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

3D Scene Reconstruction from Multiple Spherical Stereo Pairs

Hansung Kim · Adrian Hilton

Received: date / Accepted: date

Abstract We propose a 3D environment modelling method using multiple pairs of high-resolution spherical images. Spherical images of a scene are captured using a rotating line scan camera. Reconstruction is based on stereo image pairs with a vertical displacement between camera views. A 3D mesh model for each pair of spherical images is reconstructed by stereo matching. For accurate surface reconstruction, we propose a PDE-based disparity estimation method which produces continuous depth fields with sharp depth discontinuities even in occluded and highly textured regions. A full environment model is constructed by fusion of partial reconstruction from spherical stereo pairs at multiple widely spaced locations. To avoid camera calibration steps for all camera locations, we calculate 3D rigid transforms between capture points using feature matching and register all meshes into a unified coordinate system. Finally a complete 3D model of the environment is generated by selecting the most reliable observations among overlapped surface measurements considering surface visibility, orientation and distance from the camera. We analyse the characteristics and behaviour of errors for spherical stereo imaging. Performance of the proposed algorithm is evaluated against ground-truth from the Middlebury stereo test bed and LIDAR scans. Results are also compared with conventional structure-from-motion algorithms. The final composite model is rendered from a wide range of viewpoints with high quality textures.

Hansung Kim · Adrian Hilton
Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, Surrey, GU2 7XH, UK
Tel.: +44-1483-683958
Fax: +44-1483-686031
E-mail: h.kim, a.hilton@surrey.ac.uk

Keywords 3D reconstruction · Environment modelling · Disparity estimation · 3D registration and mesh integration

1 Introduction

Scene reconstruction has been an important research topic in computer and robot vision over the past decade (Akbarzadeh et al., 2006; Desouza and Kak, 2002). The problem of generating visually realistic graphical models of scenes from cameras has been addressed through computer vision techniques. However, there are several problems in environment modelling which are different from the modelling of common objects. The biggest problem is that normal cameras with a limited field-of-view capture only a partial observation of the surrounding environment. Reconstruction of a complete model of the 3D environment requires a large number of views to capture the scene and occluded regions. Reconstruction of scene models from multiple images or video acquired with a standard camera has been the focus of considerable research. However, the limited field-of-view presents a challenging problem to ensure complete scene coverage for reconstruction.

In this research, we propose to reconstruct full static 3D environment models from multiple pairs of spherical stereo images. A spherical camera captures the full surrounding scene visible from the camera location. Acquisition of stereo pairs of spherical images allows dense reconstruction of the surrounding scene. Integration of reconstructions from multiple locations allows a complete 3D scene model to be acquired from a relatively small number of spherical images. In this research a spherical line scan camera is used to capture high-resolution (12574×5658, 70Mpixels) spherical stereo image pairs

with a vertical baseline between views. To reconstruct robust and accurate depth from stereo pairs dense floating-point disparity maps are estimated using a partial differential equation (PDE) based stereo matching algorithm. This incorporates a novel cost function to ensure accurate reconstruction for edges, depth discontinuities and regions of uniform appearance. Automatic registration of reconstructions from multiple stereo pairs at widely spread locations is performed using SURF feature matching (Bay et al., 2008) and a RANSAC-based algorithm to calculate 3D rigid transforms between multiple viewpoints. Finally a complete 3D model of the environment is generated by integration of partial scene models based on surface reliability from individual stereo pairs. This ensures both preservation of surface detail from stereo reconstruction and reliable reconstruction of a complete 3D scene model.

Preliminary versions of the approach presented in this paper previously appeared in conference proceedings (Kim and Hilton, 2009, 2010). In the work of Kim and Hilton (2009), we proposed to use a spherical camera for environment capture and to use a standard PDE-based disparity estimation for scene reconstruction. Kim and Hilton (2010) extended this approach to multi-view capture and registration. In this paper, we extend the PDE based disparity estimation to achieve robust scene reconstruction and integrate this approach with reliability-based multiple view registration to introduce a complete workflow for outdoor scene reconstruction. The performance of our proposed approach is quantitatively evaluated against accurate ground-truth measurement using active Light Detection and Ranging (LIDAR) sensing. Scene reconstruction from spherical stereo image pairs of large scenes achieves accuracy of within <1% of the scene dimensions (<10cm over 10m) whilst also providing high-resolution colour appearance for visualisation applications. The following are the main contributions of this paper:

- We present a complete workflow for reconstruction of static 3D scene models.
- We introduce a robust multi-resolution PDE-based disparity estimation method. PDE-based disparity estimation produces floating-point disparity fields to obtain accurate and smooth depth. We extend the PDE formulation to handle problems of occlusion at depth discontinuities and over-segmentation in highly textured regions. A hierarchical structure is used to allow solution for large images and to accelerate the calculation. This approach is demonstrated to achieve accurate disparity estimation on the Middlebury stereo benchmark datasets.
- We introduce an efficient approach to reliably integrate reconstructions from multiple stereo pairs whilst preserving the fine surface detail of individual stereo reconstructions for large-scale scene models. This integration algorithm exploits the estimated reliability of surface reconstruction based on surface visibility, orientation and distance to the camera.
- We evaluate the accuracy of reconstruction against ground-truth models scanned by a LIDAR sensor and analyse the characteristics and behaviour of errors for spherical stereo imaging.

The rest of this paper is organised as follows: Section 2 outlines related previous works. Section 3 presents the capture system using a line scan camera and spherical stereo geometry for depth reconstruction. Section 4 presents a novel PDE-based disparity estimation algorithm for reliable correspondence in the presence of occlusion and highly textured regions, and Section 5 presents the reliability based integration algorithm to register and combine reconstructions from widely spaced locations into one complete 3D surface model of the scene with high-resolution appearance. Experimental results and discussion are given in Section 6, and Section 7 presents the conclusion from this work. Supplemental video is also available showing results of reconstruction for multiple scenes at: `http://www.youtube.com/watch?v=x3KdI8ZWZiQ`.

2 Related Work

2.1 Environment modelling

Outdoor environment modelling can be classified into two different categories according to the input sources: active methods using range sensors and passive methods using only images.

Active techniques utilise ultrasonic or laser scanners to measure distance to an object or surface. LIDAR is one of the most popular depth ranging techniques, which measures the range by the time delay between transmission of a pulse and detection of the reflected signal (Lemmens, 2007). Asai et al. (2005) developed a 3D reconstruction system for outdoor areas using a laser rangefinder and an omnidirectional multi-camera system which can capture a wide-angle high-resolution image. They captured range and colour images at 68 points in their university campus and merged them into a common 3D space.

