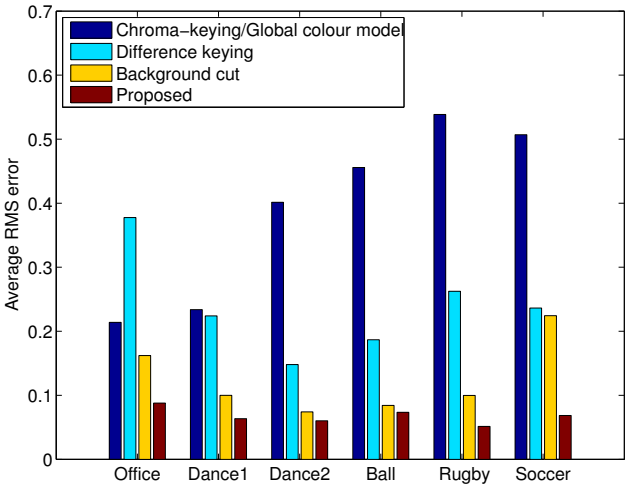


Fig. 13 Quantitative evaluation of segmentation accuracy.

9.1.2 Quantitative evaluation

In order to quantify the segmentation accuracy of the different algorithms, ground truth data was generated for all sequences. Ground truth was produced by manually segmenting, for each dataset, a set of images regularly sampled during the length of the sequence and equally distributed among all cameras. At least 30 images were annotated per dataset. The error measure used for evaluation is the root mean squared (RMS) error averaged over all images used for evaluation. The results, shown in Fig. 13, confirm the previous qualitative observations. Approaches based solely on a global colour model usually performing worst, while the second worst performer is normally difference keying, except in the case of the Office sequence where it is the worst performer; this is explained by the presence of large shadows and a relatively simple background colour distribution. Background cut performs significantly better than the previous two methods, while the proposed technique is consistently ranked first. It should be noted that different im-

Fig. 11 Segmentation results with indoor datasets for different techniques (see supplementary video for full sequences).

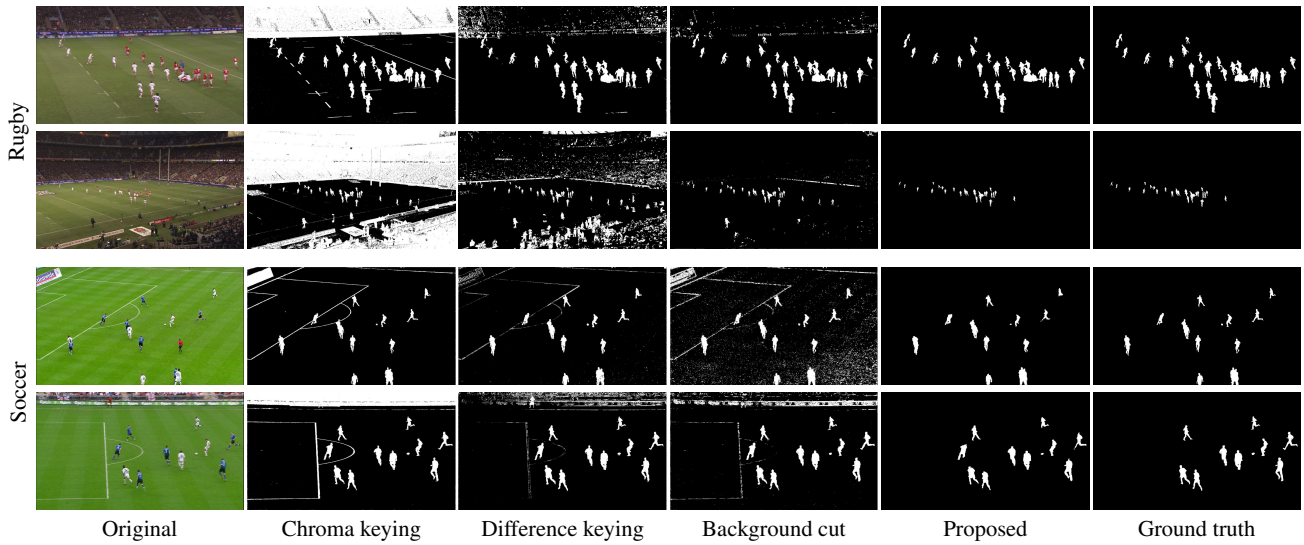


Fig. 12 Segmentation results with sport datasets for different techniques (see supplementary video for full sequences).

plementations of the proposed method are possible based on the choice of prior introduced in Section 5.4. Our experience is that the impact on segmentation is very small both qualitatively and quantitatively, and for this reason we decide not to include a comparison. These methods and their effects on reconstruction will be discussed in the next section.

9.2 Reconstruction results

In the case of indoor data, the following methods have been used for comparison:

Visual hull: This uses a standard shape-from-silhouette technique (Laurentini, 1994) described in Section 4. The segmentation used to compute the visual hull is obtained from difference keying; in the case of indoor data captured under controlled illumination this tends to produce a conservative segmentation suitable for visual hull computation without the need for introduction of additional tolerance during reconstruction. The visual hull will be used as an initial estimate for all other methods described below.

Starck (Starck and Hilton, 2007): This is a volumetric optimisation technique which uses graph-cut to find an optimum dense reconstruction maximising stereo photo-consistency and satisfying silhouette and sparse feature constraints. The method has been widely used for wide baseline studio reconstruction with a chroma-key blue screen and application to free-viewpoint video. It requires initialisation with the visual hull.

Furukawa (Furukawa and Ponce, 2010): This approach uses a three stage approach where sparse matches are first identified to define seed points which are then expanded to neighbouring regions using a growing approach and finally filtered to remove outliers based on a visibility constraint. This approach is one of the top performers according to the Middlebury multi-view evaluation page. We used the Patch-based Multi-view Stereo Software (PMVS - Version 2) which has been made publicly available[4]. Initialisation with the visual hull is optional with this technique. For a fair comparison, like other techniques, this technique was initialised with the visual hull.

In the case of sports data, techniques such as (Starck and Hilton, 2007) and (Furukawa and Ponce, 2010) are not suitable due to large calibration errors and poor resolution which prevent extraction of robust silhouette and sparse constraints and convergence to a global solution. Instead, the following simpler methods have been used for comparison:

Visual hull: This is the same as previously described.

[4] `http://grail.cs.washington.edu/software/pmvs/`

Conservative visual hull (Kilner et al., 2007): This is an error-tolerant implementation of the visual hull algorithm where a voxel is deemed occupied if its projection lies within a given threshold from the foreground in all images. This is equivalent to dilating the foreground regions prior to conventional visual hull computation.

