

# Dynamic 3D Scene Reconstruction in Outdoor Environments

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

Centre for Vision, Speech and Signal Processing, University of Surrey, GU2 7XH, UK

{h.kim, m.farooqui, t.takai, j.guillemaut, a.hilton}@surrey.ac.uk

<http://www.ee.surrey.ac.uk/CVSSP/>

## Abstract

*A number of systems have been developed for dynamic 3D reconstruction from multiple view videos over the past decade. In this paper we present a system for multiple view reconstruction of dynamic outdoor scenes transferring studio technology to uncontrolled environments.*

*A synchronised portable multiple camera system is composed of off-the-shelf HD cameras for dynamic scene capture. For foreground extraction, we propose a multi-view trimap propagation method which is robust against dynamic changes in appearance between views and over time. This allows us to apply state-of-the-art natural image matting algorithms for multi-view sequences with minimal interaction. Optimal 3D surface of the foreground models are reconstructed by integrating multi-view shape cues and features.*

*For background modelling, we use a line scan camera with a fish eye lens to capture a full environment with high resolution. The environment model is reconstructed from a spherical stereo image pair with sub-pixel correspondence.*

*Finally the foreground and background models are merged into a 3D world coordinate and the composite model is rendered from arbitrary viewpoints. We show that the proposed system generates high quality scene images with dynamic virtual camera actions.*

## 1. Introduction

Since Kanade et al. proposed the concept of “Virtualized Reality” as a new visual medium for free-view rendering of pre-recorded scenes in a controlled environment [1], many multiple camera acquisition systems and computer vision algorithms for robust surface reconstruction and high-quality view synthesis have been developed [2-4]. The concept of using multiple cameras for 3D production opens up the potential for the “3D Virtual Studio” in which the dynamic shape and appearance can be captured as a 3D computer graphics model. This is now attracting considerable interest as a production tool in film, broadcast and games [5][6]. Traditionally the use of this 3D production has centred on the virtual studio in which a

camera films live action against a controlled environment such as blue screen or static background with ambient lighting. Guillemaut et al. [7] reconstruct 3D soccer and rugby matches from multiple camera views in a stadium environment with relatively controlled backgrounds and illumination. Recently, Halser et al [8] and Shaheen et al. [9] reconstructed outdoor actions without environment or illumination constraints with multiple portable cameras, but they used skeleton-based human models from motion capture or laser scan.

The transfer of studio technology to outdoor capture introduces several problems. The first problem is movability of capture systems. State-of-the-art multiple camera studio facilities use high-quality 3CCD fixed cameras with hard-wired units for synchronisation and data transfer [10]. Systems for outdoor capture should be easy to move and setup to cope with changes in weather or capturing environment and avoid the use of wired connections for power, synchronisation or data transfer.

Second, outdoor images require more robust matting and matching methods in surface reconstruction because of uncontrolled lighting, cluttered or moving backgrounds and video compression. Most studio-based systems assume good silhouette extraction and feature matching between views, and use them as constraints in model reconstruction [2-4][10]. Many powerful natural image matting algorithms were recently developed but they require manual interaction to define key frame trimaps at regular intervals (10-20frames) [11-13]. For multiple view capture it is prohibitively time consuming to interactively define trimaps in all views. Robust multiple view trimap propagation techniques are introduced to allow application of natural image matting across multiple views from a small number of manually defined key-frame trimaps in a single view (1-2 trimaps/200 frames). The feature matching problem can be overcome by using matching costs which are robust against radiometric differences [14].

Finally, environment modelling has been considered as a separate issue because normal cameras provide only limited coverage of the surrounding environment. There have been trials to use multiple cameras or moving cameras to reconstruct an environment, but they could not recover full 3D geometry [15, 16]. The most common way to capture the

full 3D space is to use a catadioptric omnidirectional camera or fisheye lenses [17, 18], but they use only one CCD to capture the full 3D space so that the resolution of partial images from the full view is low compared with the resolution of the multiple view video cameras used to capture the foreground scene. We have proposed to use a spherical stereo from a line scan camera for reconstructing a 3D environment with high resolution [19].

