Outdoor Dynamic 3D Scene Reconstruction

Hansung Kim, Jean-Yves Guillemaut, *Member, IEEE*, Takeshi Takai, Muhammad Sarim
and Adrian Hilton, *Member, IEEE*

***Abstract*—Existing systems for 3D reconstruction from multiple view video use controlled indoor environments with uniform illumination and backgrounds to allow accurate segmentation of dynamic foreground objects. In this paper we present a portable system for 3D reconstruction of dynamic outdoor scenes which require relatively large capture volumes with complex backgrounds and non-uniform illumination. This is motivated by the demand for 3D reconstruction of natural outdoor scenes to support film and broadcast production. Limitations of existing multiple view 3D reconstruction techniques for use in outdoor scenes are identified. Outdoor 3D scene reconstruction is performed in three stages: (1) 3D background scene modelling using spherical stereo image capture; (2) multiple view segmentation of dynamic foreground objects by simultaneous video matting across multiple views; and (3) robust 3D foreground reconstruction and multiple view segmentation refinement in the presence of segmentation and calibration errors. Evaluation is performed on several outdoor productions with complex dynamic scenes including people and animals. Results demonstrate that the proposed approach overcomes limitations of previous indoor multiple view reconstruction approaches enabling high-quality free-viewpoint rendering and 3D reference models for production.**

***Index Terms*—3D Reconstruction, Computer vision, Free-viewpoint video, Image-based reconstruction, Image-based rendering.**

I. Introduction

KANADE et al. [1] pioneered the use of multiple view video acquisition for 3D modelling and rendering with the Virtualized Reality system using a 51 camera dome. Subsequently many approaches have been proposed for surface reconstruction and free-viewpoint rendering from multiple view video acquisition in controlled indoor studio environments [2]–[6]. State-of-the-art methods for multiple view surface reconstruction from images under controlled studio conditions have achieved a level of accuracy comparable to surface measurement using active sensors [7]. The concept of using multiple cameras for 3D production opens up the potential for the 3D Virtual Studio in which dynamic shape and appearance of an actor's performance can be captured as a 3D computer graphics (CG) model. This has attracted considerable interest as a production tool in film, broadcast and games [8]–[10] as it allows the combination of real and computer generated elements in 3D. Currently elements of many films are produced by capturing live-action in a multi-camera studio with controlled uniform illumination and uniform chroma-key blue or green backgrounds to facilitate foreground segmentation. Live action elements are then rendered with computer-generated or captured background scenes. The controlled indoor studio conditions enable accurate multiple view segmentation of the foreground actor performance allowing subsequent 3D reconstruction.

Location based outdoor shooting remains an important part of film production to capture scenes which cannot be filmed in a studio or live action coverage together with the environment. Synchronisation of background and foreground actions is a significant advantage of outdoor capture. In this paper we present a system for 3D live action capture using multiple camera systems in complex environments and illumination. This presents a challenging problem as backgrounds are often complex natural scenes which have overlap in colour distribution with the foreground scene.

The proposed system builds on state-of-the-art multiple camera studio systems introducing robust techniques to enable 3D reconstruction of actor performance in natural scenes. 3D background scenes are reconstructed from multiple spherical image pairs. Foreground action is separated from the background scene by the introduction of a multiple view video matting technique.This builds on previous work in single view video matting for natural scenes which are widely used in film production, introducing geometric and appearance constraints between views which allow foreground segmentation across multiple views with the same level of interaction as single view video matting. Techniques for multiple view reconstruction of 3D foreground models are then employed to overcome small errors in camera calibration and segmentation which may occur in natural scenes. This pipeline of 3D background modelling, dynamic foreground segmentation and 3D foreground reconstruction allows capture of live action in uncontrolled outdoor environments.

Evaluation is presented for 3D reconstruction of live action in several uncontrolled outdoor scenes with multiple people and animals. Comparison of 3D foreground reconstruction with state-of-the-art multiple view reconstruction techniques for indoor scenes demonstrates significant improvement in reconstruction quality.

A. Requirements for Outdoor Scene Capture

In this section we analyse the specific requirements for 3D reconstruction of dynamic outdoor scenes from multiple view video. These requirements are contrasted with techniques for indoor scene reconstruction to identify the limitations of existing approaches. An analysis of the design of multiple camera systems for indoor studio capture was previously presented in [11] providing recommendations for the number of cameras and their configuration. In contrast outdoor

This work was supported by the the EU IST FP7 project i3Dpost and TSB/EPSRC project i3Dlive.

H. Kim, J.-Y. Guillemaut, T. Takai, M. Sarim and A. Hilton are with the Centre for Vision, Speech, and Signal Processing, University of Surrey, Guildford, Surrey, U.K. (e-mail: h.kim@surrey.ac.uk; j.guillemaut@surrey.ac.uk; t.takai@surrey.ac.uk; msarim@fuuast.edu.pk; a.hilton@surrey.ac.uk).

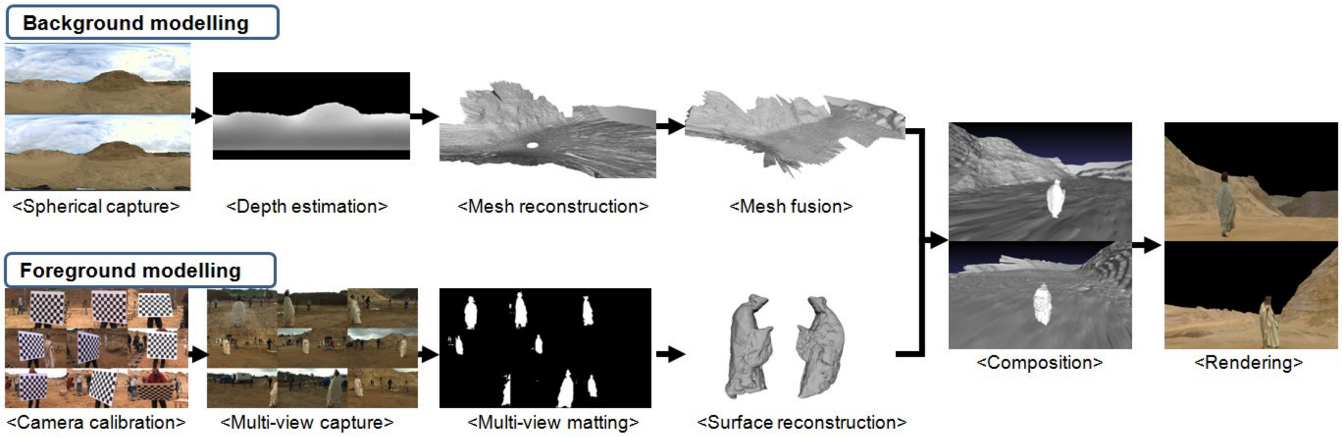


Fig. 1. Data capture and processing chain for the outdoor dynamic scene reconstruction illustrated on the Queen dataset

scene capture for film and broadcast production requires the following considerations:

- **Large capture volume:** Indoor capture of actor performance typically have a capture area of less than $5m^2$ with $2m$ height. Outdoor capture of live action commonly requires a relatively large area $> 10m^2$ or in the case of stadium sports such as soccer up to $100m \times 50m$.
- **Natural scene backgrounds:** In general outdoor capture will have natural image backgrounds which may be dynamic and do not provide a high-contrast to the foreground action. This requires simultaneous 3D background scene modelling and introduction of methods for multiple view segmentation which do not rely on controlled chroma-key backgrounds.
- **Uncontrolled illumination:** Outdoor scene illumination changes according to time-of-day and weather resulting in dynamic illumination which may also be strongly directional causing strong shadows, shading and specularities. Only limited control is offered by supplementary illumination onset.
- **Portable capture equipment:** Setup and reconfiguration of multiple camera equipment needs to be rapid to accommodate production requirements for principal photography. The requirement for camera cables for power, synchronisation and data transfer significantly increases setup time and reduces flexibility.
- **Camera calibration errors:** Calibration accuracy is reduced by large capture volumes, non-ideal camera mounting and incidental movement of cameras. In practice access onset to perform camera calibration may be limited.
- **Fast scene motion:** Live action including animals may include faster movements than typically encountered in a studio environment resulting in motion blur and large changes in shape and appearance between successive video frames.

In this paper we present and evaluate a system for outdoor scene capture which aims to address the above problems.