Stereo refinement: This performs a view-dependent dense stereo reconstruction based only on photo-consistency (no colour, contrast or smoothness term) as defined in Section 5.3. While simple, this method provides a good indication of what can be achieved using stereo information.

The following three different implementations of the proposed algorithm have been tested:

Proposed-standard: This uses the standard fronto-parallel (locally constant) depth prior described in Section 5.4 with no additional visual hull prior ($\lambda_{\mathrm{VH}} = 0$ in Eq. (21)).

Proposed-iso: This uses the iso-surface based depth prior introduced in Section 5.4 with no additional visual hull prior ($\lambda_{\mathrm{VH}} = 0$ in Eq. (21)).

Proposed-hybrid: This uses the standard fronto-parallel prior described in Section 5.4 combined with the visual hull prior introduced in Section 6 ($\lambda_{\mathrm{VH}} = 1$ in Eq. (21)).

It should be noted that we did not include the case where the iso-surface prior is combined with the visual hull prior. While there is nothing in principle preventing the simultaneous use of these two priors, we found that there is no benefit in doing so since the visual hull prior provides sufficient information to disambiguate the problem. In fact use of the visual hull prior with the iso-surface prior is likely to be detrimental in this case as it will increase the run-time compared to a standard depth prior. Note also that in the case of indoor data, there would be no benefit in using the iso-surface or visual hull priors since there are strong matching cues. As a result, only the first implementation was used in this case. In the case of the sports sequences, which present very weak matching cues, all three approaches are compared. For indoor data, three iterations of optimisation of the energy defined in Eq. (21) are performed in order to improve multi-view consistency. In the case of sports data, which is already constrained using additional priors, a single iteration is performed.

9.2.1 Qualitative evaluation

Fig. 14 and Fig. 15 show some examples of depth maps estimated in the case of the Office and Rugby datasets respectively. In the case of the sports sequences, a single iteration of the algorithm was sufficient due to the use of additional priors to constrain the reconstruction. In the case of indoor datasets, we do not use any additional prior but perform several iterations (three) in order to improve depth

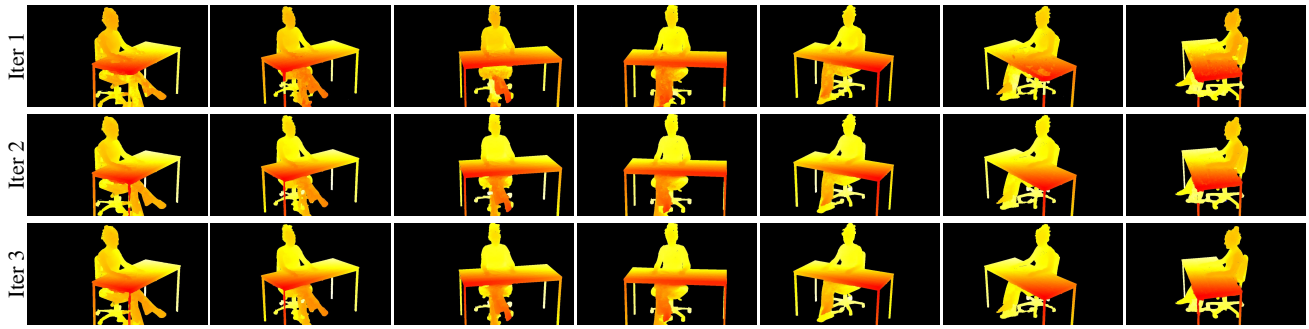


Fig. 14 Depth maps obtained in the case of the Office dataset. Rows show the evolution of the estimated depth maps with respect to the number of iterations performed. Columns correspond to different camera locations.

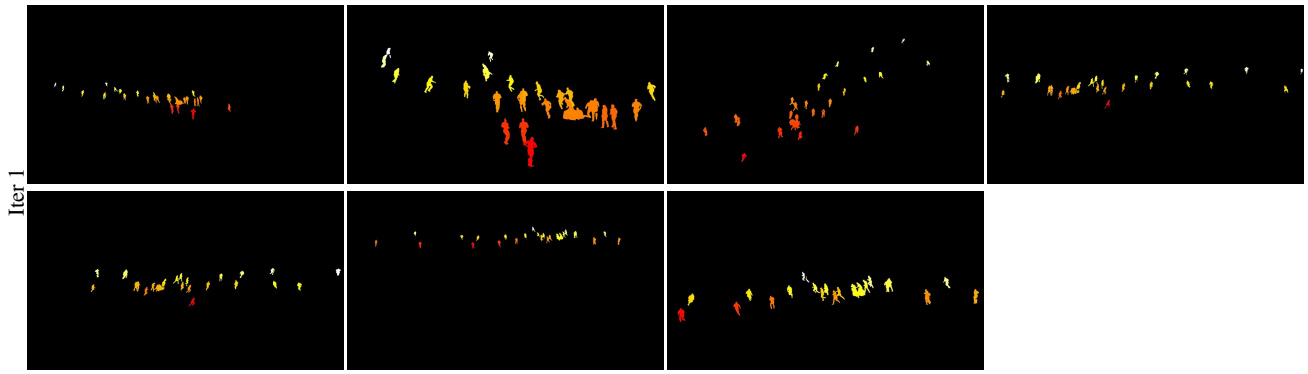


Fig. 15 Depth maps obtained in the case of the Rugby dataset. A single iteration is sufficient to obtain consistent depth maps due to the use of the additional visual hull prior. Columns correspond to different camera locations.

map consistency. Note how the multi-view consistency term $E_{\text{aux}}(d)$ reduces noise in the estimated depth maps and reduces artefacts occurring on poorly textured regions such as the legs of the table. The final meshes obtained (both untextured and textured) rendered from a novel viewpoint are shown in Fig. 16 and Fig. 17 in the case of indoor and sports data respectively.

In the case of indoor data, the visual hull produces large protrusions in shadowed areas which were erroneously segmented as foreground (Dance2 and Ball sequences). Both (Starck and Hilton, 2007) and (Furukawa and Ponce, 2010) produce an improved reconstruction, however they both fail to produce an adequate reconstruction in the shadow regions due to their inability to refine the segmentation. In contrast, the proposed technique is able to refine the segmentation and produce a much improved reconstruction of the foreground. To evaluate the performance of the techniques with high quality segmentation input, these techniques were also applied to the segmentation output of the proposed approach. These techniques are referred to as *Visual hull + mattes*, *Starck + mattes* and *Furukawa + mattes*. Use of the high quality segmentation produced by the proposed technique significantly improves the quality of foreground reconstruction, in particular for the Office and Dance1 sequences. (Starck and Hilton, 2007), (Furukawa and Ponce, 2010) and the proposed technique perform generally very well. An apparent limitation of (Starck and Hilton, 2007) is its difficulty to reconstruct thin or flat structures such as the desk where it tends to produce holes or truncation. These errors are a drawback of the volumetric formulation which is biased towards smaller surface areas. In contrast, (Furukawa and Ponce, 2010) and the proposed technique do not suffer from these artefacts. They are both able to reconstruct fine details such as cloth wrinkling. A significant advantage of the proposed technique is its ability to jointly perform the segmentation and reconstruction, thus allowing operation with much lower quality input segmentation.