In this paper, we propose a dynamic 3D reconstruction system for outdoor capture. The static environment is captured by two rotating spherical cameras and the dynamic scene is recorded by portable HD cameras. The environment is reconstructed by spherical stereo geometry and the dynamic scene by multi-view matting and global surface optimisation. Finally dynamic foreground scene and static background scene are merged into one 3D coordinate system and the full 3D scene is rendered from arbitrary viewpoints. The main contributions of this paper are:

- We introduce a portable capture system to allow off-the-shelf HD cameras to be used for outdoor wide-baseline multi-view capture. The cameras are wireless, synchronised and calibrated by simple methods.
- We reconstruct a full 3D environment from a high-resolution spherical colour image pair acquired with a line scan camera. PDE-based floating-point disparity estimation method is proposed to recover smooth depth fields with sub-pixel disparity.
- We propose a multiple view trimap propagation algorithm which is robust to changes in appearance between views and over time. This allows us to apply powerful state-of-the-art natural matting algorithms for multiple view sequences with minimum user interaction.
- Finally, we provide a rendering interface which merges static background model from the spherical camera and dynamic foreground model from multiple cameras into a common 3D space, so that the full 3D geometry and texture can be rendered from any viewpoint.

## 2. Capture System

### 2.1. Environment capture system

For static background capture, we use a line scan camera system which synthesizes a full spherical view from a set of images taken by an input camera rotating around a vertical axis [19]. A spherical image is generated by mosaicing rays from the rotating slits. Strips are taken from sampling the rays on a hemisphere at its centre of projection, and stitched together into a new image. We attached a Nikon 16mm f/2.8 AF fisheye lens to the system and it generates images with maximum resolution of 10752x5376. The scene is captured with the camera at two different heights to recover depth information of the scene through stereo geometry.



Figure 1: Spherical stereo pair for background (Top and bottom)



(a) FallingDown (53rd frame of 143 frames)



(b) Handshake (100th frame of 175 frames)

Figure 2: Multi-view capture of dynamic scene

One of the traditional problems of spherical stereo imaging using fisheye lenses is relatively low resolution of the image and complex search along conic curves for stereo matching. Line scan imaging provides high resolution images and the stereo matching process can be simplified to a 1D search along the scan line in the image, which covers the full 3D space if the two capture points are vertically aligned. Figure 1 shows a stereo image pair captured with a vertical baseline of 60cm, which has a maximum disparity of 240 pixels.

### 2.2. Dynamic scene capture system

The multiple camera system comprises eight HDV camcorders, Canon XH G1, and provides compressed MPEG2 streams with 1920x1080 resolution at 25Hz progressive scan. We attached a Canon 4.5-90mm f/1.6-35

lens to each camera. The cameras can be synchronised by genlock, but we do not use it because it requires an external timing source and cables to all cameras. Instead of that, they are synchronised using time code that is synchronised between cameras in advance. The cameras can be controlled by a PC with IEEE 1394 cables, but we use a remote controller in order to avoid any cables. The cameras are placed on tripods located around the capture volume. The captured scenes are recorded to HDV tapes and transferred to disk for processing offline.

The intrinsic parameters of cameras are estimated by using a checker board [20]. The extrinsic parameters are estimated by wand-based calibration using bundle adjustment from positions of coloured balls [3]. We use particle filtering to track the balls to cope with instability of unconstrained backgrounds of outdoor scenes. Figure 2 shows examples of the multi-view capture at the same moment.

### 3. Static Environment Modelling

#### 3.1. Spherical stereo imaging

We use spherical stereo geometry for reconstructing the full 3D scene structure. Figure 3 shows an epipolar plane which is defined by a 3D point and the two camera positions. The angles of the projection of the point  $p$  onto the spherical image pair displaced along the  $y$ -axis are  $\theta_t$  and  $\theta_b$ , respectively, the angle disparity  $d$  of point  $p$  can be defined as the difference of the angles of  $\theta_t$  and  $\theta_b$  as:

$$d = \theta_t - \theta_b \quad (1)$$

From the relationship between two cameras, the distances of the point  $p$  from the two cameras are calculated as follows.

$$\begin{aligned} r_t &= B / \left( \frac{\sin \theta_t}{\tan(\theta_t + d)} - \cos \theta_t \right) \\ r_b &= B / \left( \cos \theta_b - \frac{\sin \theta_b}{\tan(\theta_b - d)} \right) \end{aligned} \quad (2)$$

Therefore, if the two images are vertically aligned and the correspondence of scene points from the spherical stereo image pairs is known, we can compute the disparity of the point with Eq. (1) and its distance from the spherical camera with Eq. (2).

#### 3.2. Sub-pixel disparity estimation

A number of studies have been reported on the stereo correspondence problem over the past three decades [21]. However, most current disparity estimation algorithms

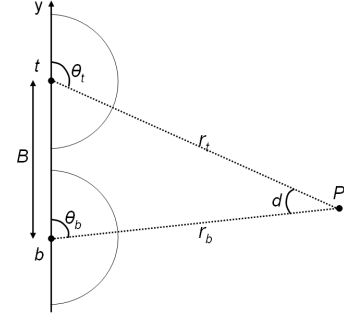


Figure 3: Spherical stereo geometry

produce discrete disparity fields which are not sufficient to find smooth surface depth. For example, when image size is 1600x1200 and baseline distance is 20cm, a 0.5 pixel disparity error for the point at a distance of 5m leads to a depth error of 6cm, and the point at 8m leads to a 15cm depth error. Spherical imaging has serious radial distortion and this quantisation error can cause stepwise ringing artefacts on reconstructed surfaces. Therefore, we need a sub-pixel disparity field to generate smooth surface and minimize depth errors. We use a PDE-based method which solves the correspondence problem in a continuous domain and produces floating-point disparity fields.