Active sensing techniques yield accurate depth information, but there are problems with respect to hardware cost, materials in the environment and temporal/spatial consistency with an imaging sensor. Accurate registration of photometric and geometric information is important for visualisation applications such as

Table 1 Image-based environment modelling techniques

Ref	Input	Strategy	Reconstruction method	Output
Vu et al. (2009)	Multiple images	MVS	Graph-cut + variational	3D mesh
Furukawa et al. (2010)	Multiple images	MVS	Patch matching	Point cloud
Agarwal et al. (2009)	Multiple images	SfM	Feature matching	Point cloud
Goesele et al. (2007)	Multiple images	SfM + MVS	Stereo matching	Depth maps
Frahm et al. (2010)	Multiple images	SfM + MVS	Plane sweeping	3D mesh
Cornelis et al. (2008)	Stereo video + GPS	SfM + Stereo	Dynamic programming	Simplified mesh
Pollefeys et al. (2008)	Multiple videos + GPS	SfM + MVS	Plane sweeping	3D mesh
Kang and Szeliski (1997)	Omnidirectinal images	MVS	Stereo matching	3D mesh
Lhuillier (2008)	Omnidirectinal videos	SfM	Bundle adjustment	Point cloud
Micusik and Kosecka (2009)	Omnidirectinal videos	MVS	Superpixel stereo	Simplified mesh

visual effects in film production. If the active sensing technique is used, image sensing should be performed separately. Sequential sensing cannot be used in a dynamic environment and simultaneous sensing from different locations requires calibration and registration to align depth and image information. Recently, Banno and Ikeuchi (2010) proposed a semi-automatic texturing method for a dense model captured by a range sensor using two spherical images.

Image-based methods are less sensitive to the environment and require a simpler and less expensive setup for reconstruction of 3D geometry. They are also inherently temporally and spatially consistent with images because they extract depth information from captured images. Table 1 gives an overview of a representative set of existing approaches to image-based scene modelling. Accurate outdoor scene reconstruction from multi-view images has been the focus of extensive research (Goesele et al., 2007; Furukawa and Ponce, 2010; Vu et al., 2009; Salman and Yvinec, 2009). Strecha et al. (2008) created a benchmarking site for the quantitative evaluation of multi-view stereo (MVS) algorithms. However, the first problem of multi-view stereo is the relatively small field of view (FOV.) Coverage of the surrounding environment requires a large number of overlapping images of the scene. The second problem is calibration of multiple cameras. Strecha et al. (2008) provided accurate calibration data calculated using markers attached to buildings and LIDAR scanning for the data sets. This requires accurate calibration of all cameras in advance which can be problematic.

Structure from motion (SfM) is a technique to simultaneously recover 3D structure of a scene and the pose of a camera from a video (Dellaert et al., 2000; Pollefeys et al., 2000; Cornelis et al., 2008). SfM originally dealt with multiple images from a single camera, but has recently been extended to image collections from arbitrary cameras without prior knowledge. Snavely et al. (2006) developed Bundler, a SfM software for unordered image collections. Bundler is used for 3D point cloud reconstruction and image registration on large sets of photos of popular sites gathered from the internet and photo collections. Camera parameters are automatically extracted from images and used to initialise a MVS algorithm (Furukawa and Ponce, 2010).

Agarwal et al. (2009) reconstructed full 3D street models from 150,000 photos downloaded from the internet using grid computing with 500 cores over 24 hours. This work is impressive but requires extensive resources for parallel computing and data transfer. Frahm et al. (2010) overcame this problem by using geometric and appearance constraints to obtain a highly parallel implementation on modern graphics processors and multicore architectures. Pollefeys et al. (2008) also used 3,000 video frames to reconstruct one building and 170,000 frames for a small town.

2.2 Reconstruction from spherical images

One of the easiest ways to generate a seamless background from a single viewpoint is synthesising panoramic representations. Apple's QuickTime VR (Chen, 1995) captures a 360° panoramic image of a scene with a camera panning horizontally at a fixed position. The overlap in images is registered first by the user and then stitched together by the software using a matching algorithm. Benosman and Devars (1998) obtain a depth map by rotating two linear image sensors with respect to an axis to generate two cylindrical projection images. Although the panoramic synthesis generates a seamless 360° view in the horizontal direction, it cannot cover the full 3D space because of the limited vertical field of view of the cameras.

The most common way to capture the full 3D space instantaneously is to use a catadioptric omnidirectional camera which uses a mirror combined with a CCD. Micusik et al. (2004) proposed a 3D metric reconstruction of the surrounding scene from two or more uncalibrated omnidirectional images. Lhuillier (2008) proposed a complete system for SfM using omnidirectional images. However, these catadioptric omnidirectional cameras have a large number of systematic parameters in-

cluding the camera and mirror calibration. Another problem of these cameras is the limited resolution. They use only one CCD to capture the full 3D space, so that the resolution of partial images from the full view is relatively low. In order to overcome this resolution problem, the MIT City group developed a spherical capture system[1] which captures the scene with a normal camera and automatically stitches photos together to generate a high-resolution spherical images (Teller et al., 2003). Point Grey developed an omnidirectional multi-camera system, the Ladybug[2], which consists of six XGA color CCDs to overcome the resolution problem. Asai et al. (2005) used this Ladybug for outdoor environment modelling by combining it with a range sensor. Google also developed their own omnidirectional multi-camera system to reconstruct and render street models (Anguelov et al., 2010).

Li (2006) proposed an alternative method for high resolution spherical image acquisition using two fisheye lenses pointing in opposite directions and fusing the two hemispherical images into a single image to construct an immersive virtual environment. He used a spherical projection to reformulate the conventional stereo vision algorithm so as to realise spherical stereo for a pair of spherical images. However, problems of spherical stereo imaging with fisheye lenses are distortion and complex search along a conic curve for stereo matching.

Instead of merging two spherical images from fisheye lenses, Kang and Szeliski (1997) composited cylindrical images from sequences of images taken while a camera is rotated 360° about the vertical axis. They reconstruct 3D models using feature tracking, SfM, and multi-baseline stereo. Feldman and Weinshall (2005) used the Cross Slits(X-Slits) projection with a rotating fisheye camera to generate a high quality spherical image and to reduce the dimension of the plenoptic function. In the X-Slits camera, the projection model is defined by two slits and the projection ray of every 3-D point is defined by the line that passes through the point and intersects both slits (Zomet et al., 2003). The image of a point is the intersection of the projection ray with the image surface. Nayar and Karmarkar (2000) also proposed a similar rotational line scan camera and Haala and Kada (2005) used a line scan panoramic camera system to generate texture for city models. Even though 2D image-based techniques provide a wide view, the background scene can be distorted in rendering according to the viewpoint because it is produced by simple mapping to 2D planes. This is tolerable in some areas, but it is not appropriate for many applications which require a realistic background and dynamic change of view.