In the case of sports data, as expected the visual hull produces large truncations due to 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 techniques yield a smooth reconstruction with accurate player boundaries and the elimination of phantom volumes. The proposed approach based on the standard depth prior tends to produce view-dependent reconstructions biased towards flat models which result in errors when combined. The proposed ap-

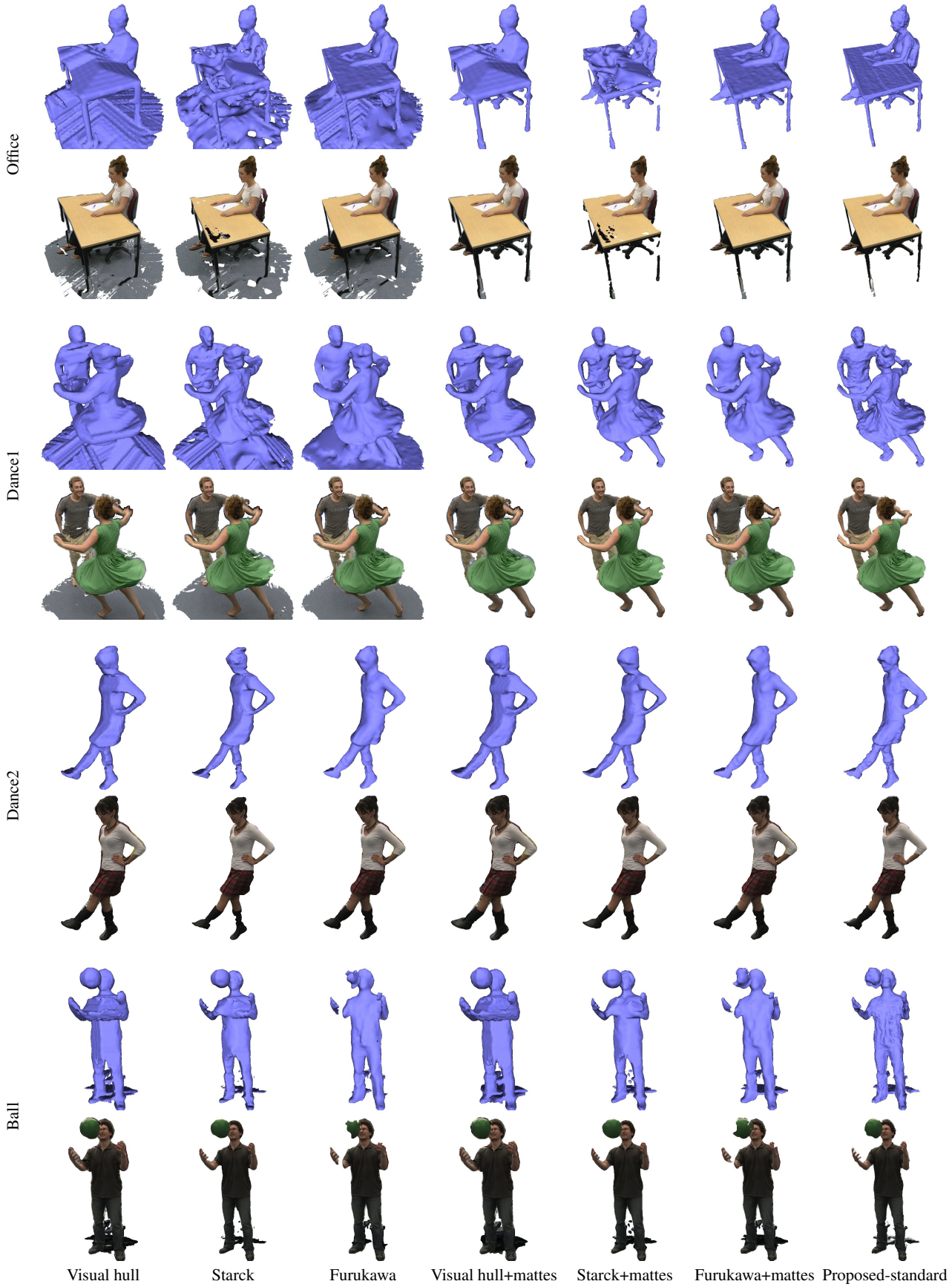


Fig. 16 Reconstruction results with indoor datasets for different techniques (see supplementary video for full sequences).