First, we tested our previous algorithm [19] which minimises the energy functional:

$$\begin{aligned} E(d(x, y)) &= \int_{\Omega} (I_t(x, y) - I_b(x, y + d(x, y)))^2 dx dy \\ &+ \lambda \int_{\Omega} \psi(\nabla d(x, y), \nabla I_t(x, y)) dx dy \\ \nabla(\psi(\nabla d, \nabla I_t)) &= g(|\nabla I_t|^2) \nabla d \end{aligned} \quad (3)$$

where  $\Omega$  is an image plane,  $\lambda$  a weighting factor of the smoothing term, and  $g(\bullet)$  a regularisation function. The solution of Eq. (3) can be obtained by calculating the corresponding PDE of Eq. (4)

$$\begin{aligned} \frac{\partial d}{\partial t} &= \lambda \text{div}(g(|\nabla I_t(x, y)|^2) \nabla d(x, y)) \\ &+ (I_t(x, y) - I_b(x, y + d)) \frac{\partial I_b(x, y + d)}{\partial y} \end{aligned} \quad (4)$$

This method produces accurate and smooth depth fields across most regions, but it has serious limitations related to occlusion around depth discontinuity regions. Therefore we modified the PDE system as follows so that it can deal with occlusion.

$$\begin{aligned} \frac{\partial d_t}{\partial t} &= \lambda \text{div}(g(|\nabla I_t(x, y)|^2) \nabla d_t(x, y)) \\ &+ H(1 - O_t(x, y))(I_t(x, y) - I_b(x, y + d_t)) \frac{\partial I_b(x, y + d_t)}{\partial y} \\ O_t(x) &= |d_t(x, y) - d_b(x, y + d_t)| \end{aligned} \quad (5)$$

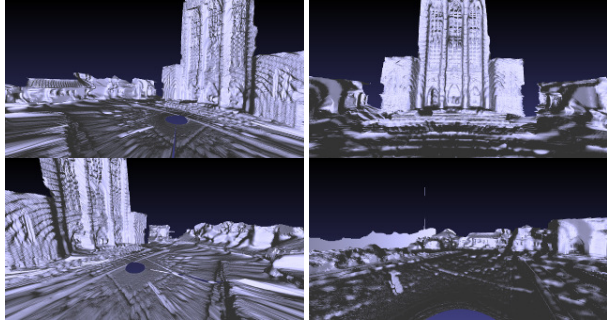
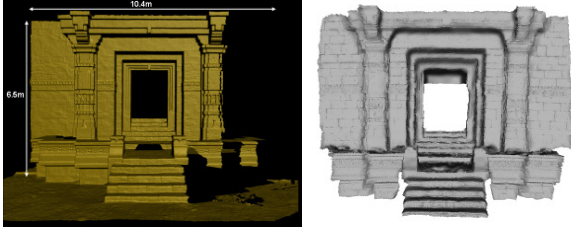


Figure 4: Snapshots of Reconstructed geometry



(a) Ground-truth by LIDAR scan (b) Reconstructed model  
Figure 5: Evaluation against ground-truth

where  $H(\bullet)$  is a unit step function. In visible regions, normal balanced diffusion equation with data term works to find the optimised solution. However, the data term produces errors in occluded region because there is no corresponding point for the occluded point. Therefore, in occluded region, only pure diffusion filtering works for smoothing disparity field and propagating correct depth information from visible regions to occluded regions. However, inserting the switch cannot guarantee the convergence of the solution around the boundary between visible and occluded regions. We set a maximum number of iterations of the solver to prevent eternal resonance. We also use a hierarchical approach to reduce computation time for large images and alleviate the local minimum problem.

### 3.3. 3D model reconstruction

The estimated dense disparity fields can be converted into depth information by utilizing camera geometry as described in Section 3.1. As a result, a 3D environment model of the real scene is reconstructed from the original texture and disparity information. The original images are described in spherical coordinates, so we convert them into the Cartesian coordinate system, and then project them to 3D space to generate a 3D mesh.

Figure 4 shows snapshots of the rendered scene of the reconstructed model at arbitrary viewpoints. The results show a natural-looking geometry of the environment.