B. Overview of outdoor 3D scene capture

A portable system is introduced for 3D capture and modelling of live action in outdoor scenes as illustrated in Figure 1. The system comprises pipelines for 3D background capture and modelling of natural scenes from multiple pairs of spherical stereo images and for 3D foreground live action capture, segmentation and reconstruction from multiple view video. The 3D foreground and background modelling are composited to allow free-viewpoint rendering of the complete 3D scene. The contributions of this system for outdoor scene capture over previous multiple view studio reconstruction are as follows:

1) **Portable capture system:** Live action capture is performed using a wireless multiple camera system capturing high-quality synchronised HD video, which can be setup and reconfigured in minutes.
2) **3D background modelling of natural scenes:** Multiple high-resolution spherical stereo image pairs are captured to reconstruct a full 3D background model of natural scenes with high-resolution appearance as well as shape sufficient for rendering background appearance from arbitrary viewpoints. This overcomes limitations of previous LIDAR(Light Detection And Ranging)-based 3D scene scans which do not capture appearance information for rendering.
3) **Multiple view foreground segmentation in natural scenes:** A novel approach is introduced for simultaneous multiple view segmentation in natural scenes. This exploits the consistency of foreground appearance between views together with multi-view geometric constraints to allow segmentation of multiple view video with the level of manual interaction required for single view video.
4) **Robust 3D foreground reconstruction:** To overcome problems of errors in camera calibration, natural scene segmentation and motion blur we employ a view-dependent multiple view reconstruction framework [12]. This requires an initial coarse reconstruction which is derived from the multiple view video segmentation. The initial multi-view segmentation and reconstruction are jointly optimised integrating information from multiple views to recover a refined scene reconstruction for high-

quality rendering. Multiple view-dependent reconstructions can also be combined to obtain a single closed-surface model.

Evaluation is performed on several challenging outdoor scenes with complex natural background, uncontrolled directional illumination, highly dynamic motion of people and animals and large capture volumes. Comparison is performed with other state-of-the-art multiple view reconstruction approaches. This evaluation demonstrates that the proposed system achieves full 3D reconstruction in outdoor scenes allowing high-quality free-viewpoint rendering and overcomes limitations of previous reconstruction approaches developed for indoor studio environments.

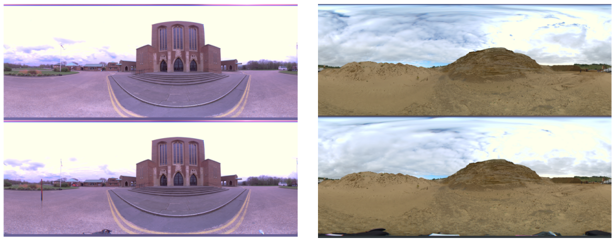
II. Related Work

There has been a large amount of research on image-based 3D reconstruction focused on two important problems: real-time reconstruction for interactive applications [2], [8]; and free-viewpoint rendering without loss of visual quality [3], [4], [13]. A review of multiple view studio design is presented in [11]. Application of multi-view studio production techniques to outdoor scenes requires additional factors to be taken into consideration as identified in Section I.

Initial transfer of studio technology has focused on stadium sports applications which present a semi-controlled environment with relatively uniform backgrounds and fixed pitch markings providing fiducials for calibration of moving cameras. The Virtualized Reality^{TM} technology [1] was used in the EyeVision system to produce virtual camera sweeps as action replays for Super Bowl XXXV in 2001. This system employed more than thirty cameras on motorised heads slaved to a single manually operated camera to capture the action of interest. Virtual fly-around effects were then produced by switching between the real camera views without interpolation of intermediate views resulting in visible jumps. The Matrix movie franchise popularised the use of such effects in film production through the *bullet time* effect which used dense rigs of hundreds of cameras with a narrow baseline spacing ($< 3°$) to give a smooth fly-around effect by switching views.

For full 3D reconstruction to enable view interpolation in sports, Ohta et al. [14] proposed a simplified geometric representation of soccer players using planar billboards for real-time transmission and view synthesis. Guillemaut et al. [12] proposed a segmentation and reconstruction techniques tolerant to errors in calibration and segmentation, and reconstructed free-viewpoint video in soccer and rugby matches from multiple moving cameras. An alternative approach to full 3D reconstruction was proposed by Germann et al. [15] who approximated 3D shape of the player's body articulated textured billboards attached to the skeleton structure. This required manual interaction to pose the skeleton for free-viewpoint video rendering. This technology was commercialised by Liberovision to allow synthesis of virtual transitions between broadcast cameras for soccer[1].

General outdoor scenes are challenging due to the variable lighting and non-uniform backgrounds. They may also impose more restrictions on the camera setup. Halser et al. [16] and Shaheen et al. [17] reconstructed outdoor actions without environment or illumination constraints using multiple portable cameras by tracking a skeleton-based human model. Ballan et al. [18] proposed a non-photorealistic view-transition method for multi-view videos of complex scenes using rough foreground and background modelling with billboard and camera tracking using structure-from-motion (SfM) techniques.



(a) Cathedral (b) Quarry

Fig. 2. Spherical stereo pairs captured (top and bottom)

In this paper we present a system to reconstruct general live-action in natural scene without prior knowledge of scene structure.

III. Portable Capture System

A. Environment capture system

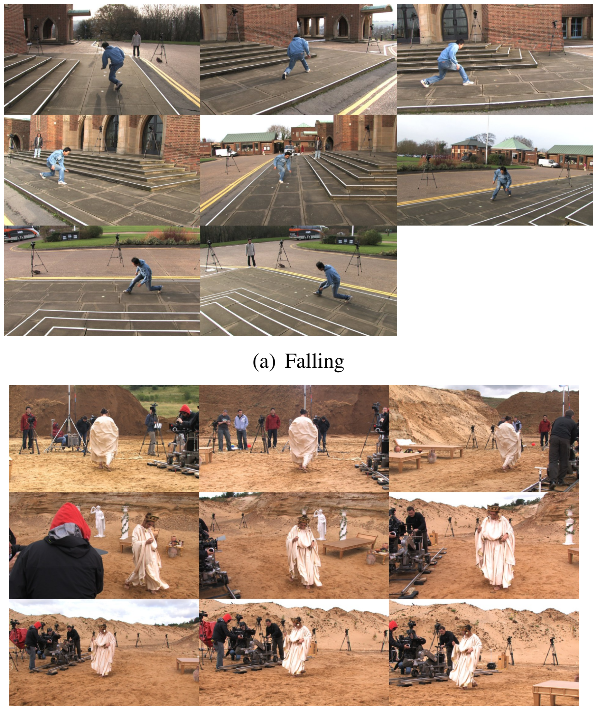
Static background capture and reconstruction is performed using an off-the-shelf line-scan camera, Spheron[2], with a fisheye lens. A full spherical view is generated by mosaicing rays from a vertical slit at the centre of a rotating lens. The camera rotates rotates about an axis passing through its optical centre. The imaging geometry of the line-scan capture can be modelled as a conventional perspective projection because all the rays in the spherical image intersect at a single 3D point. We attach a Nikon 16mm f/2.8 AF fisheye lens to the system and capture images with a maximum resolution of 12574×5658. The scene is captured with the camera at two different heights to recover depth information of the scene through stereo geometry. Figure 2 shows stereo image pairs captured with a vertical baseline of 60cm resulting in a maximum disparity of 240 pixels.

There are several advantages of using a line-scan camera to acquire spherical images for environment modelling. First, we can acquire high resolution details of the background scene with a high-dynamic range. Second, the stereo matching can be simplified to a 1D search along the scan line in the image if the two capture points are vertically aligned. Error in the alignment can be corrected by rectification as proposed by Banno and Ikeuchi [19]. Finally, a relatively simple calibration is required for depth reconstruction and registration.

If pixels on the vertical line are evenly mapped to the [0, π] range, the disparity d between projection points $p_t(x_t, y_t)$ and $p_b(x_b, y_b)$ of a 3D point P is defined as the difference of the vertical angles of the projected points θ_t and θ_b as in Eq. (1). The distance r_t between the top camera and the 3D point P can be calculate as in Eq. (2) using only angular disparity

[1] LiberoVision http://www.liberovision.com

[2] Spheron http://www.spheron.com/en/intruvision/solutions/spherocam-hdr.html



(a) Falling

(b) Queen

Fig. 3. Examples of multi-view outdoor capture

and baseline distance B by triangulation.