[1] City group, `http://city.csail.mit.edu/`

[2] Pointgrey, `http://www.ptgrey.com/`

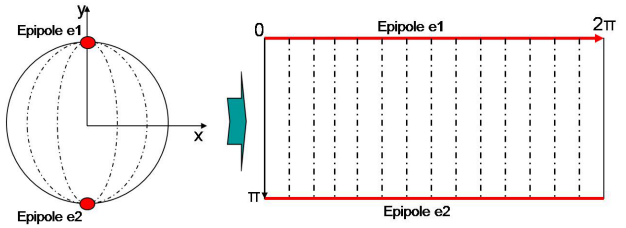
2.3 3D registration and fusion

In previous approaches using active sensors or stereo reconstruction from narrow field-of-view conventional cameras, it is common to reconstruct parts of the scene and merge them into one single coordinate system to large scale scenes. This is performed in two stages: registration of multiple reconstructions into a common coordinate system and integration of the overlapping reconstructions into a single surface model. Fisher (2007) provides a summary of 3D registration and fusion methods on his on-line compendium of computer vision.

Iterative closest point (ICP) has been widely adopted to register two point sets into a common coordinate system (Besl and McKay, 1992). The ICP algorithm finds a rigid 3D transformation (rotation and translation) between two overlapping clouds of points by iteratively minimizing squared-error of registration between the nearest points from one set to the other. The ICP has been modified and extended to surfaces (Rusinkiewicz and Levoy, 2001; Granger et al., 2001; Aiger et al., 2008) and multiple sets (Chen and Medioni, 1992; Soucy and Laurendeau, 1995; Williams and Bennamoun, 2001). However, registration results in multiple overlapping points sets or meshes with different levels of accuracy. The ICP algorithm performs just registration of multiple meshes regardless of overlaps, surface boundaries and reconstruction errors.

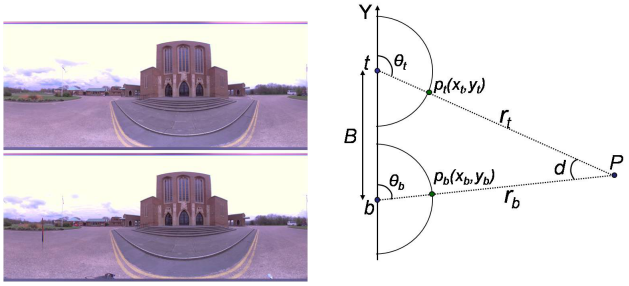
A number of algorithms have been proposed to merge partially overlapping meshes into one single surface mesh by optimising overlapped surfaces. Merrell et al. (2007) proposed a visibility-based depth-map fusion algorithm, which is used for urban 3D reconstruction (Pollefeys et al., 2008). Depth layers are reconstructed by comparing visibility and reliability between neighbouring views. Micusik and Kosecka (2009) took a similar approach with reconstructions from a Ladybug camera to fuse depth maps into dominant planes. Furukawa also proposed a view-clustering algorithm to extend his patch-based multi-view stereo (PMVS) algorithm (Furukawa and Ponce, 2010) to large scale reconstruction (Furukawa et al., 2010). The problem of these approaches is that fusing multiple overlapping layers can result in loss of fine surface details due to registration errors and noise.

In this paper, we are interested in removing multiple outliers and noise on surfaces while keeping geometrical surface details. We propose a reliable surface selection algorithm based on surface visibility and reliability for mesh registration and refinement.



(a) (b)

Fig. 1 Spherical image representation: (a)Spherical imaging; (b)Latitude-longitude imaging



(a) (b)

Fig. 2 Spherical stereo: (a)Spherical stereo pair (top-bottom); (b)Spherical stereo geometry

3 Spherical Stereo Acquisition and Reconstruction

In this work, we use a commercial off-the-shelf line-scan camera[3] with a fisheye lens in order to capture the full environment as a high resolution spherical image. A full spherical view is generated by mosaicing rays from a vertical slit at the centre of a rotating fisheye lens. The camera samples the rays on a hemisphere at its centre of projection and stitches the rays from the rotating slits together into a new image. The camera rotates on the axis passing through its optical centre, therefore the imaging geometry of the line-scan capture can be regarded as conventional perspective projection because all the rays in the spherical image intersect at a single 3D point.

In order to recover depth information from a spherical image pair, the scene is captured with the camera at two different heights. In the line scan imaging, we can regard the epipoles as the two poles and extend the spherical image by latitude-longitude sampling as illustrated in Fig. 1. In this latitude-longitude geometry, the great circles intersecting at the epipoles become parallel straight lines. Therefore, the conventional correlation-based matching over a 1D search range can be used to compute the disparity of spherical stereo images if they are vertically aligned. Ensuring that the displacement between the camera views is parallel to the direction of the line scan camera pixel array makes the epipolar lines correspond to pixel columns in the spherical images. This relies on mechanical precision of the line scan camera. In this work, we scanned a scene at a lower position and raised up the pole of the tripod to a higher position for the second scan. Evaluation demonstrates that this mechanism results in an alignment within 2 pixels for corresponding vertical scan lines in the resulting spherical stereo image pair. If a less accurate camera system is used, errors can be corrected to some extent by rectification (Banno and Ikeuchi, 2010) or X-slit projection (Feldman and Weinshall, 2005).

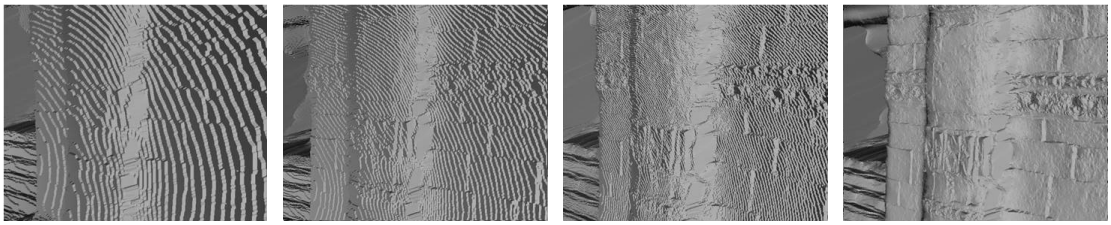
There are three advantages of using this line scan cameras for stereo imaging. First, we can acquire high resolution images because it captures an image by stitching image columns from a line scan camera. The maximum resolution of the image provided by the line-scan camera used in this work is 12574×5658 which is sufficient to capture the full environment with high resolution details. Second, the stereo matching can be simplified to a 1D search along the vertical scan line as discussed above, while normal spherical images require a complex search along conic curves or rectification of the images. Finally, a relatively simple calibration is required. Depth reconstruction only requires knowledge of the baseline distance between the stereo image pair and correction of radial distortion in the vertical direction. Radial distortion is rectified using a 1D lookup table to evenly map pixels on the vertical central line to the $[0, \pi]$ range. Lens distortion parameters are static values so that this mapping can be calculated for the lens in advance. Figure 2 (a) shows an example of a vertical stereo pair captured by the line scan camera system.