For objective evaluation, we chose one object in the captured test images and compared its depth with ground-truth range data scanned by a LIDAR sensor. Figure

5 (a) shows the ground-truth model by LIDAR scan and Fig. 5 (b) is the reconstructed model by the proposed algorithm. We can see that the reconstructed model shows very fine structure with smooth surface. The average depth error over the whole common area was  $-0.20\text{cm}$  with  $15.2\text{cm}$  standard deviation. However, most errors occur at depth discontinuities. The errors in the depth-discontinuity region are mainly from the difference of field-of-views (FOV) of the LIDAR sensor and spherical imaging. The average depth error in uniform appearance region was  $-0.38\text{cm}$  with  $0.41\text{cm}$  standard deviation.

## 4. Multi-view Video Matting

State-of-the-art video matting algorithms require the labour intensive task of a drawing trimap for each image to be applied to multiple view sequences. This section presents a novel framework for wide-baseline multi-view video matting given only a sparse set of manually defined key frames in a single view. We assume that all the cameras capture the same foreground scene from different directions.

Trimaps are constructed by spatio-temporal propagation of high confidence trimap labels using a Bayesian inference framework from a sparse set of key frame trimaps  $\{T^1_i\}_{i=1}^k$  for a single view. The key frame trimaps  $T^1_i$  specify the definite foreground  $F$  and background  $B$  pixels with the corresponding trimap confidence  $C^1_i = 1$  while the remaining pixels are labeled as unknown  $U$  with a confidence of 0. High confidence pixels are used to statistically model the temporally static global foreground and background appearance in the key frames represented as  $\{M^{SF}, M^{SB}\}$ . To process the views at time  $t$ , we adaptively model the temporal foreground and background variations, due to illumination and shadows, from the global static models represented as global dynamic models  $\{M^{DF}(t, \tau), M^{DB}(t, \tau)\}$  over a temporal window  $(t-\tau, t+I)$ . A local pixel-wise background model  $\{M^{LB}\}$  is also constructed to capture the pixel-wise background variations using background sequence. All models are represented by mixture of Gaussians in colour space and can be constructed using any state-of-the-art clustering algorithms. Each component of a model is assigned a confidence  $\psi_i$  estimated from the confidence of the member pixels. For a new frame  $I^v_t$  at time  $t$  in view  $v$ , the trimap label for a pixel  $q$  is propagated using the MAP (maximum a posteriori) estimation of label based on the global foreground  $M^{GF} = \{M^{SF} \cup M^{DF}\}$ , background  $M^{GB} = \{M^{SB} \cup M^{DB}\}$  and local background  $M^{LB}(q)$  models. The posterior probability of a pixel  $q$  belonging to the  $i^{\text{th}}$  component of a model  $M_i(\mu_i, \Sigma_i)$  is given by Bayes rule:

$$P(\mu_i, \Sigma_i | x = q) = \frac{P(x = q | \mu_i, \Sigma_i) P(\mu_i, \Sigma_i)}{P(x = q)}. \quad (6)$$

The term  $P(\mu_i, \Sigma_i)$  is the prior for the cluster and is given by the cluster confidence  $\psi_i$ . The term  $P(x=q)$  is parameter independent and can be ignored in optimization. To estimate the most likely cluster  $M_{ml}$  with MAP estimates  $(\mu_{ml}, \Sigma_{ml})$ , Eq. (6) is maximized over the entire component space of model  $M$ :

$$(\mu_{ml}, \Sigma_{ml})_{M_{ml}} = \arg \max_M P(x = q | \mu_i, \Sigma_i) \psi_i \quad (7)$$

Since the clusters have multivariate Gaussian distributions, the MAP estimates correspond to the minimum squared Mahalanobis distance  $Q$  for the global foreground  $Q_{min}^{GF}$ , global background  $Q_{min}^{GB}$  and local background  $Q_{min}^{LB}$ . The squared Mahalanobis distance follows the chi-square distribution over  $f$  degrees of freedom that is  $Q \sim \chi^2(f)$ , where  $f=3$ , given by the dimension of the colour space. The inferential statistics based on  $\chi^2$  are used to infer the trimap label of the pixel  $q$ . Three different null hypotheses ( $H_0^{GF}, H_0^{GB}, H_0^{LB}$ ) are defined stating the membership of pixel  $q$  to models ( $M_{mb}^{GF}, M_{mb}^{GB}, M_{ml}^{LB}$ ) at critical value of  $\chi^2_{\beta, f}$  with significance level of  $\beta=0.05$  (95% confidence). Initially only the foreground pixel labels with high-confidence are propagated as there is no local foreground model. High-confidence foreground pixels are inferred as follows:

$$T_i^v(q) = \begin{cases} F : (H_0^{GF}) \wedge (\sim H_0^{GB}) \wedge (\sim H_0^{LB}) \\ U : otherwise \end{cases} \quad (8)$$