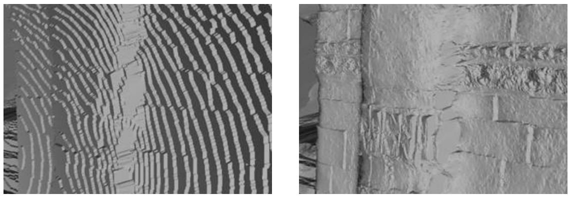
$$d(p_t) = \theta_t - \theta_b = (y_t - y_b) \times \pi / image\_height \quad (1)$$

$$r_t(p_t) = B / \left(\frac{\sin \theta_t}{\tan(\theta_t + d(p_t))} - \cos \theta_t \right) \quad (2)$$

Reconstructed geometries are registered and merged using feature matching between views, so that only the baseline distance information is required in outdoor capture. Uniform mapping of projected pixels into the [0, π] range can be easily performed with a 1D lookup table because the lens distortion is modelled by pre-calibrated fixed internal parameters.

B. Dynamic scene capture system

The dynamic foreground scene is captured with a portable multiple HD camera system. The camera system comprises HDV camcorders, Canon XH G1s[3], which have f=4.5–90 mm, F/1.6–3.5 lens and three 1/3-in CCDs providing uncompressed HD-SDI at 1920×1080 resolution, with 4:2:2 chroma subsampling. We have two options for camera synchronisation and video recording according to the capture environment. If power is available and cabling is possible, all cameras are synchronised by an external genlock reference synchronisation signal and uncompressed images are transferred to an external disk recorder, DVS Xway system[4], using HD-SDI with embedded time-code. In outdoor settings where power is not available and cabling between cameras is undesirable, all camera timecodes are synchronised to a master camera in advance. Cameras are then operated in a free-running mode using their internal clocks for synchronisation and video is captured to tape in compressed MPEG2 format. This allows multiple cameras to be used without cables whilst maintaining synchronisation enabling rapid setup and reconfiguration with minimal impact onset. Synchronised multi-view image sequences can be extracted from the tapes with stored time-code for processing off-line. Frame drift induced by using internal time-code is approximately 1 frame per hour in our experiments when cameras are free-running, which is acceptable in most applications. Since the cameras and the genlock synchronisation signal generator can be driven by batteries, this system is fully portable and can produce synchronised multi-view image sequences of actions in any outdoor environment. Figure 3 shows examples of the captured multi-view images in outdoor environments.



(a) Integer disparity (b) Continuous disparity

Fig. 4. Comparison of surface reconstructions

Camera calibration is an essential process for 3D reconstruction from multiple view video. The intrinsic parameters of cameras are estimated using a checker board. Extrinsic parameters are estimated by wand-based calibration using bundle adjustment between observed locations of coloured markers. Wand calibration of camera array takes 1-2 minutes onset. For moving cameras, such as the principal film camera, through-the-lens calibration can be employed to register the moving camera with the multiple static witness cameras and estimate the intrinsic parameters [20], [21].

IV. Environment Modelling

A. 3D reconstruction from a spherical stereo image pair

As mentioned in Section III-A, 3D information can be extracted from 2D images if we know camera parameters and point correspondences between views. A number of studies have been reported on the stereo correspondence problem over the past three decades [22]. Most current disparity estimation algorithms solve the correspondence problem on a discrete domain such as integer, half- or quarter-pixel levels. This results in quantisation error and is not sufficient to recover a smooth surface for environment modelling because: (i) the fisheye lens has wide field-of-view (FOV) with a relatively small disparity range; (ii) small disparity change may produce large depth error because of the distance from the camera to the background scene; (iii) the captured image has a large radial distortion. Variational approaches can be a solution for this quantisation problem because they solve the correspondence problem in a continuous domain. The difference between discrete and continuous disparity fields in environment reconstructions is shown in Fig. 4. The quantisation error causes step-like artefacts in the result from integer disparity while variational methods produce smooth surfaces with fine details.

[3] Canon http://www.canon.co.uk/For_Home/Product_Finder/Camcorders/professional/XH_G1s/

[4] Pronto3 http://www.dvs.de/products/video-systems/pronto3.html

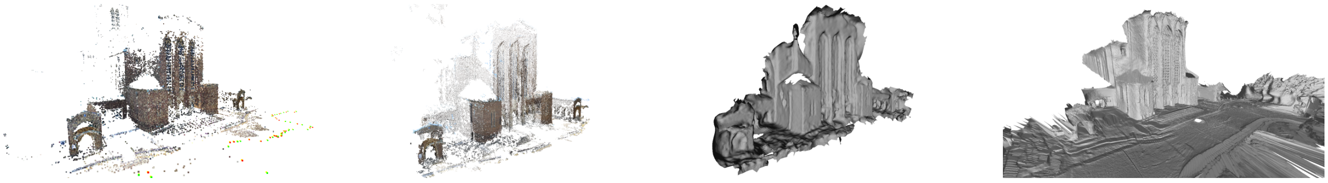

(a) Bundler (b) Bundler+PMVS (c) PMVS+Poisson (d) Proposed method

Fig. 5. Environment modelling from multiple images of Cathedral

Partial differential equation (PDE)-based variational methods provide a continuous approach to solve the energy minimisation problem of Eq. (3) involving a data term and a smoothing term as an equivalent nonlinear diffusion equation with an additional reaction term of Eq. (4) [23]. In Eq.(3) and (4), λ is a weight for the data term, $I(p)$ is a pixel value of the point p, d is a 2D disparity vector, $\nabla := (\partial x, \partial y)^T$ denotes a spatial gradient operator, $\Psi(\cdot)$ is a potential function for diffusion filtering and $g(\cdot)$ is a diffusion tensor.

$$\begin{aligned} E(d_t) &= \lambda \int_\Omega (I_1(p) - I_2(p+d_1))^2 \, dx \\ &\quad + \int_\Omega \Psi(\nabla d_1, \nabla I_1) dx \qquad (3) \\ \frac{\partial d}{\partial t} &= div(g(\nabla I_1, \nabla d_1)\nabla d_1) \\ &\quad + \lambda(I_1(p) - I_2(p+d_1)) \frac{\partial I_2(p+d_1)}{\partial d} \qquad (4) \end{aligned}$$

In designing the diffusion tensor, we adopt an anisotropic image/disparity driven method, Eq. (5), which produces continuous depth fields with sharp object boundaries even in occluded and highly textured regions. In Eq. (5), L denotes the identity matrix and the term $\nabla I \nabla I^T$ is the structure tensor of Nagel and Enkelmann's method [24] for anisotropic diffusion filtering. $s(\nabla d)$ is a monotonically increasing function which converges to 1 for ε =0.37 and controls image gradients in highly textured regions. Further details can be found in [25].

$$g(\nabla I, \nabla d) = f(\nabla I, s(\nabla d))(\nabla I \nabla I^T + L) \qquad (5)$$

$$f(\nabla I, s(\nabla d)) = \frac{1}{1 + s(\nabla d \cdot |\nabla I|^2)^2}$$

$$s(\nabla d) = -ln(\varepsilon + (1-\varepsilon) \cdot e^{-|\nabla d|})$$

In solving the PDE, we use a hierarchical structure which starts from low resolution images and recursively refines the result at higher levels in order to reduce the computation time and avoid local minima.

Relying on a pure image-based approach can be risky in real outdoor scene reconstruction. Stereo matching does not guarantee perfect correspondence because real scenes include non-Lambertian surfaces (glass, water) resulting in different specular reflections on the surface in the stereo image pair. Reflection or transparency of windows result in false depth for the glass. Wide featureless regions or repetitive textures also make the matching algorithm converge to local minima. Therefore, we allow manual user interaction on the result to correct such errors in real production. Small erroneous regions are manually marked as occluded and we set the weighting term λ in Eq. (4) to zero so that only smoothing is performed for the regions in disparity estimation. If the erroneous regions are large and planar, they can be approximated into planes using the RANSAC(Random Sample Consensus)-based plane fitting algorithm [26] in mesh reconstruction.

B. Mesh fusion for stereo reconstruction at multiple locations

Reconstruction from a spherical stereo image pair provides a good environment model from the captured location, but there is no way to get information about occluded regions behind any object from a single input image. In order to overcome this problem, we capture spherical stereo image pairs of the background scene from multiple locations and merge partial reconstructions from each view point into a common 3D scene structure using mesh fusion.