To define the disparity between spherical stereo image pairs, let us focus on an epipolar plane which is defined by a 3D point and the two camera positions, as shown in Fig. 2 (b). If we assume the angles of the projection of the point P onto the spherical image pair displaced along the y-axis are θ_t and θ_b, respectively, the angle disparity d of point $p_t(x_t, y_t)$ can be defined as the difference of the angles as:

$$d(p_t) = \theta_t - \theta_b = (y_t - y_b) \times \pi / h \tag{1}$$

where h is the image height in pixels, y_t and y_b are y-coordinates of the projection points p_t and p_b, respectively. The distance of the scene point P from the two cameras is calculated by triangulation as illustrated in Fig. 2 (b). If B is the baseline distance between the camera centers of projection and r_t and r_b are the distance

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



(a) (b) (c) (d)

Fig. 3 Precision in surface reconstruction: (a)Integer disparity; (b)Half-pixel disparity; (c)Quarter-pixel disparity; (d)Floating-point disparity

from P to the top and bottom cameras, respectively, then:

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

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

4 PDE-based Disparity Estimation

Disparity estimation is one of the most important steps in image-based 3D reconstruction. Scharstein and Szeliski (2002) present a taxonomy of existing stereo algorithms together with a benchmarking site for their quantitative evaluation. Most disparity estimation algorithms, including recent approaches based on graph-cut (GC) (Kolmogorov and Zabih, 2001) and belief-propagation (BP) (Sun et al., 2003; Yang et al., 2008), 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.

Spherical stereo image pairs typically have relatively small variations in disparity and strong radial distortion because of the wide FOV of the fisheye lens. The narrow baseline (<1m) relative to the scene depth (>10m) also results in small disparity range for scene reconstruction. Figure 3 shows the difference in surface reconstructions from discrete and floating-point disparity fields for the "Gate" dataset used for evaluation in Section 6.3. For discrete disparity estimation in Fig. 3 (a)-(c), there is a loss of surface detail and stepwise artifacts with depth quantisation. These examples are reconstructed from high resolution spherical images of 12574×5658. The quantisation artifact will be even more obvious with lower resolution images. This artifact can be reduced by increasing the number of depth layers, e.g. calculating in 1/8 pixel levels, but this drastically increases memory consumption and computational cost due to increased search range. Variational approaches (Kim and Sohn, 2003b; Alvarez et al., 2002; Slesareva et al., 2005) which calculate disparity on a continuous domain can be a solution to avoid quantisation artifacts. They allow optimisation of stereo disparity on a continuous domain as seen in Fig. 3 (d). Variational approaches are still limited in accuracy of disparity estimation due to the resolution of original images (Szeliski and Scharstein, 2004) and also require a discrete step size for solver implementation. The step size can be adaptively refined to evaluate a smooth disparity field with fine surface details without an additional consumption of memory.

4.1 Background and problem definition

In variational approaches, the disparity vector fields are extracted by minimizing an energy functional involving a fidelity term $E_f(\cdot)$ and a smoothing term $E_s(\cdot)$ such as:

$$\begin{aligned} E(d_t) &= E_f(d_t) + E_s(d_t) \\ &= \lambda \int_\Omega \left(I_t(p) - I_b(p + d_t)\right)^2 dx \\ &\quad + \int_\Omega \Psi(\nabla d_t, \nabla I_t) dx \end{aligned} \quad (3)$$

where λ is a weight from the fidelity term, $p \in \Omega$ is an open bounded set of R^2, $I(p)$ is a pixel value of the point p, d_t is a 2D disparity vector, and $\nabla := (\partial x, \partial y)^T$ denotes a spatial gradient operator. If we set the gradient of the potential function $\Psi(\nabla d_t, \nabla I_t)$ as Eq. (4), the minimisation problem can be solved by the associated Euler-Lagrange equation in Eq. (5) with Neumann boundary conditions (Alvarez et al., 2002).

$$\nabla(\Psi(\nabla d_t, \nabla I_t)) = g(\nabla I_t, \nabla d_t)\nabla d_t \quad (4)$$

$$\begin{aligned} -\nabla E(d_t) = div(g(\nabla I_t, \nabla d_t)\nabla d_t) \\ +\lambda(I_t(p) - I_b(p + d_t))\frac{\partial I_b(p + d_t)}{\partial d} = 0 \end{aligned} \quad (5)$$

We obtain the solution of Eq. (5) by calculating the asymptotic state ($t \to \infty$) of the PDE:

$$\frac{\partial d}{\partial t} = div(g(\nabla I_t, \nabla d_t)\nabla d_t) + \lambda(I_t(p) - I_b(p + d_t))\frac{\partial I_b(p + d_t)}{\partial d} \quad (6)$$

This PDE corresponds to the nonlinear diffusion equation with an additional reaction term (Weickert, 1997), and $g(\cdot)$ is a diffusion tensor which controls the direction and amount of diffusion filtering.

Design of the energy function which keeps smooth disparity fields for continuous surfaces while preserving sharp object boundaries has been the subject of extensive research. In this section, we deal with three problems in the PDE-based disparity estimation.

The first problem is over-segmentation in highly textured regions. Many researchers have focused on producing sharp depth discontinuities because the diffusion filtering tends to over-smooth object boundaries (Slesareva et al., 2005; Zimmer et al., 2008). Image gradient has been widely used to control the diffusion (Alvarez et al., 2002; Kim and Sohn, 2003b), but this also affects diffusion in planar regions with textured appearance because it is hard to differentiate between high gradient due to discontinuities and changes in appearance due to texture or shadows. Sun et al. (2008) proposed a joint image/flow-driven optical flow to obtain sharp object boundaries without over-segmentation. In Section 4.2, we take Sun et al.'s idea as inspiration to design a new diffusivity function controlled by a combination of image and disparity gradients.

The second problem is due to stereo occlusion. Variational methods are popular for estimating optical flow where displacements are relatively small between images so that the occlusion is negligible. However, stereo images generally include large displacements and the occluded regions result in distortions of the disparity field because there is no valid correspondence between the image pair. An occlusion detector based on bi-directional matching is added to the PDE functional in order to control the balance of fidelity and smoothing terms in Section 4.3.

Finally, local minima and computational complexity are serious problems in disparity estimation for high resolution images. The maximum disparity in the stereo pairs used in this work is 320 pixel with a resolution of 12574×5658. The PDE-based method cannot converge to the optimal solution with such a large displacement. Moreover, the computational load of PDE-based methods depend on the image size because of the iterative solver. To overcome these problems, we propose a multi-resolution approach using successive estimates at lower resolution to initialise successively higher resolutions. This approach is presented in Section 4.4.