Equation 8 labels foreground pixels only where there is no ambiguity between foreground and background, resulting in holes in the foreground. Given the initial foreground labeling we can estimate a local foreground  $M^{LF}(q)$  model for a pixel  $q$  labeled as unknown  $U$  from the foreground pixels in the neighbourhood  $R(q)$ . For the local foreground model we can then define a null hypothesis  $H_0^{LF}$  for pixel  $q$  belonging to the local foreground model and infer the trimap labels for all unknown pixels as:

$$T_i^v(q) = \begin{cases} F : (H_0^f) \wedge (\sim H_0^b) \\ B : (H_0^b) \wedge (\sim H_0^f) \\ U : otherwise \end{cases} \quad (9)$$

A confidence map  $C_i^v$  is associated to the trimap  $T_i^v$  by assigning a confidence to each foreground and background which is used to estimate the colour model confidence. Confidence for a foreground pixel  $q$  is formulated using the confidence of the most likely foreground cluster  $\psi_{ml}^f$  and the

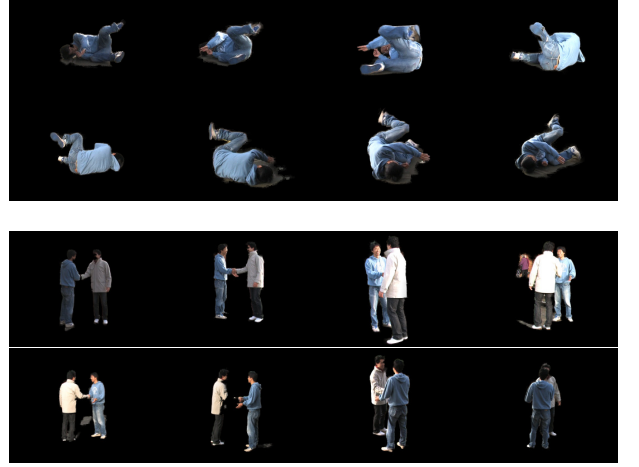


Figure 6: Extracted foreground regions in multiple views

minimum squared Mahalanobis foreground and background distance as:

$$C_i^v(q) = \psi_{ml}^f \left( 1 - e^{-Q_{min}^b / \chi_{\beta, f}^2} \right) e^{-Q_{min}^f / \chi_{\beta, f}^2}. \quad (9)$$

The confidence to the background pixel is assigned by interchanging the foreground and background parameters.

Once the final trimap is estimated, the alpha matte can be estimated by using existing natural image matting algorithm such as closed-form or non-parametric techniques [12][13]. The foreground silhouette is estimated by thresholding the alpha matte.

Figure 6 shows the cropped segmentation results of 8 views in Fig. 2. The proposed algorithm produces accurate foreground boundary in all 8 views using only a single hand drawn trimap of a key-frame from a single view, selected from 100-200 frames per view. Small errors in the matte may occur when there are large overlaps in foreground and background appearance. Such errors are not generally consistent between views and can be eliminated in reconstruction [22]. The moving background regions are automatically carved out by other view silhouettes in visual hull reconstruction. Additional key-frames could be added to correct the small errors in matting. Throughout this work no manual correction was used.

## 5. Dynamic Scene Reconstruction

### 5.1. Surface reconstruction of people

For dynamic human modelling from multiple images and silhouettes, we use a global optimization technique to extract the optimal 3D surface of the model by integrating multiple shape cues and features for robust wide-baseline reconstruction [3][23].

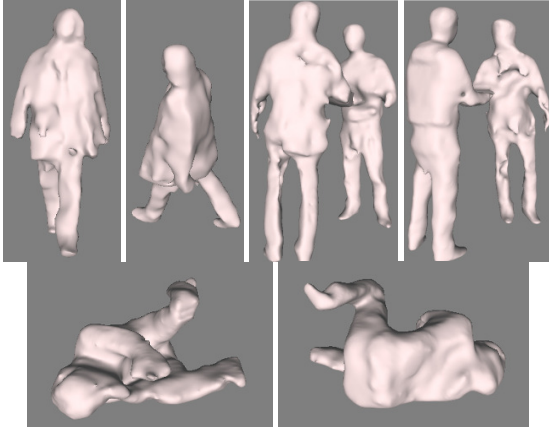


Figure 7: Reconstructed 3D models from arbitrary viewpoints

Once the extent of the scene is defined by reconstructing the visual-hull, surface features are matched between views to derive constraints on the location of the scene surface. A Canny-Deriche edge detector is used for feature detection, and each feature contour in an image is matched with the appearance in an adjacent camera view by considering the camera epipolar geometry. Correspondence is verified by enforcing left-right consistency between views.