Mesh fusion consists of two steps: mesh registration and reliable surface extraction. First, all meshes are registered into a common coordinate system using the iterative closest point (ICP) algorithm [27]. To automate the initialisation of the ICP registration, we use SURF (Speeded Up Robust Feature) matching [28] between captured images for different stereo pairs. The resulting matches are used as reference points for 3D matching by projecting them into 3D space with the estimated depth field. However, these points are not sufficiently reliable to be used as references for the ICP algorithms because two possible sources of error exist: one from SURF matching error between image pairs because of radial distortion and the other from depth estimation error. Therefore we use a RANSAC algorithm [29] to calculate an optimised 3D rigid transform between two meshes excluding outliers. Finally, a complete 3D model is generated as a single mesh by selecting the most reliable observations among overlapping surface measurements considering surface visibility, orientation and distance from the camera.

This approach is advantageous for environment modelling compared to other Multi-View Stereo (MVS) and SfM-based methods using narrow field-of-view cameras because a dense structured mesh is reconstructed from a small number of narrow-baseline stereo pairs without preceding calibration steps. Another advantage of the reliable surface extraction approach is preservation of surface detail from stereo reconstruction while other mesh fusion methods such as Poisson reconstruction [30] or range image merging [31] may incur loss of details on the original surface because these algorithms generate a combined surface from data in overlapping regions.

Figure 5 shows a comparison of the reconstruction results with other MVS and SfM-based methods: Bundler [32], Patch-based Multi-view Stereo (PMVS) [5], and Poisson reconstruction [30]. The comparative methods were applied to 92 photos with a resolution of 2272×1704, and the mesh fusion is generated from 3 pairs of spherical stereo images with a

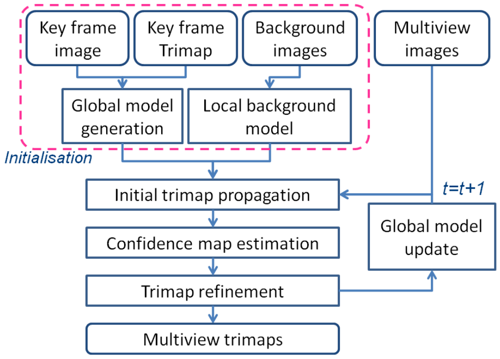


Fig. 6. Trimap label propagation for wide-baseline video matting

resolution of 6284×2794. We can see that the proposed mesh fusion method generates much denser points over a wider range from a smaller image capture dataset.

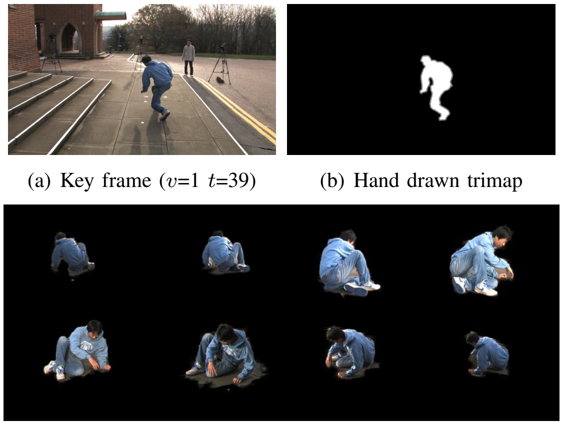
V. Dynamic Foreground Scene Reconstruction

A. Multiple view video matting

State-of-the-art natural image matting algorithms provide accurate silhouettes against changing backgrounds and camera noise but require labour-intensive trimaps to be defined for key-frame images in a video sequence. This is prohibitively expensive for application to multiple view image sequences. In this work we extend video matting to multiple views by propagating key-frame foreground information input on a single view. This approach exploits the similarity in foreground appearance between views [33]. Trimaps for multiple view videos are constructed by spatial propagation between views using the epipolar constraint and temporal propagation over time of high confidence trimap labels using a Bayesian inference framework from a sparse set of key frame trimaps for a single view. This techniques allows application of natural image matting across multiple views from a small number of manually defined key-frame trimaps T^k in a single view (typically 1-2 trimaps in one view for 200 frames×8 cameras). This approach ensures that multi-view matting is no more labour intensive than single view matting commonly used in post-production.

An overview of the multi-view matting framework is presented in Fig. 6. The statistics of foreground and background scene appearance are represented by four models: multi-view global foreground model $M^{GF}(t)$, multi-view global background model $M^{GB}(t)$, single-view per-pixel local foreground model $M^{LF}(p,v,t)$, and single-view per-pixel local background model $M^{LB}(p,v,t)$ for view v at each pixel p at time t. Gaussian Mixture Models (GMM) are used to represent the statistics of colour appearance for both the global and local models. All models are dynamically updated using estimated foreground and background label confidence to incorporate new observations.

Global appearance models $M^{GF}(t)$ and $M^{GB}(t)$ are initialised from the known key-frame trimap pixel labels using the mean shift clustering algorithm [34]. Initial trimap labels are propagated into other views by maximum a posterior (MAP) estimates. In each propagated image, a confidence map of the initial labels is constructed from the maximum likelihood foreground and background model component over both local and global models. The likelihood is measured by minimum squared Mahalanobis distance. Finally the local per-pixel foreground model $M^{LF}(p,v,t)$ from the pixels in a local neighbourhood with foreground trimap labels is constructed based on the confidence map in order to fill holes and refine boundaries. Generated trimaps are used as input to Levin's natural image matting algorithm [35] and the global models are updated with the results before moving to the next frames.


(a) Key frame (v=1 t=39)

(b) Hand drawn trimap

(c) Matting results (8 views, cropped, t=81)

Fig. 7. Multiview matting results

An additional advantage of using this trimap propagation algorithm within our system is that camera calibration information is available for propagation. Using the calibration information, global model propagation can be restricted to the corresponding epipolar line. To take into account calibration errors, we dilate the epipolar lines by a few pixels. Using this epipolar constraint, we can increase both processing speed and accuracy in the trimap propagation. If the result shows significant errors, additional key-frames can be easily added to prevent error propagation.

Figure 7 shows the cropped segmentation results of the Falling sequence in Fig. 3. In practice, only one key-frame trimap was used for this sequence of 8 views × 100 frames, i.e., 800 frames.

B. Surface reconstruction

The first stage of the reconstruction process consists in extracting an approximate representation of the dynamic foreground object using shape-from-silhouette [36]. This technique computes a 3D model by back-projecting the foreground silhouettes into 3D space and intersecting the set of visual cones obtained. The reconstructed model is referred to as the visual-hull and is guaranteed to contain the foreground surface assuming perfect input calibration and segmentation.

The visual-hull is then refined by the view-dependent depth estimation using additional visual cues from multiple cameras. Traditionally reconstruction is performed sequentially using the segmentation input to infer depth information. This works well in situations where high accuracy calibration and segmentation can be achieved but can result in errors in outdoor scenes where calibration and segmentation are less accurate. In order to increase robustness to these errors, we use the technique described in [12] to jointly refine the segmentation and estimate depth in a view-dependent manner for each input camera. This was shown to reduce propagation of errors between the two stages and reduce ambiguities by simultaneously using all available cues.

Joint segmentation and depth estimation defines a labelling problem where we seek the mappings $l : \mathcal{P} \rightarrow \mathcal{L}$ and $d : \mathcal{P} \rightarrow \mathcal{D}$, respectively assigning a layer label l_p and a depth label d_p to every pixel p in a given image, where $\mathcal{P}$ denotes the set of pixels in the reference image and $\mathcal{L}$ and $\mathcal{D}$ are discrete sets of labels representing the different layer and depth hypotheses. $\mathcal{L}$ consists of a single background layer and multiple foreground layers defined by the visual hull connected components representing different people in the scene. The set of depth labels $\mathcal{D}$ is formed of depth values d_i obtained by discretising the 3D space together with an unknown label $\mathcal{U}$ accounting for occlusions. The visual hull is used to initialise the layered depth estimation process and restrict the set of possible depth labels to its interior thus considerably reducing the number of possible labels.