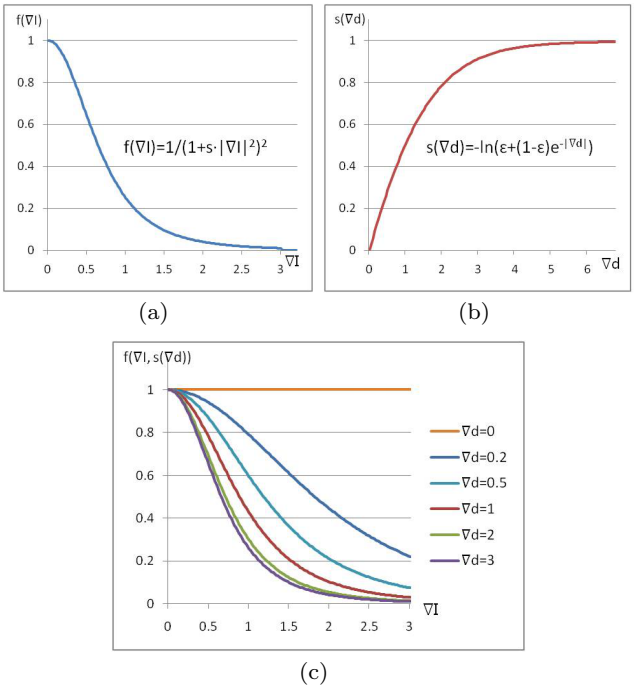


Fig. 4 Behaviour of diffusivity function (Eq. (7)): (a)Geman and McClure's function for s=1; (b)$s(\nabla d)$ for $\varepsilon = 1/e$; (c)Final diffusivity function

4.2 Diffusion tensor for highly textured region

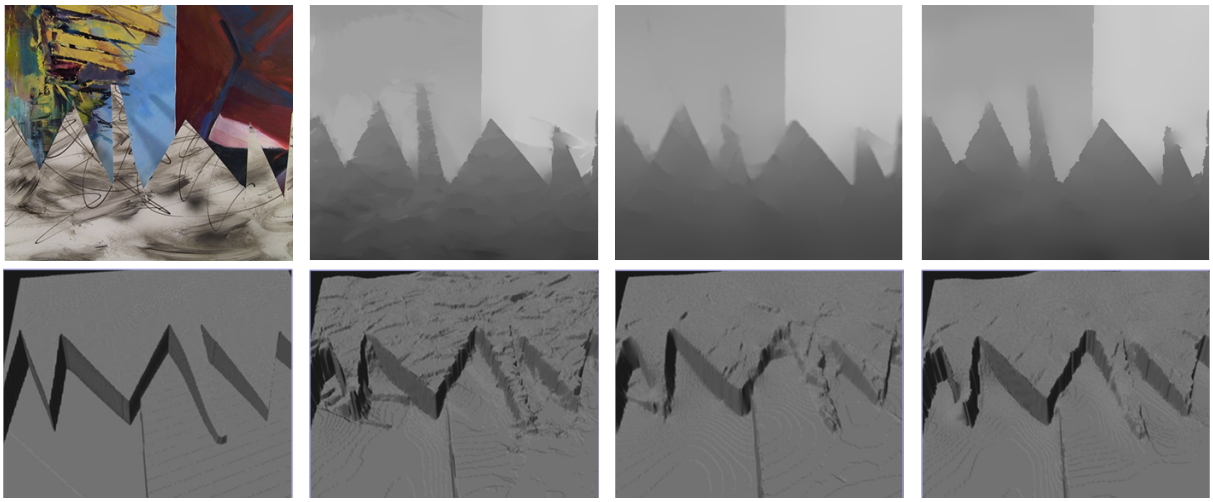
Our approach modifies Nagel and Enkelmann (1986)'s anisotropic diffusion tensor and Geman and McClure (1985)'s diffusivity function to define a new diffusion tensor controlled by image and disparity gradients which handles the over-segmentation problem while preserving sharp object boundaries as follows.

$$g(\nabla I, \nabla d) = f(\nabla I, s(\nabla d))(\nabla I \nabla I^T + L) \quad (7)$$

$$f(\nabla I, s(\nabla d)) = \frac{1}{(1 + s(\nabla d)|\nabla I|^2)^2} \quad (8)$$

$$s(\nabla d) = -ln(\varepsilon + (1 - \varepsilon) \cdot e^{-|\nabla d|}) \quad (9)$$

In Eq. (7), L denotes the identity matrix and the term $\nabla I \nabla I^T$ is the structure tensor of Nagel and Enkelmann's method for anisotropic diffusion filtering. Equation (8) is a form of Geman and McClure's diffusivity function which behaves as illustrated in Fig. 4 (a) when s=1. This is modified with Eq. (9) so that it is also scaled by the gradient of the disparity field. Equation (9) is a monotonically increasing function which converges to 1 for $\varepsilon = 1/e$ as shown in Fig. 4 (b). As a result, the filtering direction is decided by the image gradient ∇I, and the amount of smoothing is decided



(a) (b) (c) (d)

Fig. 5 Comparison of disparity estimation for different formulations of the diffusion tensor (top: original image and disparity maps, bottom: rendering of reconstructed model from upper direction): (a)Ground truth; (b)Image-driven method; (c)Disparity-driven method; (d)Proposed method

by both image and disparity gradients (∇I and ∇d) as shown in Fig. 4 (c). In regions with high disparity gradient, the diffusivity function is more sensitive to the image gradient because the image gradient is more likely to be a depth discontinuity. In the region with low disparity gradient, on the contrary, more smoothing is performed regardless of image gradient because the image gradient has a higher possibility to be texture. In order for Eq.(6) to be an energy gradient, we regard ∇d in Eq. (9) as a constant ∇d^0 for each iteration step in the PDE linearisation.

Figure 5 shows results of different diffusivity functions. The original image in Fig. 5 (a) consists of three planes with strong textures. We simulated a Geman and McClure's diffusivity function with isotropic diffusion (Kim and Sohn, 2003b) for the image-driven variational method, and Zimmer et al. (2008)'s algorithm for the disparity-driven method. The image-driven method Fig. 5 (b) results in incorrect discontinuities on the planar surface due to the high image gradient resulting from the texture pattern. On the other hand, the disparity-driven method Fig. 5 (c) produces smooth surfaces, but we can see diffusion of the fields at object boundaries. The proposed approach Fig. 5 (d) overcomes the limitations of both image-based and disparity-based methods producing a smooth disparity field on the planar surface whilst correctly preserving discontinuities at object boundaries.

4.3 Occlusion handling

Occlusion in stereo imaging results in areas which are visible from one view but not the other due to the displacement of the viewing position. In the occluded areas, there is no valid correspondence between the image points and many stereo matching algorithms find the most similar image regions (Scharstein and Szeliski, 2002). The occlusion problem can be negligible in some application areas but it is a critical problem in 3D reconstruction because forcing false matching in occluded regions induces distortion in the depth fields.