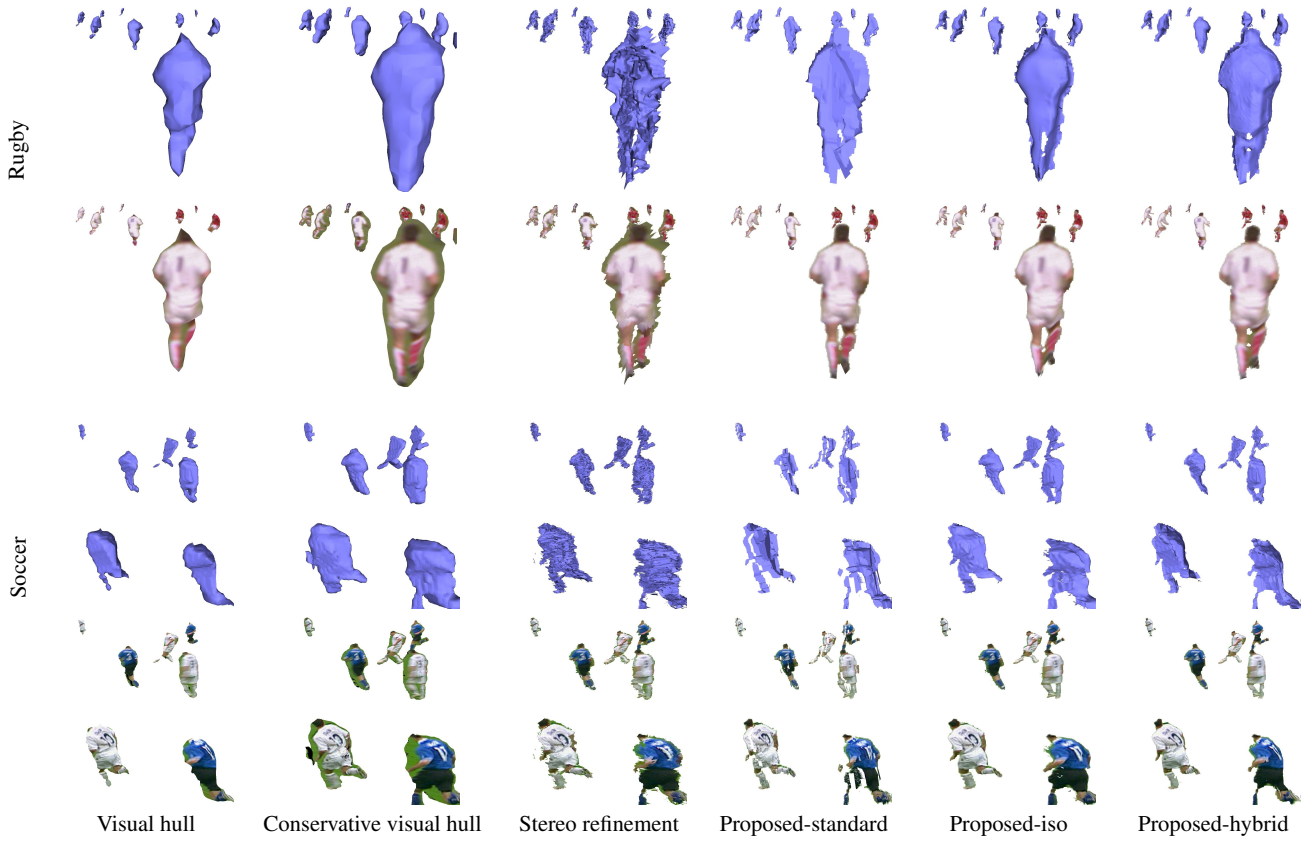


Fig. 17 Reconstruction results with sport datasets for different techniques (see supplementary video for full sequences).

proach based on the iso-surface prior introduced in Section 5.4 produces models which are naturally more consistent due to their bias towards the conservative visual hull iso-surfaces, however the models may appear inflated due to the conservativeness of the prior introduced. Finally, the proposed hybrid approach which combines the standard prior with an additional visual hull prior is naturally consistent and does not suffer from the inflation observed in the previous case. This produces the most consistent results among the three methods compared.

9.2.2 Quantitative evaluation

Evaluation of the algorithm for reconstruction with the types of datasets considered so far is not feasible due to the difficulty in capturing ground truth data for video sequences. Instead, the performance of the proposed algorithm is evaluated using the Middlebury multi-view dataset[5] following the methodology described in (Seitz et al., 2006) which considers both accuracy and completeness. Results where evaluated for all six datasets, i.e. the dino and temple datasets for all camera configurations (sparse, ring and full). See Fig. 18 for some images of the models obtained. A summary of the results obtained for the proposed technique in the case of the standard depth prior are shown in Table 2. For reference, results reported using (Starck and Hilton, 2007) and (Furukawa and Ponce, 2010) were also included. Full results can be seen on the Middlebury evaluation webpage.

The results show that the proposed algorithm is able to produce accurate and complete reconstructions across the range of camera configurations considered. In particular, the algorithm performs extremely well for the full datasets. It is ranked fifth in terms of accuracy for the Temple Full dataset (better than (Furukawa and Ponce, 2010)) and third in terms of accuracy for the Dino Full dataset (in both cases only a few hundredths of a mm behind the top performers). The algorithm is also top performer for the Dino Full dataset in terms of completeness (the only algorithm achieving a score of 100%). For sparser datasets, the algorithm is still able to achieve good performance, being usually ranked in the middle of the table. This is a rather good performance bearing in mind that algorithms are usually separated by very small differences in accuracy or completeness. The proposed approach performs significantly better than (Starck and Hilton, 2007) for all datasets. It should be noted that unlike other algorithms in the Middlebury multi-view evaluation which only address reconstruction, our approach addresses the joint problem of reconstruction and segmentation.

[5] http://vision.middlebury.edu/mview/

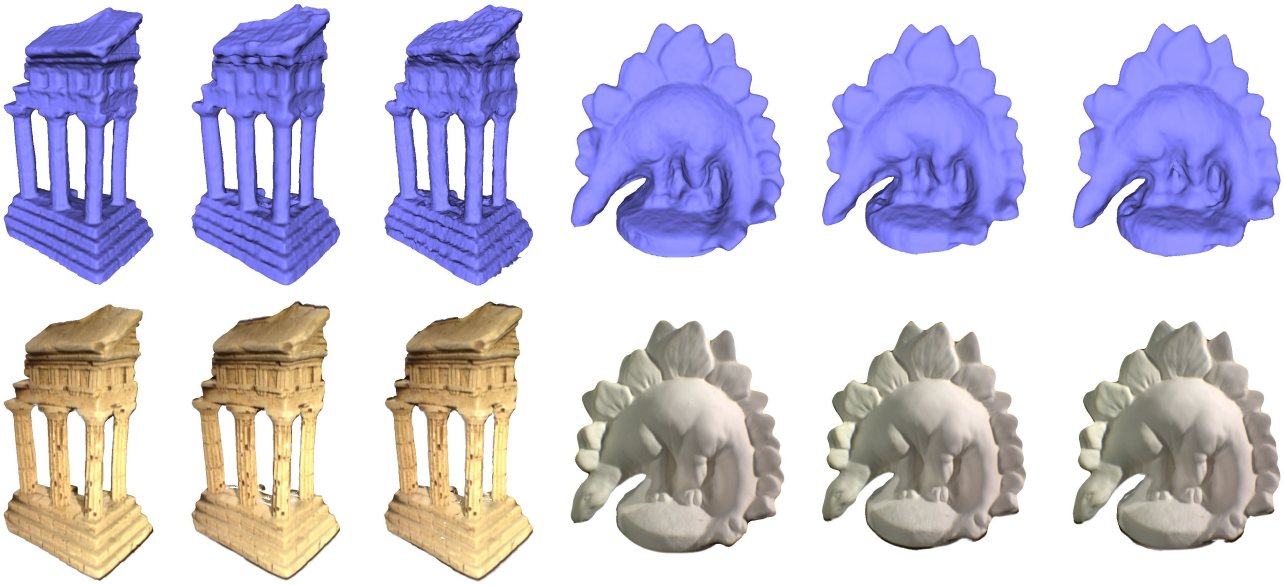


Fig. 18 Reconstruction results with Middlebury multi-view datasets.

Table 2 Quantitative evaluation of the accuracy and completeness of the reconstruction on Middlebury multi-view datasets; the superscript number indicates the rank of the method.