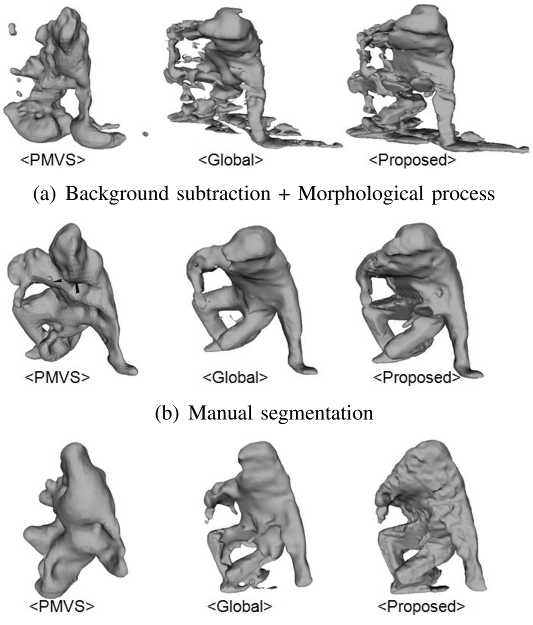
Computation of the optimum labelling (l, d) is formulated as an energy minimisation problem of the cost function

$$E(l,d) = w_{\text{colour}}E_{\text{colour}}(l) + w_{\text{contrast}}E_{\text{contrast}}(l) + w_{\text{match}}E_{\text{match}}(d) + w_{\text{smooth}}E_{\text{smooth}}(l,d), \quad (6)$$

where the energy terms correspond to various cues derived from layer colour models, contrast, photo-consistency and smoothness priors and whose relative contributions are controlled by the parameters w_{colour}, w_{contrast}, w_{match} and w_{smooth}. Detailed description of each energy term can be found in [12].

Optimisation of the energy defined in Eq. (6) is NP-hard, however an approximate solution with strong optimality properties can be computed using the α-expansion algorithm based on graph-cuts [37], [38]. To improve multi-view consistency, optimisation is performed in several iterations, using depth maps from neighbouring views as depth priors.

The final stage of the dynamic model generation consists in fusing the view-dependent depth maps into a mesh representation. To merge the different representations, each layered depth representation is converted into an oriented point cloud with one oriented point per pixel. Fusion is then performed using Poisson surface reconstruction [30] producing a watertight mesh representation. The visual hull is used to provide additional oriented points in unconstrained occluded areas preventing the formation of protrusions in these areas.

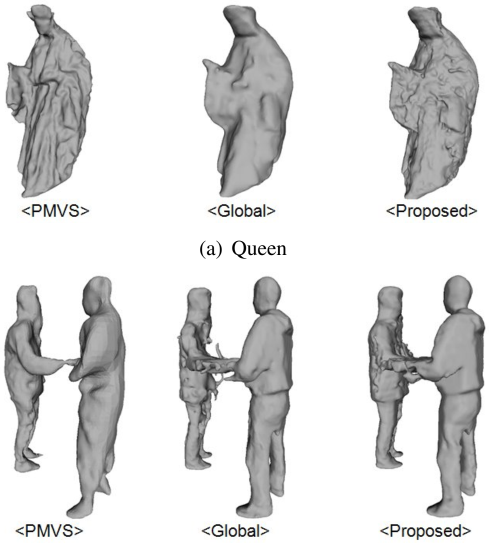


(a) Background subtraction + Morphological process

(b) Manual segmentation

(c) Multiview matting

Fig. 8. Reconstruction of Falling scene according to segmentation methods (Left: PMVS [5], Middle: Global optimisation [4], Right: Proposed method)



(a) Queen

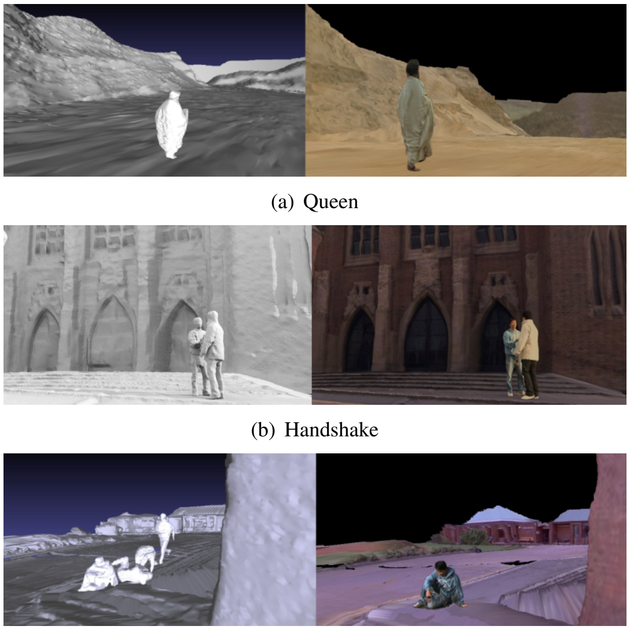
(b) Handshake

Fig. 9. Reconstruction with the proposed matting for other sequences (Left: PMVS, Middle: Global optimisation, Right: Proposed method)

C. Performance evaluation

We compared the reconstruction performance of the view-dependent reconstruction with global optimisation [4] and PMVS [5] methods in outdoor scenes. Figure 8 shows the reconstruction results of the Falling scene with different segmentation techniques. The Falling scene is a difficult case because of directional lighting, few features, self-occlusion and fast movement. Background subtraction is fast but produces errors in reconstruction due to strong shadow and false negative segmentation errors. Reconstruction with ground-truth manual segmentation shows good results. The global optimisation shows the best performance and PMVS is still poor because of self-occlusion of the scene. However, this segmentation requires 30 mins of manual drawing step per frame for 8 camera views, which is unrealistic in real applications. The proposed multiple view matting also produced good results while the segmentation is automatically generated from one hand-drawn trimap in different frame (Fig. 7(b)). This shows that the proposed trimap propagation makes a significant contribution to practical reconstruction in outdoor scenes within a realistic timeline with minimal user interaction..

Figure 9 shows reconstruction results for other sets with the multi-view matting. The Queen sequence (Fig. 9(a)) has relatively good lighting conditions (due to uniform ambient illumination on a cloudy day), good segmentation and slow



(a) Queen

(b) Handshake

(c) Falling

Fig. 10. Model composition

action. There are also plenty of features for stereo correspondence from creases on the clothing. All methods show good reconstructions but PMVS in particular shows more clothing detail for this example. The Handshake scene (Fig. 9(b)) has strong directional lighting from the sun and large occlusions between the two actors. The actors wear unpatterned creaseless clothes. PMVS fails in point cloud reconstruction on such smooth regions and results in bumpy surfaces. The global optimisation method produces a good surface for the uniform regions but shows serious errors in occluded regions where large holes appear. The proposed view-dependent approach gives more accurate reconstruction of all surface regions compared to either PMVS or global reconstruction.

Overall the multi-view matting and view-dependent reconstruction approach has been found to give more reliable reconstruction with minimal user interaction for outdoor scene captures than the other state-of-the-art approaches evaluated.

VI. Model composition and rendering

The reconstructed background and foreground models are merged into a common 3D coordinate system with the same scale. Coordinate alignment is achieved by setting the origins of foreground coordinates as the location of the line-scan camera in calibration, because both models are constructed at real world scale.

In texture mapping, we use different methods for background and foreground models as the background has a single texture map whereas view-dependent texturing of the dynamic foreground gives improved visual detail. The mesh grid of the background model is generated from the disparity map which has the same coordinates as the original images. Therefore, we directly map textures from the corresponding patches in the original image to the mesh. We use UV mapping of 3D points onto the image texture.

Dynamic foreground models reconstructed from multiple cameras provide only partial texture information for the model. Due to occlusion and changes in surface appearance caused by viewing angles, care must be taken to select the appropriate camera to use as a texture for each mesh face. We adopt a view-dependent texture mapping technique to assign camera images to each face with the best visibility [39]. Figure 10 shows results of composition and texture mapping.