In variational stereo methods, some researchers have designed energy functionals which handle occlusion regions and converge to a minimum solution. Strecha et al. (2004) proposed to compute visibility of pixels to avoid this problem. The visibility is modelled as a mixture problem by introducing a set of hidden variables which are sequentially updated in the EM algorithm. Similarly, Ben-Ari and Sochen (2007) detected occlusions using a level-set method and performed disparity estimation only for unoccluded regions. Alvarez et al. (2007) proposed a symmetrical dense optical flow energy functional which modifies Eq. (3) by adding a bi-directional disparity checking term. Ince and Konrad (2008) also proposed a similar bi-directional disparity checking method, but they put the bi-directional matching penalty in the fidelity term.

In this work, we take a similar approach to Ince and Konrad (2008)'s work for occlusion handling, but make it simpler by adding the bi-directional matching function as a weighting factor of the fidelity term in the iterative solver. In order to penalise the fidelity term in

occluded regions, we change the weighting factor λ of Eq. (6) to a function of bi-directional disparity matching as follows:

$$\frac{\partial d_t}{\partial t} = div(g(\nabla I_t, \nabla d_t)\nabla d_t(p)) + h\left(|d_t^0(p) + d_b^0(p + d_t^0)|\right) \cdot (I_t(p) - I_b(p + d_t))\frac{\partial I_b(p + d_t)}{\partial d} \quad (10)$$

$$h(x) = \frac{\lambda_1}{(1 + x^2)^2} \quad (11)$$

Equation (11) is the monotonically decreasing function proposed by Geman and McClure (1985), illustrated in Fig. 4 (a). In visible regions, the fidelity and smoothing terms are balanced to find the optimal solution. In occluded regions, anisotropic diffusion filtering increases smoothness of the disparity field and propagates reliable depth information from visible regions to occluded regions.

To solve Eq. (10), we discretise the parabolic system by finite differences and the computationally expensive solution of the nonlinear system is avoided by using a first-order Taylor series expansion in an implicit discretisation:

$$I(p + d^{k+1}) \approx I(p + d^k) + (d^{k+1} - d^k)\frac{\partial I(p + d^k)}{\partial p} + e^k(p + d^k) \quad (12)$$

Finally, the regularized disparity field can be found in a recursive manner by updating the discretised field of Eq. (10) (Johnson, 1988).

Figure 6 shows a comparison of reconstruction results with and without occlusion handling. Figure 6 (a) is parts of the "Gate" images used for evaluation in Section 6.3. The black regions in occlusion map represent regions where the bi-directional matching error in Eq. (10) is larger than 1 pixel. Fig. 6 (b) and (c) are parts of the estimated disparity fields and reconstructed model without and with the occlusion handling term, respectively. This example demonstrates that the disparity field without the occlusion handling is blurred and corrupts the structure boundaries, while the reconstruction with the proposed occlusion handling produces sharper depth discontinuities with clear boundaries.

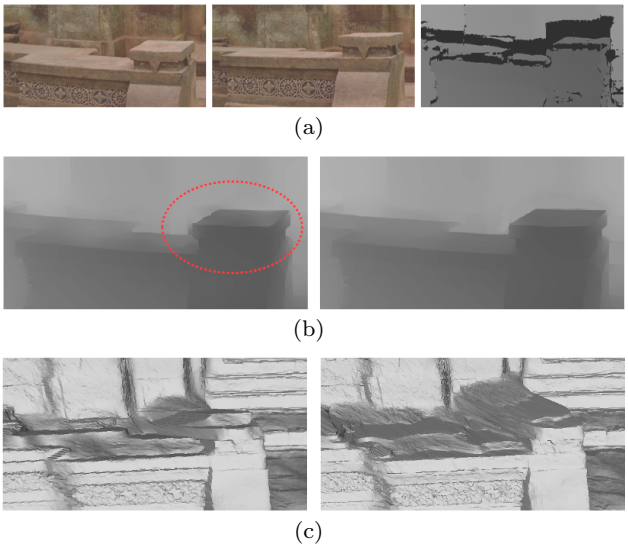


Fig. 6 Comparison of stereo reconstruction without and with occlusion handling: (a)Stereo pair (Left: top image, Centre: bottom image, Right: occlusion map); (b)Estimated disparity fields d_t (Left: without occlusion handling, Right: with occlusion handling); (c)Reconstructed model (Left: without occlusion handling, Right: with occlusion handling)

4.4 Hierarchical approach and initial estimation

In general, the computational load for the iterative solver and convergence to local minima are the most serious problems in variational methods because the energy functional in Eq. (3) can be non-convex due to its fidelity term. Alvarez et al. (2002) used a scale-space approach and Brox et al. (2004) used a warping method to avoid convergence to local minima for large displacements.

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. Multi-resolution images are expanded using a Gaussian pyramid (Burt, 1981) to construct a G-level hierarchical image structure, which involves low-pass filtering and down-sampling the image by a factor of 2. At each level, the input disparity field from the previous level is up-sampled and used as an initial field for calculating the disparity field at that level. This hierarchical approach also has the merit of reducing the computation time for large images. The hierarchical approach with G=4 has been found to give an order of 8 reduction in computational time over single resolution disparity estimation.

However, variational methods implemented with a hierarchical structure, in general, have difficulty in reconstructing small details with strong depth discontinuities when initialized from a constant depth, since the required image details are lost at coarser levels of the pyramid. Therefore the number of hierarchy should be carefully chosen according to the characteristics of scenes so that it can avoid local minimum while keeping small details. In practice we restricted the maximum level to G=4 for 12574×5658 spherical images and to

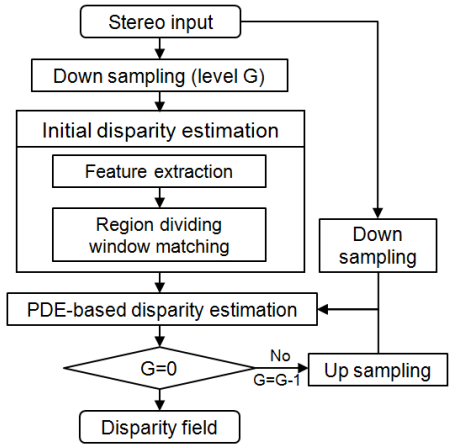


Fig. 7 Block diagram for disparity estimation

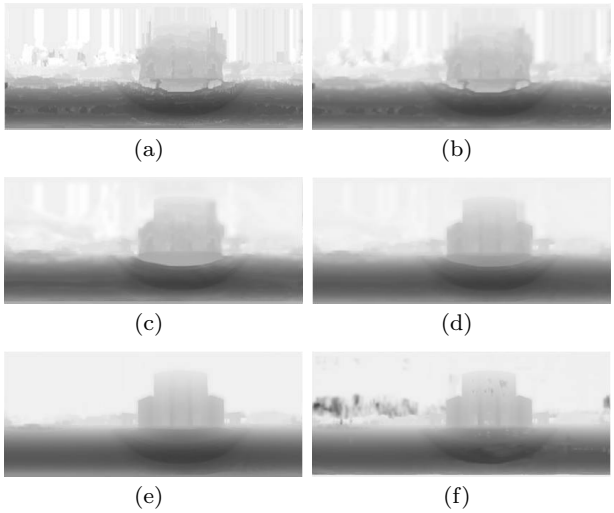


Fig. 8 Hierarchical disparity estimation results at each step: (a)Initial disparity (786×354); (b)G=3 (786×354); (c)G=2 (1571×708); (d)G=1 (3142×1415); (e)Final disparity (G=0, 6284×2829); (f)Non-hierarchical result