	Temple Full		Temple Ring		Temple Sparse		Dino Full		Dino Ring		Dino Sparse	
	Acc	Comp	Acc	Comp	Acc	Comp	Acc	Comp	Acc	Comp	Acc	Comp
Starck	N/A	N/A	N/A	N/A	1.27^{26}	87.7^{23}	N/A	N/A	N/A	N/A	1.01^{26}	90.7^{25}
Furukawa	0.49^{7}	99.6^{5}	0.47^{2}	99.6^{2}	0.63^{5}	99.3^{1}	0.33^{2}	99.8^{3}	0.28^{1}	99.8^{1}	0.37^{1}	99.2^{3}
Proposed-standard	0.43^{5}	99.0^{7}	0.71^{24}	97.6^{23}	0.86^{16}	96.2^{8}	0.35^{3}	100^{1}	0.58^{22}	99.5^{7}	0.68^{20}	98.0^{10}

9.3 Free-viewpoint video results

9.3.1 Qualitative evaluation

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 or to add value to the match analysis. Some examples of images from the free-viewpoint video sequences we generated for sports can be seen in Fig. 19; full sequences can be seen in the supplementary video. The background stadium models are manually generated using either images from the captured sequences (soccer sequence) or synthetic images (rugby sequence). High quality rendering from view-dependent geometry and images is made possible by the use of the proposed 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. In order to maximise the degree of realism, the scene is augmented with virtual shadows. A shadow mapping algorithm is used to simulate a virtual light source and cast soft shadows generated by players onto the ground.

9.3.2 Quantitative evaluation

A quantitative evaluation of free-viewpoint rendering accuracy was performed on the *soccer-eval* dataset which consists of 15 closely spaced cameras. One camera (shown in yellow in Fig. 10(e)) was excluded from reconstruction and used to provide ground truth. Three sets of reconstruction cameras were defined by excluding one, two or three of the neighbouring cameras (see Fig. 10), thus defining sets with increasing baseline. All 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, PSNR and appearance scores defined in (Kilner et al., 2009). Results are shown in Fig. 20. 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 techniques all score high on all evaluation criteria with the proposed hybrid method being consistently the top performer. Although scores are comparably high for these techniques, the visual quality is usually greater for

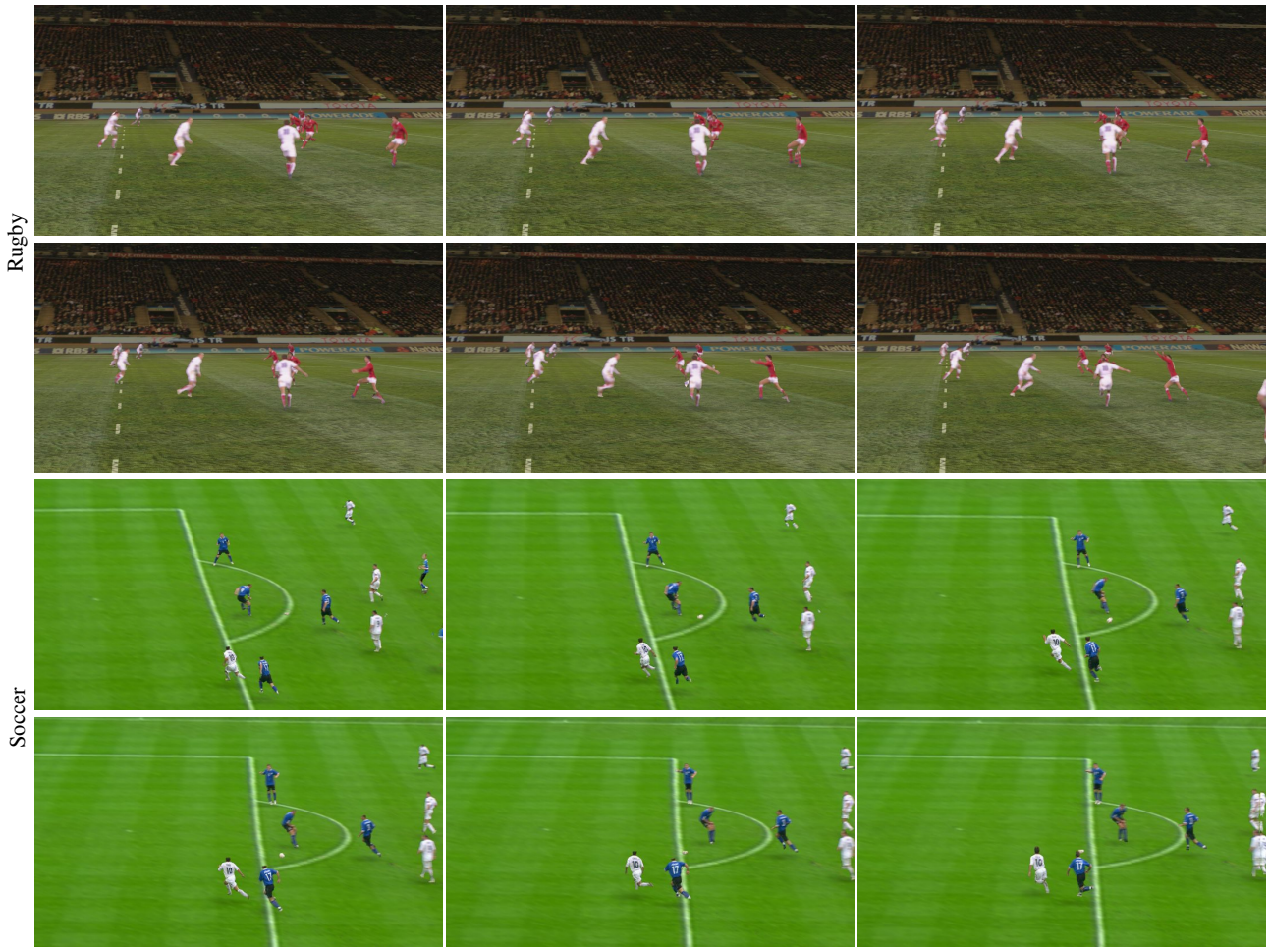


Fig. 19 Sample frames from the free-viewpoint video sequences generated using the proposed technique (see supplementary video for full sequences).

the proposed technique using the additional visual hull prior (Proposed-hybrid) closely followed by the proposed techniques using the iso-surface and standard depth priors (see supplementary 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. These visual quality improvements can be best seen in the accompanying video.

10 Conclusions and future work

This paper introduced a novel approach for simultaneous multi-layer segmentation and reconstruction of complex dynamic scenes captured under challenging outdoor or natural background indoor conditions. 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 through-the-lens camera calibration of large scale sports scenes. A general framework for iteratively enforcing multiple view consistency between the view-dependent reconstructions has been introduced, together with a method for robustly merging reconstructions from multiple views separated by a wide baseline when calibration accuracy is sufficiently high.