Dense surface reconstruction is then performed inside the volume defined by the visual hull reconstruction. The volume is discretised and refined with a maximum-flow/minimum-cut problem [24] on a graph defined in the volume. Each voxel forms a node in the graph with adjacent voxels connected by graph edges. Edges are weighted by a cost defined by the consistency in appearance between camera images. The maximum flow on the graph saturates the set of edges where the cost is minimized and the consistency is maximized. The final surface can then be extracted as the set of saturated edges cutting the graph. In the graph-cut optimization, the feature contours are used as constraints to derive a surface that passes through the reconstructed feature contours and reproduces the initial set of silhouette images. Finally, the surface is extracted from the volume reconstruction as a triangulated mesh.

Figure 7 shows the results of surface reconstruction from the captured outdoor scene. We can see some errors on the reconstructed surfaces. Holes and amputated parts are from segmentation errors because the initial visual hull set the maximum volume of the model and silhouette information works as hard constraints. Bumpy surface errors are from surface optimization. Matching errors in surface optimization were caused by camera noise, compression errors, uniform textures and non-Lambertian surfaces. However, the results still look very close to natural 3D models. Small surface errors can be concealed by texture mapping.

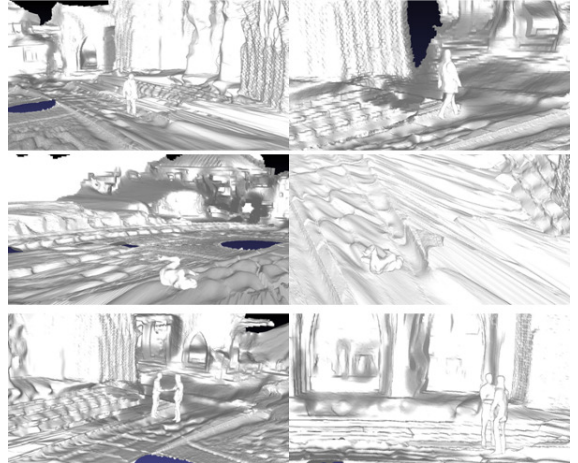


Figure 8: Registration of foreground and background models

## 5.2. Model composition and final rendering

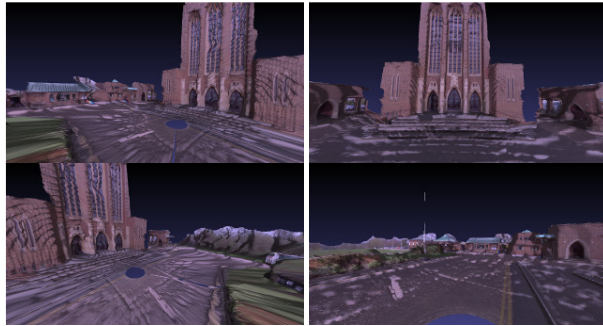
Finally, the reconstructed foreground models are merged into the background model. We capture the background scene with the spherical cameras slightly out of the foreground capture volume in advance. The reasons are; 1) we do not want to include multiple cameras in our background model, 2) depth errors from disparity estimation diverge to infinite around polar regions in Eq. (2). Therefore, the origins of background and foreground coordinates are not consistent. However, they can be easily aligned by rotating on the  $y$ -axis and shifting in the  $x$  and  $z$  directions because both coordinates are constructed in real world scale. Figure 8 shows the composite model.

In texture mapping, we use different methods for background and foreground models. 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 which matches the 3D point onto a texture.

However, dynamic foreground models are reconstructed from multiple cameras and they provide only partial texture information of the model. Due to occlusion and changes in a 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 adopted view-dependent texture mapping technique to assign camera images to each face with the best visibility [25]. The textures derived from selected cameras are then composited onto the surface by blending the textures.

Figure 9 shows results of texture mapping from the same viewpoints as Fig. 4 and Fig. 7, respectively.

For final rendering, we developed a virtual camera controller to enable arbitrary viewpoint rendering of the model with texture mapping. It can generate various virtual



(a) Background model (Fig. 4)



(b) Foreground model (Fig. 7)

Figure 9: Results of texture mapping

camera actions such as interpolation between two real camera positions, dolly shot from a certain viewpoint, and time-slice shot to generate dynamic 360° of a certain frame. Users can also design their own camera actions. The renderer finally creates a rendered composite video. Snapshots of the rendered video with various virtual camera actions are shown in Fig. 10.