G=3 for 6284×2829 images (which means the coarsest image has a resolution of 786×354) and provide an initial disparity field for the coarsest level. A magnitude-extended Laplacian pyramid can also be a solution to avoid loss of small objects (Sizintsev, 2008).

The initial disparity field is generated by window-based matching with a region-dividing technique (Kim and Sohn, 2003a) based on the ordering constraint (Yuille and Poggio, 1984). The ordering constraint states that if an object A appears to the left of an object B in the left image, then object A will also appear to the left of object B in the right image. The ordering constraint can be violated when a thin object is placed close to the cameras. In practice the constraint is rarely violated for vertical stereo pairs of environments. The approach performs point matching in order of the possibility of correct matching (measured by magnitude of a Sobel edge detector) and divides the region into sub-regions at the true matching point. After the region splits into two sub-regions, the search ranges of the points in each sub-region are restricted to the corresponding sub-region. For matching criteria, we found that the Mean Absolute Error (MAE) works better for the images from controlled environment such as indoor or synthetic images with controlled illuminations, and the Normalised Cross Correlation (NCC) is better for outdoor scenes where the brightness may change, as observed by Hirschmüller and Scharstein (2008).

Figure 7 shows a block diagram and Table 2 is a list of parameters for the whole disparity estimation process. If λ_1 is too large, the PDE solver diverges. On the other hand, if λ_1 is too small, disparity fields are over-smoothed. Gradient and time step sizes are used for descritisation of the PDE solver (Johnson, 1988). They are related to the convergence and speed of the PDE solver. They were experimentally decided but fixed to these values for all experiments.

Table 2 Parameters for disparity estimation

Parameter	Values
Level of hierarchy	G=3
Window size	W=16x16
Weight for fidelity term	λ_1 = 0.0005
Gradient step size	δ_I=3 / δ_d=1
Time step size	τ=0.0001

Figure 8 shows an example of initialisation and refinements at each level of the Cathedral scene in Fig. Fig. 2 (a). The initial estimation by the region-dividing technique provides a rough initial disparity field at the lowest resolution (Fig. 8 (a)), which is refined to the optimal solution with the proposed method for increasing resolution in the hierarchy (Fig. 8 (b)-(e)).

Results are compared with non-hierarchical processing. The proposed algorithm failed to find disparity fields for full resolution images without initial disparity estimation because variational methods cannot be directly applied to large displacements (Alvarez et al., 2002; Brox et al., 2004). The maximum disparity of the test images is 120 pixels. Initialising the non-hierarchical implementation of the proposed algorithm with the initial disparity field results in convergence to local minima around regions with erroneous initial disparity fields as shown in Fig. 8 (f). Application of the proposed hierarchical approach correctly recovers disparity even with errors in the initialisation as shown in Fig. 8 (e). We also compared processing times of the algorithms on a normal PC. It took around 6 hours for the resolution of 6284×2829 images with NCC-based initial disparity estimation, while it took less than an hour with a three level (G=3) hierarchy and produced more accu-

rate results. This demonstrates that the proposed hierarchical approach can improve both computational efficiency and convergence for regions with large disparity overcoming the problem of convergence to local minima.

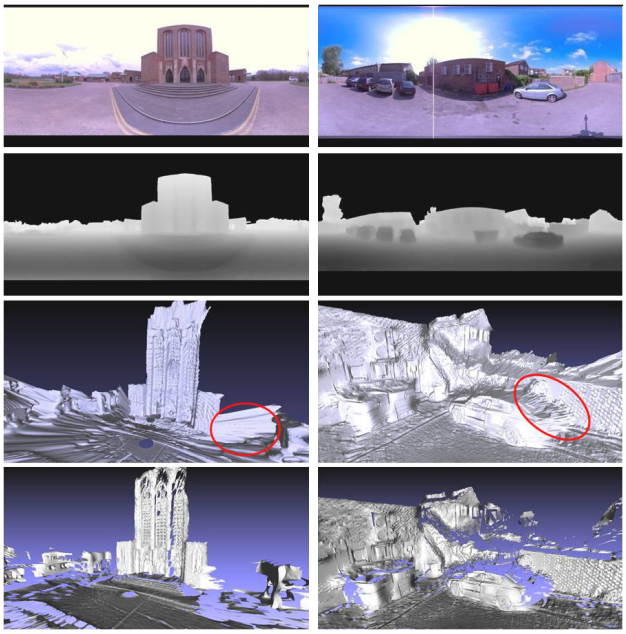
5 3D Model Reconstruction

5.1 Single stereo pair reconstruction

The estimated dense disparity fields can be converted into depth information by the spherical stereo geometry as described in Section 3. A mesh model of the scene is then obtained by sampling vertices as a M-pixel grid and triangulating adjacent vertices from the original texture and depth information. The triangulation is generated as a regular mesh grid on 2D plane and the mesh grid is projected to 3D space to create a 3D mesh model. The vertex points are described in spherical coordinates, so we convert them into the Cartesian coordinate system.

Figure 9 shows examples of reconstruction from spherical stereo image pairs. Most of the sky regions are automatically removed because they have zero disparity which means infinite distance. However, the sky regions sometimes show visually annoying artefacts in reconstruction because of the lack of texture or moving clouds. Pollefeys et al. (2008) proposed an automatic sky removal algorithm using learning colour distribution of the sky offline via k-means clustering but it still requires manual refinement in various cases. Powerful learning and recognition algorithms such as SuperParsing (Tighe and Lazebnik, 2010) can improve the performance but we simply manually removed the sky regions in this research.