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

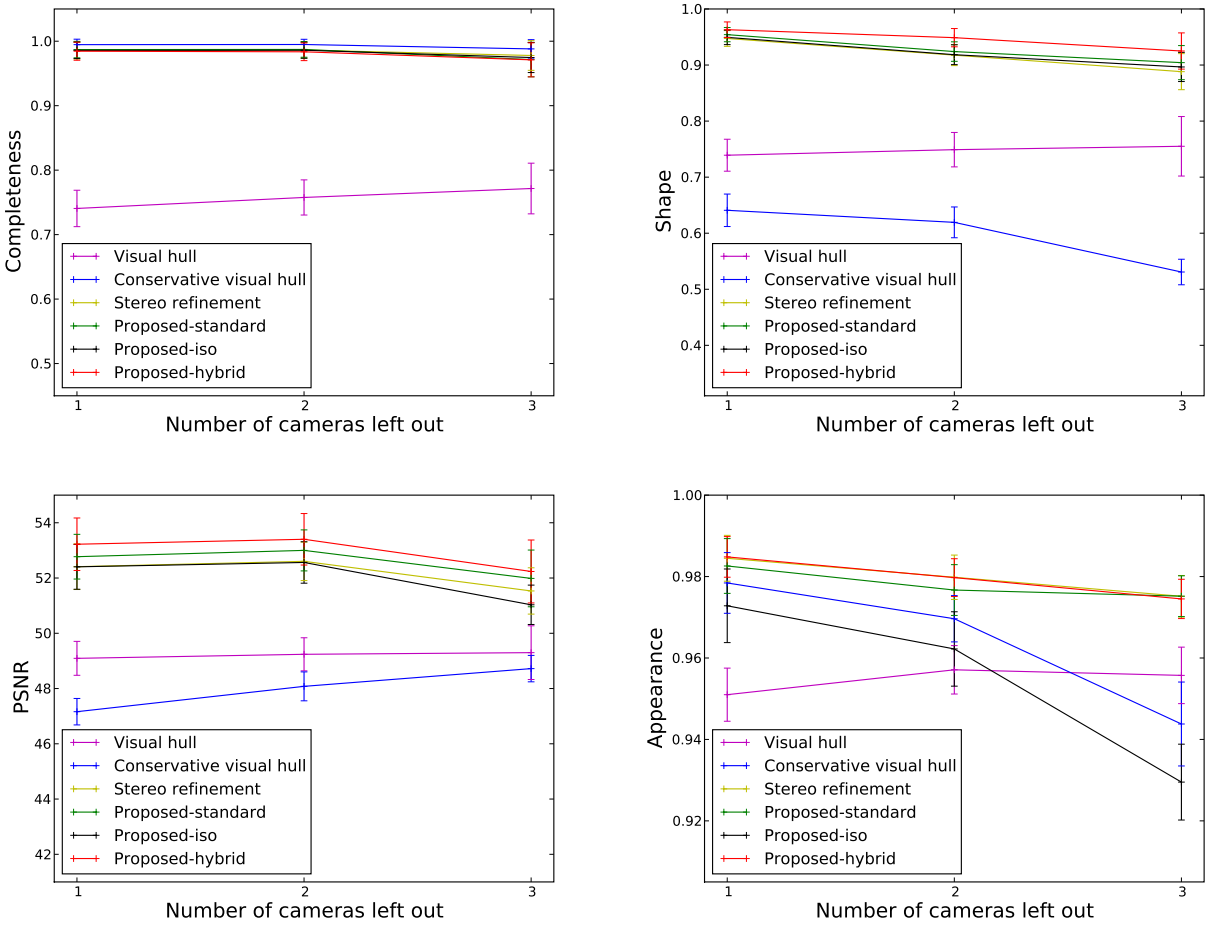


Fig. 20 Quantitative evaluation of free-viewpoint video performance.

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.

Experiments with a range of complex indoor sequences with natural background demonstrated the system's ability to achieve a level of accuracy similar to state-of-the-art algorithms such as (Furukawa and Ponce, 2010) without requirement for accurate input segmentation. This is a significant advantage as it allows fully automatic reconstructions of foreground scenes in the presence of cluttered natural backgrounds. Quantitative evaluation using the Middlebury multi-view datasets showed the method to be accurate, being ranked among the top 5 for two of the datasets. This demonstrates the ability of the technique to operate in a range of conditions from indoor environments to challenging outdoor environments without the need for manual intervention.

Acknowledgements This work was supported by the Technology Strategy Board and the Engineering and Physical Sciences Research Council 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). The Technology Strategy Board's role is to promote and support research into, and development and exploitation of, technology and innovation for the benefit of UK business, in order to increase economic growth and improve the quality of life. `www.innovateuk.org`. The authors would like to acknowledge the BBC for providing the sport data, the project partners at BBC R&D, The Foundry, Snell and Hawk-Eye Innovations for technical discussions, Daniel Scharstein for kindly performing the evaluation with the Middlebury datasets, and Martin Klaudiny and Cemre Zor for acting in the Ball and Dance2 sequences.

Appendix

We now give a proof that the energy defined in Eq. (21) satisfies the regularity condition required for graph-cut optimisation via α-expansion (Kolmogorov and Zabih, 2004). For that, it is sufficient to demonstrate that each pair-wise en-

ergy term (i.e. each contrast and smoothness term defined in Eq. (7) and Eq. (20)) is a metric (Boykov et al., 2001).

Each contrast term is independent of the depth labels. Additionally, the function C being constant for a given pixel pair, each term defines a Potts model with respect to the layer labels. This type of function is well known to be a metric.

Let us now show that each smoothness term defines a metric. Let us introduce a label d_∞ with finite depth $d_\infty \geq d_{\max} + \max_{\mathbf{p}} d_{\mathbf{p}}$. We can re-parametrise the energy by replacing the unknown label $\mathscr{U}$ by the label d_∞ and defining the equivalent smoothness term:

$$e_{\text{smooth}}(l_{\mathbf{p}}, d_{\mathbf{p}}, l_{\mathbf{q}}, d_{\mathbf{q}}) = \begin{cases} \min(|d_{\mathbf{p}} - d_{\mathbf{q}}|, d_{\max}) & \text{if } l_{\mathbf{p}} = l_{\mathbf{q}}, \\ d_{\max} & \text{otherwise.} \end{cases} \tag{26}$$

The function e_{smooth} is a metric if it satisfies:

$$e_{\text{smooth}}(l_\alpha, d_\alpha, l_\beta, d_\beta) = 0 \Leftrightarrow l_\alpha = l_\beta \text{ and } d_\alpha = d_\beta, \tag{27}$$

$$e_{\text{smooth}}(l_\alpha, d_\alpha, l_\beta, d_\beta) = e_{\text{smooth}}(l_\beta, d_\beta, l_\alpha, d_\alpha) \geq 0, \tag{28}$$

$$\begin{aligned} e_{\text{smooth}}(l_\alpha, d_\alpha, l_\beta, d_\beta) \leq\ & e_{\text{smooth}}(l_\alpha, d_\alpha, l_\gamma, d_\gamma) \\ & + e_{\text{smooth}}(l_\gamma, d_\gamma, l_\beta, d_\beta), \end{aligned} \tag{29}$$

for any labels (l_α, d_α), (l_β, d_β) and (l_γ, d_γ).