## 6. Discussion and conclusion

In this paper, we introduced a 3D scene reconstruction system for outdoor capture. We have addressed the practical problems of outdoor scene capture and proposed robust solution to segment and reconstruct dynamic scene elements.

The outdoor scene is captured with a set of portable cameras, which are synchronised and do not require any cabling. However, MPEG compression reduces the quality of the captured images compared to studio capture requiring more robust reconstruction methods.

Background scene is captured by spherical cameras and reconstructed by spherical stereo geometry. The proposed method reconstructs accurate depth information. This gives a reconstruction with sufficient quality for rendering novel views. Fusion of spherical stereo scene reconstruction for multiple views is required in future work to fill holes in the

reconstructed scene due to occlusion.

To reconstruct the dynamic foreground from multi-view video sequences we introduce a multiple view trimap propagation algorithm. This approach allows trimaps to be propagated across multiple views given a small number of manually specified key-frames trimaps in a single view. This approach allows state-of-the-art natural image matting techniques to be used for multiple view sequences without the prohibitive time consuming manual interaction for every view. Typically 1-2 keyframes are specified in a single view for matting multiple view sequences with 100-200frames per view.

Finally, we reconstructed the foreground model from multiple videos by the global surface optimization method and merged the model into the background geometry. The final composite model can be rendered from any viewpoint with high quality textures.

The proposed system tries to transfer the multiple view camera system out of the studio and adapt it for use in real outdoor environments with natural scene backgrounds and uncontrolled illumination. Further research is required to refine the reconstructed foreground and background models to achieve a visual quality comparable to the captured images for production. The system presented and algorithms introduced in this work provide a framework for many future systems and applications such as 3D film, sport and documentary productions.

## Acknowledgement

This research was executed with the financial support of the EU IST FP7 project i3Dpost.

## References

- [1] T. Kanade, P. W. Rander, and P. J. Narayanan. Virtualized Reality: Constructing Virtual Worlds from Real Scenes. *IEEE Multimedia*, 4(1):34-47, 1997.
- [2] T. Matsuyama, X. Wu, T. Takai, and S. Nobuhara. Real-Time 3DShape Reconstruction, Dynamic 3D Mesh Deformation, and High Fidelity Visualization for 3D Video. *CVIU*, 96(3):393-434, 2004.
- [3] J. Starck and A. Hilton. Surface Capture for Performance-Based Animation. *IEEE CG&A*, 27(3):21-31, 2007.
- [4] Y. Furukawa and J. Ponce. Carved Visual Hulls for Image-Based Modeling. *IJCV*, 81(1):53-67, 2008.
- [5] M. Price, J. Chandaria, O. Grau, G.A. 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. *Proc. International Broadcasting Convention*, 348-356, 2002.
- [6] O. Grau and G. Thomas. 3D image sequence acquisition for TV & film production. *Proc. 3DPVT*, 320-326, 2002.