VII. Experimental outdoor production

The proposed system is tested for outdoor production in three different scenarios: “Cathedral”, “Horse-riding” and “Queen”. All scenes were captured with the portable camera system and recorded to HDV tapes. The background scenes were captured by the line-scan camera with resolution of 6284×2794 and models were reconstructed from 1-4 pairs of spherical stereo images. Table I gives a summary of the setup and characteristics of the scenes and Fig. 11 shows the camera set up for each scene. Details of capture environments, challenging points, production results and discussions for each test production are presented in the following subsections. Free-viewpoint rendering results with dynamic virtual camera views for full sequences and the final short film of “The Midas Touch” using the proposed system can be seen in the supplementary videos available at:

http://kahlan.eps.surrey.ac.uk/hkim/tcsvt/

Low resolution versions of the videos are also available at:

http://youtu.be/C4ViJoF8oVY
http://youtu.be/zXeR6zNN59w

A. Cathedral

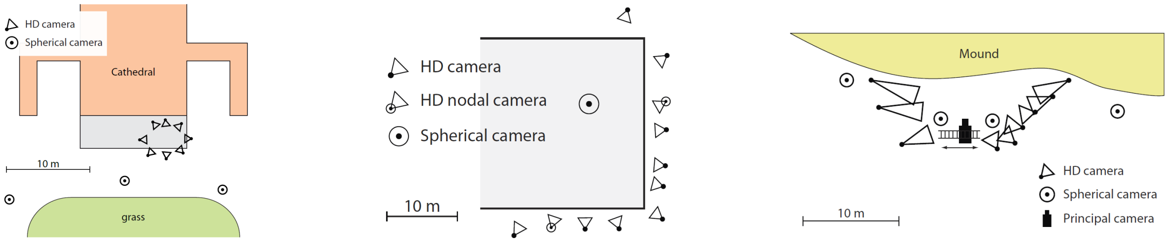
The Cathedral scene was taken in front of a cathedral as illustrated in Fig. 2(a). The background scene was captured in three locations and two actions were captured with eight surrounding portable cameras. The captured images were already shown in Fig. 3(a). The main background building has a complex structure with self-occlusions and complicated details such as sculptures. The weather was sunny, so the scene has strong shadow of the main objects and captured images have significant variation in brightness for different viewing directions. Portable multiple view capture system setup, configuration, time code synchronisation and calibration for an 8 camera system took less than 30 minutes.

One of the most serious problems in background modelling is reflection or transparency of windows. The scenes reflected on glass come from farther away than the real position of the windows and result in false depth. Therefore we manually corrected the initial disparity values for window reflections and lens flares are marked as occluded as suggested in section IV-A such that the data term has no influence in these regions. The reconstructed background model is shown in Fig. 5(d).

Two actions were filmed with the portable camera system. The first action is “Falling” with fast movement and dynamic shape changes, and the second is “Handshake” with two actors causing significant inter-actor occlusion. The view-dependent

TABLE I
CAPTURE SET UP

Scene	# of spherical capture	# of portable cameras	Coverage	Capture volume	Length (frames)	Lighting	Motion
Cathedral	3	8	360°	Small	100 (each)	Directional	Fast, occlusion
Horse-riding	1	11	180°	Large	150	Ambient	Fast
Queen	4	9	180°	Medium	300	Additional lighting	Slow



(a) Cathedral (b) Horse-riding (c) Queen

Fig. 11. Camera set up for outdoor production

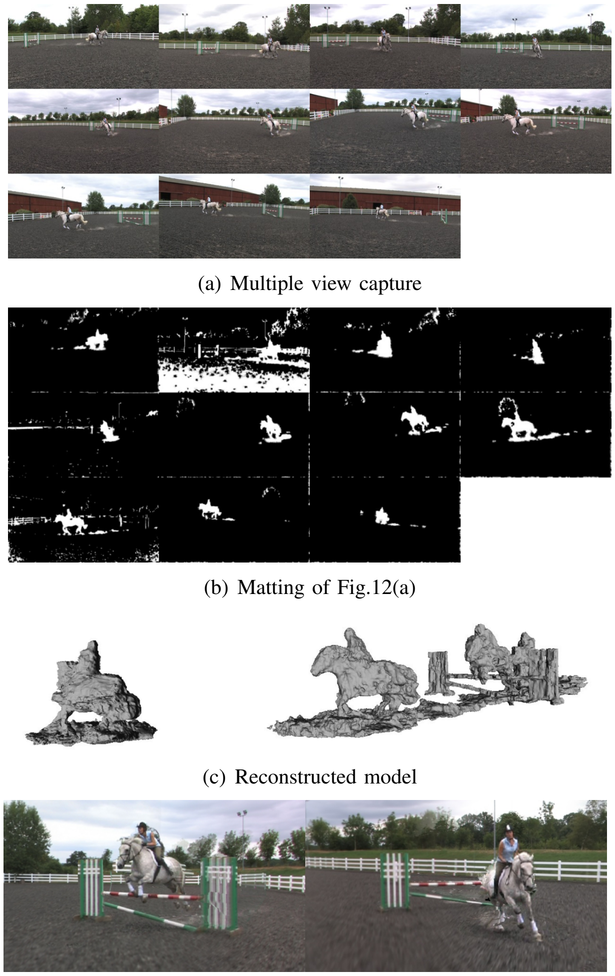
surface reconstruction algorithm produced relatively good reconstruction for both sequences of sufficient accuracy for high-quality free-viewpoint rendering. Results of model reconstruction and free-viewpoint rendering are shown in Fig. 9-10.

B. Horse-riding

Horse-riding scene was captured at an arena with an 11 camera system. The arena is a wide open area surrounded by a fence, so the background scene was captured with a spherical image pair at a single location for reconstruction. In order to cover a wide capture volume, we set 9 fixed cameras and 2 nodal cameras on three sides around one end of the arena fence. Camera parameters for the nodal cameras were estimated by feature matching with other fixed cameras and applying a perspective-2-point (P2P) solver [21]. The weather was cloudy so illumination was relatively uniform.

In background modelling, all structures outside the fence are too distant for accurate reconstruction. Therefore we approximate the background model with a rectangular cuboid by fitting the reconstructed mesh to the position of the fence.

In foreground reconstruction, the hurdle and horse-riding are reconstructed separately and merged together at the compositing stage. Figure 12 shows captured images, matting, reconstruction and free-viewpoint rendering results. To estimate foreground mattes for nodal cameras, the local background per pixel appearance model is approximated with an image mosaic from the captured video sequence. This results in larger matting errors for the nodal cameras due to the higher variance between the true and estimated local background than for static views. Results also include errors in parts of the scene due to dynamic dust from the ground which is segmented as foreground. Segmentation errors are reduced with the proposed view-dependent reconstruction.



(a) Multiple view capture

(b) Matting of Fig.12(a)

(c) Reconstructed model

(d) Free-viewpoint rendering

Fig. 12. Horse-riding results

Free-viewpoint rendering results achieve a high visual quality for the horse and rider despite the relatively wide camera framing and challenging environment. Due to segmentation errors some artefacts occur at the boundaries of horse and rider for viewpoints which are distant from the cameras. This could be corrected with additional manual interaction in the video matting. The nodal cameras allow acquisition with increased resolution for rendering by correctly framing the foreground object of interest, but as noted earlier give higher errors in segmentation and reconstruction due to inaccuracies in the background appearance model and calibration.

TABLE II
PROCESSING TIME FOR QUEEN (FOR 10 SEC, 250 FRAMES × 9 VIEWS)

Background reconstruction		Foreground reconstruction	
Step	Time (mins)	Step	Time (mins)
Spherical capture	4	Setup & calibration	20
Depth estimation	15	capture	1
Mesh reconstruction	1	Data transfer	10
Mesh Fusion	20	Initial trimap	4
		Matting	120
		Reconstruction	200
Composition		5	
Rendering		30	
Total		390	

C. Queen

The Queen sequence is a part of the final experimental film production "The Midas Touch" of EU project i3DPost[5]. The proposed system was used to support production of visual effects for the film by a production company BUF[6]. The production is divided into two capture sessions: studio capture using a multi-camera rig to create digital assets and the on-set capture of principal photography making use of the portable multi-camera system to create structured 3D representations of actor performance at an outdoor sand quarry location for post-production tasks. In the studio capture, several actors were captured to create animated character models using the studio system proposed in [11].

On-set capture was performed in a sand quarry with relatively uniform background appearance. The background was captured at 4 locations with the line-scan camera and main actions filmed with one principal RED camera[7] and 9 witness cameras. Examples of the captured scene were shown in Fig. 2(b) and Fig. 3(b). Additional lighting was used to get high quality textures. Results of model reconstruction and free-viewpoint rendering are presented in Fig. 9-10.