Clearly, e_{smooth} satisfies Eq. (27) and Eq. (28), so it remains to prove that the triangular inequality Eq. (29) is verified. Let us distinguish the following five cases:

$l_\alpha = l_\beta = l_\gamma$: In this case e_{smooth} is a truncated absolute distance which is well known to be a metric and satisfy Eq. (29).

$l_\alpha \neq l_\beta$, $l_\alpha \neq l_\gamma$, $l_\beta \neq l_\gamma$: In this case e_{smooth} is constant and Eq. (29) is trivially satisfied.

$l_\alpha = l_\gamma$, $l_\beta \neq l_\alpha$: In this case we have

$$e_{\text{smooth}}(l_\alpha, d_\alpha, l_\beta, d_\beta) = e_{\text{smooth}}(l_\gamma, d_\gamma, l_\beta, d_\beta) = d_{\max}, \tag{30}$$

and the triangular inequality is satisfied because by definition $e_{\text{smooth}}(l_\alpha, d_\alpha, l_\gamma, d_\gamma) \geq 0$.

$l_\beta = l_\gamma$, $l_\alpha \neq l_\beta$: This is equivalent to the previous case after permutation of α and β.

$l_\alpha = l_\beta$, $l_\gamma \neq l_\alpha$: In this case Eq. (29) is equivalent to $e_{\text{smooth}}(l_\alpha, d_\alpha, l_\gamma, d_\gamma) \leq 2d_{\max}$, which always holds. This completes the proof. □

References

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

Bai, X., Wang, J., Simons, D., and Sapiro, G. (2009). Video snapcut: robust video object cutout using localized classifiers. *ACM Trans. Graphics (Proc. SIGGRAPH) 28*(3).

Ballan, L., Brostow, G., Puwein, J., and Pollefeys, M. (2010). Unstructured video-based rendering: Interactive exploration of casually captured videos. *ACM Trans. Graphics (Proc. SIGGRAPH) 4*(29).

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

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

Bradley, D., Boubekeur, T., and Heidrich, W. (2008). Accurate multi-view reconstruction using robust binocular stereo and surface meshing. In *CVPR*.

Broadhurst, A., Drummond, T., and Cipolla, R. (2001). A probabilistic framework for the Space Carving algorithm. In *ICCV*, pp. 388–393.

Campbell, N., Vogiatzis, G., Hernández, C., and Cipolla, R. (2008). Using multiple hypotheses to improve depth-maps for multi-view stereo. In *ECCV*, Volume I, pp. 766–779.

Campbell, N., Vogiatzis, G., Hernández, C., and Cipolla, R. (2010). Automatic 3d object segmentation in multiple views using volumetric graph-cuts. *Image and Vision Computing 28*(1), 14–25.

Chen, S. and Williams, L. (1993). View interpolation for image synthesis. *ACM Trans. Graphics (Proc. SIGGRAPH)*, 279–288.

Chuang, Y.-Y., Agarwala, A., Curless, B., Salesin, D., and Szeliski, R. (2002). Video matting of complex scenes. *ACM Trans. Graphics (Proc. SIGGRAPH) 21*(3), 243–248.

Cohen, J., Lin, M., Manocha, D., and Ponamgi, M. (1995). I-collide: an interactive and exact collision detection system for large-scale environments. In *Proc. Symp. on Interactive 3D graphics*, pp. 189–218.

Connor, K. and Reid, I. (2003). A multiple view layered representation for dynamic novel view synthesis. In *BMVC*.

De Bonet, J. and Viola, P. (1999). Roxels: Responsibility weighted 3D volume reconstruction. In *ICCV*, Volume 1, pp. 418–425.

Debevec, P., Taylor, C., and Malik, J. (1996). Modeling and rendering architecture from photographs: a hybrid geometry- and image-based approach. In *ACM Trans. Graphics (Proc. SIGGRAPH)*, pp. 11–20.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological) 39*(1), 1–38.

Eisemann, M., De Decker, B., Magnor, M., Bekaert, P., de Aguiar, E., Ahmed, N., Theobalt, C., and Sellent, A. (2008). Floating textures. *Comput. Graphics Forum (Proc. Eurographics) 27*(2), 409–418.

Franco, J.-S. and Boyer, E. (2003). Exact polyhedral visual hulls. In *BMVC*, Volume 1, pp. 329–338.

Franco, J.-S. and Boyer, E. (2005). Fusion of multi-view silhouette cues using a space occupancy grid. In *ICCV*, pp. 1747–1753.

Furukawa, Y. and Ponce, J. (2010). Accurate, dense, and robust multiview stereopsis. *PAMI 32*(8), 1362–1376.

Germann, M., Hornung, A., Keiser, R., Ziegler, R., Würmlin, S., and Gross, M. (2010). Articulated billboards for video-based rendering. *Comput. Graphics Forum (Proc. Eurographics) 29*(2), 585–594.

Goesele, M., Curless, B., and Seitz, S. (2006). Multi-view stereo revisited. In *CVPR*, pp. 2402–2409.

Goldlücke, B. and Magnor, M. (2003). Joint 3D-reconstruction and background separation in multiple views using graph cuts. In *CVPR*, Volume 1, pp. 683–688.

Gortler, S., Grzeszczuk, R., Szeliski, R., and Cohen, M. (1996). The lumigraph. *ACM Trans. Graphics (Proc. SIGGRAPH)*, 43–54.

Grau, G., Thomas, G., Hilton, A., Kilner, J., and Starck, J. (2007). A robust free-viewpoint video system for sport scenes. In *3DTV*.

Grau, O., Prior-Jones, M., and Thomas, G. (2005). 3d modelling and rendering of studio and sport scenes for tv applications. In *WIAMIS*.

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

Guillemaut, J.-Y., Kilner, J., and Hilton, A. (2009). Robust graph-cut scene segmentation and reconstruction for free-viewpoint video of complex dynamic scenes. In *ICCV*, pp. 809–816.