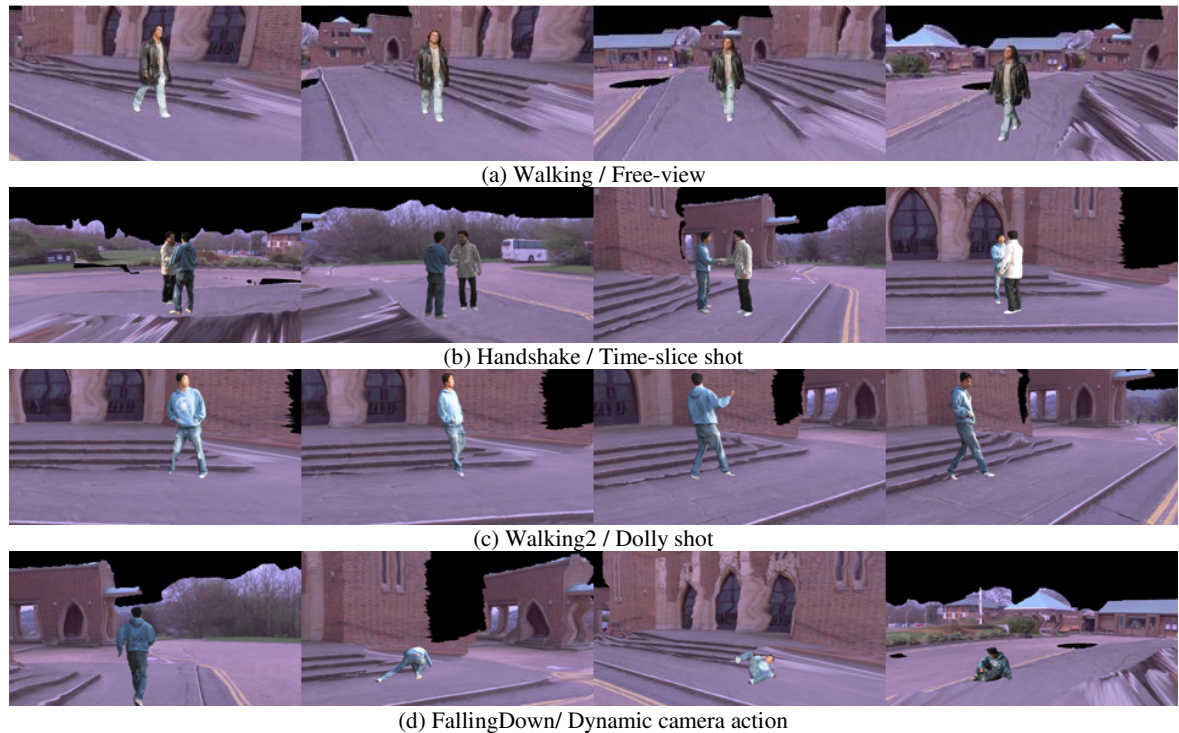


Figure 10: Full 3D scene rendering with dynamic camera actions

- [7] J.-Y. Guillemaut, J. Kilner, and A. Hilton. Robust Graph-Cut Scene Segmentation and Reconstruction for Free-Viewpoint Video of Complex Dynamic Scenes. *Proc. ICCV*, 809-816, 2009.
- [8] N. Hasler, B. Rosenhahn, T. Thormaehlen, M. Wand, H.P. Seidel. Markerless Motion Capture with Unsynchronized Moving Cameras. *Proc. CVPR*, 224-231, 2009
- [9] M. Shaheen, J. Gall, R. Strzodka, L.V. Gool, and H.P. Seidel. A Comparison of 3D Model-based Tracking Approaches for Human Motion Capture in Uncontrolled Environments. *Proc. WACV*, 1-8, 2009.
- [10] J. Starck, A. Maki, S. Nobuhara, A. Hilton and T. Matsuyama. The Multiple-Camera 3-D Production Studio. *IEEE Trans. CSVT*, 19(6):856-869, 2009.
- [11] J. Sun, W. Zhang, X. Tang, and H.-Y. Shum. Background cut. *Proc. ECCV*, 628-641, 2006.
- [12] A. Levin, D. Lischinski, and Y. Weiss. A closed form solution to natural image matting. *IEEE Trans. PAMI*, 30(2):228-242, 2008.
- [13] M. Sarim, A. Hilton, J.-Y. Guillemaut, and H. Kim. Nonparametric natural image matting. *Proc. ICIP*, 3221-3224, 2009.
- [14] H. Hirschmuller and D. Scharstein. Evaluation of Stereo Matching Costs on Images with Radiometric Differences. *IEEE Trans. PAMI*, 31(9): 582-1599, 2009.
- [15] Pollefeys, M., Koch, R., Vergauwen, M., Gool, L.V.: Automated reconstruction of 3D scenes from sequences of images. *PandRS*, 55(4): 251-267, 2000
- [16] A. Broadhurst, T. Drummond, and R. Cipolla. A probabilistic framework for the Space Carving algorithm. *Proc. ICCV*, 388-393, 2001
- [17] S. Nayar. Catadioptric Omnidirectional Camera. *proc. CVPR*, 482-488, 1997.
- [18] S. Li. Full-View Spherical Image Camera. *Proc. ICPR*, 386-390, 2006
- [19] H. Kim, and A. Hilton: 3D Environment Modelling Using Spherical Stereo Imaging. *Proc. 3DIM*, 2009
- [20] Z. Zhang. Flexible Camera Calibration by Viewing a Plane from Unknown Orientations. *Proc. ICCV*, 666-673, 1999.
- [21] D. Scharstein and R. Szeliski. A Taxonomy and Evaluation of Dense Two-frame Stereo Correspondence Algorithms. *IJCV*, 47(1):7-42, 2002.
- [22] H. Kim, R. Sakamoto, I. Kitahara, T. Toriyama and K. Kogure. Reliability-Based 3D Reconstruction in Real Environment. *Proc. ACM Multimedia*, 257-260, 2007.
- [23] J. Starck, G. Miller and A. Hilton. Volumetric Stereo with Silhouette and Feature Constraints. *Proc. BMVC*, 1189-1198, 2006.
- [24] Y. Boykov and V. Kolmogorov. Computing geodesics and minimal surfaces via graph cuts. *Proc. ICCV*, 26-33, 2003.
- [25] Debevec, Y. Yu, and G. Borshukov. Efficient view-dependent image-based rendering with projective texture mapping. *Proc. Eurographics Rendering Workshop*, 105-116, 1998.