Table II shows processing time for a 10s sequence, i.e., 2250 frames. Background reconstruction can be performed in parallel with foreground reconstruction, and set up and trimap initialisation do not depend on number of frames. It takes roughly 6 and a half hours from the capture to the final rendering. Surface reconstruction is a bottleneck, but this was performed in parallel with 10 processors because it is a frame-independent process.

The reconstructed foreground and background models have sufficient quality in their geometry and texture resolution to be used as a 3D reference for visual effects production. Figure 13 shows snapshots of the final short film. All extras in the scene were captured in the studio and registered into the scene as animated character models. The 3D registration was performed by fitting the footage of the principal camera into the reconstructed background model.

The palace model was captured from a separate location using the line-scan camera and reconstructed by the proposed background modelling technique. The lower part of the palace model was directly used for the production and the upper part was manually corrected to remove distortion in the columns using the reconstructed model as a geometric reference. The principal actress Queen reconstruction was used as a dynamic model to transform the Queen to gold.

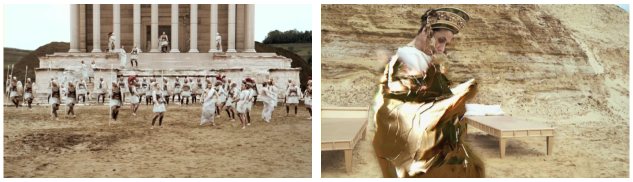

(a) Digital assets registration (b) Visual effect

Fig. 13. Snapshots of the experimental production "The Midas Touch"

Use of the proposed spherical imaging and multiple camera system to capture 3D reference data on set reduced the time required for CG modelling by an order of magnitude compared to conventional film production techniques by providing a 3D reference of both the scene and actor performance. Multiple camera studio capture provides digital doubles of actors suitable for direct use as secondary assets in production.

VIII. DESIGN OF OUTDOOR CAPTURE SYSTEMS

Outdoor 3D capture of dynamic live action scenes presents a number of practical and algorithmic challenges as identified in section I. The system introduced in this paper aims to address these challenges allowing outdoor capture of complex 3D dynamic scenes. Evaluation on natural outdoor scenes demonstrates that the approach achieves reconstruction of dynamic 3D scene models which allow high-quality free-viewpoint rendering. The development and evaluation of this system has identified important factors in the design of systems for outdoor scene capture. A number of the problems in outdoor scene capture have been addressed with the proposed system. Further development is required to fully address production requirements:

- **Portable capture:** The wireless multiple camera capture system developed allows flexible and rapid setup for dynamic scene capture. Two significant bottlenecks remain for efficient and reliable use: (i) transfer of video data from the camera-to-computer for processing is currently a time-consuming offline process; (ii) verification of the camera capture volume for live action reconstruction and calibration currently requires offline data transfer and processing such that problems are only identified offset when it is too late to correct. Video-rate data transfer, calibration and coarse reconstruction are required to verify capture quality onset.
- **Large capture volume:** Multiple view capture is performed using fixed cameras with wide framing to ensure coverage of the capture volume for reconstruction. This results in a lower image resolution for the foreground action resulting in a reduction in reconstruction accuracy. Ideally moving or nodal (pan-tilt-zoom) cameras would be employed to capture the foreground action with higher resolution. Currently this requires high-accuracy encoders or through-the-lens camera tracking which leads to increased calibration errors and issues of reliability for multiple cameras in natural scenes. A practical solution with static cameras is to increase the number of cameras and frame the required capture volume as tightly as possible

[5] i3DPost. http://www.i3dpost.eu/

[6] http://www.buf.com

[7] http://www.red.com

ensuring sufficient camera resolution and coverage in all areas for accurate reconstruction.

- **Natural backgrounds and illumination:** The approach presented for 3D background modelling and multiple view segmentation currently assumes static backgrounds. In outdoor scenes background motion may occur either locally such as trees blowing in the wind or globally due to objects moving in the background scene or clouds moving across the sun. This leads to errors in the 3D background model which require manual correction through additional key-frames. Robust methods to model and segment outdoor scenes with dynamic backgrounds are required in further work. In practice, minimising the overlap of actor appearance from that of the background reduces errors in segmentation.
- **Foreground scene reconstruction:** The view-dependent approach for joint segmentation and reconstruction refinement from multiple views is robust to increased errors in segmentation and calibration which occur in outdoor scenes. In practice due to the increased calibration error and wider camera framing improved results are obtained if the inter-camera baseline is narrower than for studio reconstruction ($< 30^\circ$). This together with the increased capture volume requires an increased number of cameras for outdoor capture $12 - 16$.

IX. Conclusion

A portable system for capture and 3D reconstruction of dynamic outdoor scenes has been introduced and evaluated. The system is designed to address the problems of outdoor reconstruction in natural scenes, with uncontrolled illumination which require relatively large capture volumes. The system comprises a pipeline for 3D background reconstruction from multiple high-resolution spherical stereo image pairs and a pipeline for foreground live action reconstruction from multiple wide-baseline video cameras. Background 3D scene reconstruction is performed using a PDE-based disparity estimation scheme which gives continuous sub-pixel estimates avoiding quantisation errors for reconstruction over large distances. Fusion of background scene reconstructions from multiple stereo pairs gives a complete 3D scene model with high-resolution image appearance for rendering novel views.

Live action 3D foreground reconstruction is performed in two stages: initial multiple view foreground segmentation; and view-dependent joint segmentation and reconstruction. To enable efficient foreground segmentation a multiple view segmentation algorithm is introduced which exploits the common foreground appearance and visual geometry between camera views to jointly segment multiple views with the level of user-input required for single view video segmentation. The initial foreground scene segmentation enables multiple view reconstruction to be performed in complex outdoor scenes. A view-dependent approach for joint refinement of multi-view reconstruction and segmentation is used to achieve robust reconstruction in the presence of errors in the initial foreground segmentation and camera calibration. View-dependent reconstruction can then be fused to obtain a watertight foreground model if required. Free-viewpoint rendering is performed by compositing the foreground reconstruction with the 3D background scene model.

Evaluation is performed on several challenging natural scenes with multiple people and animals. Results demonstrate high-quality free-viewpoint rendering providing 3D information to support production. Comparison with state-of-the-art approaches previously introduced for reconstruction of controlled indoor studio scenes demonstrates improved robustness in outdoor scenes provided by the 3D background model, joint multiple view segmentation, and view-dependent reconstruction.

The paper identifies the problems and design considerations in developing system for outdoor scene capture. A number of open-problems remain for future research including use of moving cameras to increase capture resolution, robust approaches for dynamic backgrounds and video-rate calibration together with coarse reconstruction to verify camera calibration and coverage onset. Portable multiple camera video-based capture and image-based scene modelling enabling the reconstruction of natural outdoor scenes has the potential to provide 3D reference information for use in media production.

References

[1] T. Kanade, P. W. Rander, and P. J. Narayanan, “Virtualized reality: Constructing virtual worlds from real scenes,” *IEEE Multimedia*, vol. 4, no. 1, pp. 34–47, 1997.

[2] T. Matsuyama, X. Wu, T. Takai, and S. Nobuhara, “Real-time 3d shape reconstruction, dynamic 3d mesh deformation, and high fidelity visualization for 3d video,” *Computer Vision and Image Understanding*, vol. 96, no. 3, pp. 393–434, 2004.

[3] C. L. Zitnick, S. B. Kang, M. Uyttendaele, S. Winder, and R. Szeliski, “High-quality video view interpolation using a layered representation,” in *Proc. ACM SIGGRAPH*, 2004, pp. 600–608.

[4] J. Starck and A. Hilton, “Surface capture for performance-based animation,” *IEEE Computer Graphics and Applications*, vol. 27, no. 3, pp. 21–31, 2007.

[5] Y. Furukawa and J. Ponce, “Accurate, dense, and robust multiview stereopsis,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 32, no. 8, pp. 1362–1376, 2010.

[6] E. Aguiar, C. Stoll, C. Theobalt, N. Ahmed, H. Seidel, and S. Thrun, “Performance capture from sparse multi-view video,” in *Proc. ACM SIGGRAPH*, 2008, pp. 1–10.

[7] S. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski, “A comparison and evaluation of multi-view stereo reconstruction algorithms,” in *Proc. CVPR*, 2006, pp. 519–528.