Habbecke, M. and Kobbelt, L. (2007). A surface-growing approach to multi-view stereo reconstruction. In *CVPR*.

Hernández Esteban, C. and Schmitt, F. (2004). Silhouette and stereo fusion for 3d object modeling. *Comput. Vis. Image Underst. 96*(3), 367–392.

Inamoto, N. and Saito, H. (2007). Virtual viewpoint replay for a soccer match by view interpolation from multiple cameras. *IEEE Trans. on Multimedia 9*(6), 1155–1166.

Kang, S., Li, Y., Tong, X., and Shum, H.-Y. (2006). Image-based rendering. *Foundations and Trends in Computer Graphics and Vision 2*(3), 173–258.

Kazhdan, M., Bolitho, M., and Hoppe, H. (2006). Poisson surface reconstruction. In *Symp. on Geometry Processing*, pp. 61–70.

Kilner, J., Starck, J., Guillemaut, J.-Y., and Hilton, A. (2009). Objective quality assessment in free-viewpoint video production. *SPIC 24*, 3–16.

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

Kimura, K. and Saito, H. (2005). Player viewpoint video synthesis using multiple cameras. In *CVMP*, pp. 112–121.

Kohli, P. and Torr, P. (2007). Dynamic graph cuts for efficient inference in Markov random fields. *PAMI 29*(12), 2079–2088.

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

Kolmogorov, V. and Zabih, R. (2002). Multi-camera scene reconstruction via graph cuts. In *ECCV*, Volume III, pp. 82–96.

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

Kutulakos, K. (2000). Approximate N-view stereo. In *ECCV*, Volume I, pp. 67–83.

Kutulakos, K. and Seitz, S. (2000). A theory of shape by space carving. *Int. J. Comput. Vision 38*(3), 199–218.

Laurentini, A. (1994). The visual hull concept for silhouette-based image understanding. *PAMI 16*(2), 150–162.

Levin, A., Lischinski, D., and Weiss, Y. (2008). A closed-form solution to natural image matting. *PAMI 30*(2), 228–242.

Levoy, M. and Hanrahan, P. (1996). Light field rendering. *ACM Trans. Graphics (Proc. SIGGRAPH)*, 31–42.

Li, Y., Sun, J., and Shum, H.-Y. (2005). Video object cut and paste. *ACM Trans. Graphics (Proc. SIGGRAPH) 24*(3), 595–600.

Liu, Y., Cao, X., Dai, Q., and Xu, W. (2009). Continuous depth estimation for multi-view stereo. In *CVPR*, pp. 2121–2128.

Matusik, W., Buehler, C., Raskar, R., Gortler, S. J., and McMillan, L. (2000). Image-based visual hulls. In *ACM Trans. Graphics (Proc. SIGGRAPH)*, pp. 369–374.

Matusik, W., Pfister, H., Ngan, A., Beardsley, P., Ziegler, R., and McMillan, L. (2002). Image-based 3d photography using opacity hulls. *ACM Trans. Graph. (Proc. SIGGRAPH) 21*(3), 427–437.

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

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

Mitchelson, J. and Hilton, A. (2003). Wand-based multiple-camera studio calibration. Technical Report VSSP-TR-2/2003, Centre for Vision, Speech and Signal Processing, University of Surrey. `http://www.ee.surrey.ac.uk/CVSSP/VMRG/Publications/mitchelson03cvssp-tr.pdf`.

Moezzi, S., Tai, L.-C., and Gerard, P. (1997). Virtual view generation for 3d digital video. *IEEE MultiMedia 4*(1),

18–26.

Narayanan, P., Rander, P., and Kanade, T. (1998). Constructing virtual worlds using dense stereo. In *ICCV*, pp. 3–10.

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

Rother, C., Kolmogorov, V., and Blake, A. (2004). "GrabCut" – interactive foreground extraction using iterated graph cuts. In *ACM Trans. Graphics (Proc. SIGGRAPH)*, pp. 309–314.

Roy, S. and Cox, I. (1998). A maximum-flow formulation of the N-camera stereo correspondence problem. In *ICCV*, pp. 492–499.

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

Seitz, S. and Dyer, C. (1996). View morphing. *ACM Trans. Graph. (Proc. SIGGRAPH)*, 21–30.

Seitz, S. and Dyer, C. (1999). Photorealistic scene reconstruction by voxel coloring. *IJCV 35*(2), 151–173.

Sinha, S. and Pollefeys, M. (2005). Multi-view reconstruction using photo-consistency and exact silhouette constraints: A maximum-flow formulation. In *ICCV*, Volume 1, pp. 349–356.

Slabaugh, G., Culbertson, B., Malzbender, T., and Schafer, R. (2001). A survey of methods for volumetric scene reconstruction from photographs. In *International Workshop on Volume Graphics*.

Snow, D., Viola, P., and Zabih, R. (2000). Exact voxel occupancy with graph cuts. In *CVPR*, Volume 1, pp. 345–352.

Starck, J. and Hilton, A. (2005). Virtual view synthesis of people from multiple view video sequences. *Graphical Models 67*(6), 600–620.

Starck, J. and Hilton, A. (2007). Surface capture for performance based animation. *IEEE Computer Graphics and Applications 27*(3), 21–31.

Sun, J., Zhang, W., Tang, X., and Shum, H.-Y. (2006). Background cut. In *ECCV*, Volume 3954, pp. 628–641.

Szeliski, R. and Golland, P. (1999). Stereo matching with transparency and matting. *IJCV 32*(1), 45–61.

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

Thomas, G. (2007). Real-time camera tracking using sports pitch markings. *J. Real-Time Image Proc. 2*, 117–132.

Vlasic, D., Peers, P., Baran, I., Debevec, P., Popović, J., Rusinkiewicz, S., and Matusik, W. (2009). Dynamic shape capture using multi-view photometric stereo. *ACM Trans. Graphics (Proc. SIGGRAPH Asia) 28*(5).

Vogiatzis, G., Hernandez, C., Torr, P., and Cipolla, R. (2007). Multiview stereo via volumetric graph-cuts and occlusion robust photo-consistency. *PAMI 29*(12), 2241–2246.

Waschbüsch, M., Würmlin, S., and Gross, M. (2007). 3d video billboard clouds. *Comput. Graphics Forum (Proc. Eurographics) 26*(3), 561–569.

Zhang, Z. (2000). A flexible new technique for camera calibration. *PAMI 22*(11), 1330–1334.

Zitnick, C., Kang, S., Uyttendaele, M., Winder, S., and Szeliski, R. (2004). High-quality video view interpolation using a layered representation. In *ACM Trans. Graphics (Proc. SIGGRAPH)*, pp. 600–608.