[8] M. Price, J. Chandaria, O. Grau, G. Thomas, D. Chatting, J. Thorne, G. Milnthorpe, P. Woodward, L. Bull, E.-J. Ong, A. Hilton, J. Mitchelson, and J. Starck, “Real-time production and delivery of 3d media,” in *Proc. International Broadcasting Convention*, 2002, pp. 348–356.

[9] O. Grau and G. Thomas, “3d image sequence acquisition for tv & film production,” in *Proc. 3DPVT*, 2002, pp. 320–326.

[10] P. Hillman, J. Lewis, S. Sylwan, and E. Winquist, “Issues in adapting research algorithms to -stereoscopic visual effects,” in *Proc. ICIP*, 2010, pp. 17–20.

[11] J. Starck, A. Maki, S. Nobuhara, A. Hilton, and T. Matsuyama, “The multiple-camera 3-d production studio,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 19, no. 6, pp. 856–869, 2009.

[12] J.-Y. Guillemaut and A. Hilton, “Joint multi-layer segmentation and reconstruction for free-viewpoint video applications,” *International Journal of Computer Vision*, vol. 93, no. 1, pp. 73–100, 2011.

[13] M. Eisemann, B. de Decker, M. Magnor, P. Bekaert, E. de Aguiar, N. Ahmed, C. Theobalt, and A. Sellent, “Floating textures,” *Computer Graphics Forum (Proc. Eurographics)*, vol. 27, no. 2, pp. 409–418, 2008.

[14] Y. Ohta, I. Kitahara, Y. Kameda, H. Ishikawa, and T. Koyama, “Live 3d video in soccer stadium,” *International Journal of Computer Vision*, vol. 75, no. 1, pp. 173–187, 2007.

[15] M. Germann, A. Hornung, R. Keiser, R. Ziegler, S. Würmlin, and M. Gross, "Articulated billboards for video-based rendering," *Computer Graphics Forum (Proc. Eurographics)*, vol. 29, no. 2, pp. 585–594, 2010.

[16] N. Hasler, B. Rosenhahn, T. Thormaehlen, M. Wand, and H. Seidel, "Markerless motion capture with unsynchronized moving cameras," in *Proc. CVPR*, 2009, pp. 224–231.

[17] M. Shaheen, J. Gall, R. Strzodka, L. Gool, and H. Seidel, "A comparison of 3d model-based tracking approaches for human motion capture in uncontrolled environments," in *Proc. WACV*, 2009, pp. 1–8.

[18] L. Ballan, G. J. Brostow, J. Puwein, and M. Pollefeys, "Unstructured video-based rendering: Interactive exploration of casually captured videos," in *Proc. SIGGRAPH*, 2010, pp. 1–11.

[19] A. Banno and K. Ikeuchi, "Omnidirectional texturing based on robust 3d registration through euclidean reconstruction from two spherical images," *Computer Vision and Image Understanding*, vol. 114, no. 4, pp. 491–499, 2010.

[20] Z. Zhang, "Flexible camera calibration by viewing a plane from unknown orientations," in *Proc. ICCV*, 1999, pp. 666–673.

[21] E. Imre, J.-Y. Guillemaut, , and A. Hilton, "Calibration of nodal and free-moving cameras in dynamic scenes for post-production," in *Proc. 3DIMPVT*, 2011.

[22] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *International Journal of Computer Vision*, vol. 47, no. 1, pp. 7–42, 2002.

[23] J. Weickert, "A review of nonlinear diffusion filtering," *Lecture Notes in Computer Science*, vol. 1252, pp. 3–28, 1997.

[24] H. Nagel and W. Enkelmann, "An investigation of smoothness constraints for the estimation of displacements vector fields from image sequences," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 8, pp. 565–593, 1986.

[25] H. Kim and A. Hilton, "3d modelling of static environments using multiple spherical stereo," in *Proc. RMLE workshop in ECCV*, 2010.

[26] O. Gallo, R. Manduchi, and A. Rafii, "Cc-ransac: Fitting planes in the presence of multiple surfaces in range data," *Pattern Recognition Letters*, vol. 32, pp. 403–410, 2010.

[27] P. Besl and N. McKay, "A method for registration of 3-d shapes," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 14, no. 2, pp. 239–256, 1992.

[28] H. Bay, A. Ess, T. Tuytelaars, and L. Gool, "Surf: Speeded up robust features," *Computer Vision and Image Understanding*, vol. 110, pp. 346–359, 2008.

[29] M. Fischler and R. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Communication of the ACM*, vol. 24, pp. 381–395, 1982.

[30] M. Kazhdan, M. Bolitho, and H. Hoppe, "Poisson surface reconstruction," in *Proc. SGP*, 2006, pp. 61–70.

[31] B. Curless and M. Levoy, "A volumetric method for building complex models from range images," in *Proc. SIGGRAPH*, 1996, pp. 303–312.

[32] N. Snavely, S. Seitz, and R. Szeliski, "Modeling the world from internet photo collections," *International Journal of Computer Vision*, vol. 80, no. 2, pp. 189–210, 2008.

[33] M. Sarim, A. Hilton, J.-Y. Guillemaut, T. Takai, and H. Kim, "Natural image matting for multiple wide-baseline views," in *Proc. ICIP*, 2010, pp. 2233–2236.

[34] D. Comaniciu and P. Meer, "Mean shift: a robust approach toward feature space analysis," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, pp. 603–619, 2002.

[35] A. Levin, D. Lischinski, and Y. Weiss, "A closed form solution to natural image matting," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 30, no. 2, pp. 228–242, 2008.

[36] A. Laurentini, "The visual hull concept for silhouette-based image understanding," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 16, no. 2, pp. 150–162, 1994.

[37] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 23, no. 11, pp. 1222–1239, 2001.

[38] Y. Boykov and V. Kolmogorov, "An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 26, no. 9, pp. 1124–1137, 2004.

[39] P. Debevec, Y. Yu, and G. Borshukov, "Efficient view- dependent image-based rendering with projective texturemapping," in *Proc. Eurographics Rendering Workshop*, 1998, pp. 105–116.

Hansung Kim received the BS degree in radio communication engineering in 1998, and the MS and Ph.D degrees in electronic and electrical engineering from Yonsei University, Korea, in 2001 and 2005, respectively. He was employed as a research fellow with Advanced Telecommunications Research Institute International (ATR), Japan, from 2005 to 2008. He is currently a research fellow of Centre for Vision, Speech and Signal Processing (CVSSP) in University of Surrey, UK. His research interests include 3D computer vision, image-based modeling and media production.

Jean-Yves Guillemaut (M03) received an MEng degree from the Ecole Centrale de Nantes, France, in 2001, and a PhD degree from the University of Surrey, U.K., in 2005. He is currently a Lecturer with the Centre for Vision, Speech and Signal Processing, University of Surrey, U.K. His research interests includes free-viewpoint video and 3D TV, image/video-based scene reconstruction and rendering, image/video segmentation and matting, camera calibration, and active appearance models for face recognition.

Takeshi Takai received the BEng degree in electrical engineering from Doshisha University in 1998, the MEng degree in engineering from Nara Institute Science and Technology in 2000, and the PhD degree in informatics from Kyoto University in 2005. His research interests are in computer vision and computer graphics and include visualization of the essence from the real world. He was a research fellow in CVSSP, University of Surrey, and is now a research fellow in Kyoto University.

Muhammad Sarim received the Bsc (Hons.) and Msc degrees in Physics from University of Karachi, Karachi, Pakistan in 1999 and 2000 respectively. He received the PhD degree in CVSSP, University of Surrey in 2010. He is a assistant professor of Computer Science at Federal Urdu University of Arts, Science and Technology, Karachi, Pakistan. He was a lecturer of Physics at University of Karachi in 2001-2004, and Federal Urdu University of Arts, Science and Technology in 2005-2007. His research interests include wide-baseline image/video matting and segmentation for 3D scene reconstruction.

Adrian Hilton (M96) received the B.S.(Hons.) and D.Phil. degrees from University of Sussex, UK, in 1988 and 1992, respectively. He is a professor of Computer Vision and Graphics and Director of the Centre for Vision, Speech and Signal Processing at the University of Surrey, UK. His research interest is robust computer vision to model and understand real world scenes. Contributions include technologies for the first hand-held 3D scanner, modelling of people from images and 3D video for games, broadcast and film production. He currently leads research investigating the use of computer vision for applications in entertainment content production, visual interaction and clinical analysis.