The results show a natural-looking environment around the captured location. Changes in the viewpoints result in distortion of the scene because of self-occlusion from the centre of projection of the spherical image as seen in the circled regions in the third row of Fig. 9. This erroneous surface is generated to extract an estimate of a surface by extrapolation given no observation. We define this extrapolated surface around a depth discontinuity region as a false surface. There is no way to get information about occluded regions behind any object seen from a single location. This occlusion problem occurs not only between objects but even on the same object at step discontinuities. False surfaces occur along a radial direction from the camera centre of projection and can be removed by evaluating the angle between the surface normal and direction to the centre of projection in the disparity field. However, due to the inherent smoothing of the disparity field in the PDE solution false surface may not be orthogonal and a threshold is required to remove them. The fourth row of Fig. 9 shows the surface refinement with the fixed angle threshold of 83°. It failed to clearly remove false surfaces and also damaged other true surfaces such as the ground or details on the surfaces.



(a) Cathedral (b) Car park

Fig. 9 Reconstruction from a single spherical stereo pair (Top row: Captured spherical image, Second row: Estimated disparity field mapped to grey scale Third row: Reconstructed mesh model, Fourth row: Surface refinement with normal vectors)

In this section, we address two problems of the single stereo pair reconstruction: 1) How to overcome the occlusion problem; 2) How to remove false surfaces and retrieve true surface. In order to solve the occlusion problem, we need more information about the scene structure, shape and appearance from multiple viewpoints. Merging reconstructions from multiple stereo pairs into a common 3D scene structure is a possible solution. However, integrating multiple surfaces raises a problem of surface overlap and requires a surface refinement process because the overlapping surfaces may include false surface or inaccurately reconstructed surfaces which have lower reliability. In the following subsections, we provide a brief survey of existing surface registration and fusion algorithms, and propose a novel technique to generate a complete mesh structure of the scene from multiple meshes.

5.2 Integration of reconstructions from multiple spherical stereo pairs

A number of multiple-view reconstruction algorithms have been developed as listed on Strechas benchmarking site (Strecha et al., 2008) and the Middlebury multi-view reconstruction benchmarking site[4]. However, these approaches require calibration for each capture point. Existing methods introduced in Section 2.1 and 2.3 generally use wide-baseline feature matching to recover camera parameters and reconstruct sparse 3D point clouds from a large number of images. Our approach is advantageous because a dense structured mesh is reconstructed from narrow-baseline stereo pairs with simple calibration steps and it requires a relatively small number of images.

5.2.1 Mesh registration

As addressed in Section 2.3, the ICP algorithm can provide a solution to extract a 3D rigid transform between reconstructed meshes from different capture points. The algorithm selects the closest points as correspondences and calculates the rigid transformation, i.e., rotation and translation (R,t), minimising the energy:

$$E_R(R,t) = \sum_{i}^{N_m} \sum_{j}^{N_d} w_{i,j} \|m_i - (Rd_j + t)\|^2 \qquad (13)$$

where N_m and N_d are the number of points in the model set m and reference set d, respectively, and $w_{i,j}$ are the weights for a point match.

In order to automate the initialisation of the ICP registration, we use SURF feature matching (Bay et al., 2008) 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 reliable enough to be used in ICP registration if the capture points are far from each other because two possible sources of errors exist: errors in SURF matching between widely spaced image pairs with radial distortion; and errors in the reconstructed depth from a single narrow-baseline spherical stereo image pair.

A robust wide-baseline registration algorithm is therefore proposed using RANSAC (Fischler and Bolles, 1982) to calculate an initial 3D rigid transform and eliminate outliers in SURF matching between pairs of reconstructions from spherical stereo images.

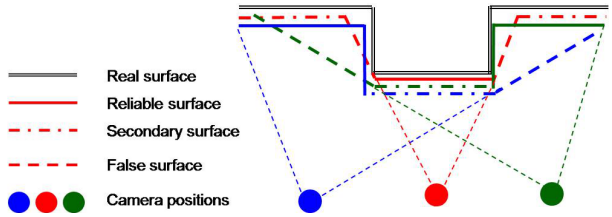
[4] Middlebury multi-view, `http://vision.middlebury.edu/mview/`

5.2.2 Reliable surface extraction

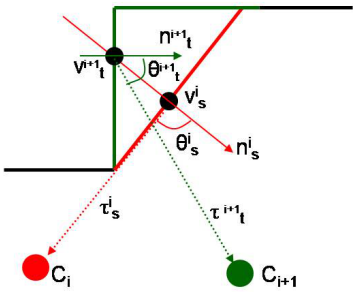
The final step is to refine registered surface estimates from spherical stereo pair reconstructions in overlapping regions based on reliability. Poisson reconstruction (Kazhdan et al., 2006), depth map or range image merging (Gargallo and Sturm, 1988) and mesh fusion (Turk and Levoy, 1994; Hilton et al., 1998) have previously been proposed as methods to produce a single mesh structure from a set of oriented points, multiple depth fields, or partial meshes. However, in the presence of measurement errors such as differences in sampling and errors in registration, loss of details on the original surface may occur because the algorithms generate combined surface from overlapping regions. False surfaces from self-occlusion may also introduce errors in integration. Furukawa et al. (2009) proposed an optimized surface boundary extraction for self-occlusion using axis alignment for Manhattan-world scenes. However, this approach also results in loss of surface detail due to the planar surface approximation. Therefore we propose a surface selection algorithm to select the most reliable surface in overlapping regions of surface reconstruction.

The most similar approach to our algorithm is the visibility-based depth-map fusion approach proposed by Merrell et al. (2007). This approach renders depth maps of neighbouring views into a reference view and generates a reliable depth by considering stability and confidence of reconstructed depth layers. The stability is measured by occlusion and free-space violation, and the confidence is calculated from errors in the plane-sweep stereo reconstruction. In this work, we use angles of the surface normal vector and distance to the camera centre of projection as a measure of reliability instead of the free-space assumption and stereo disparity matching error used in Merrell's work. This approach is advantageous in finding the relationship between vertices and the camera centre in the spherical imaging system because the vertices are generated from a narrow baseline vertical image pair so the membership of vertices from each capture points are clear, while plane-sweeping reconstructs vertices from sparse images sets.

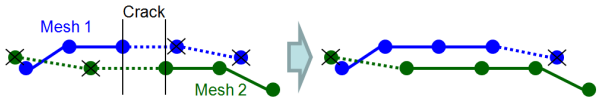
Figure 10 (a) shows an illustration of a real surface and overlapped surfaces reconstructed from three camera pairs. The overlapped surfaces include false surfaces from self-occlusion and less reliable (secondary) surfaces. We assume that the reconstructed surface is more reliable when the surface normal vector and camera viewing direction are aligned, and the distance to the camera is closer. Our approach is to choose the most reliable surface in any region and discard all false and less reliable overlapping surface reconstructions.



(a) Overlapped surfaces from 3 cameras



(b) Visibility constraint violation



(c) Morphological dilation process

Fig. 10 Reliable surface extraction

Figure 10 (b) shows notations used for evaluating reliabilities of overlapping surfaces. Let us assume we have multiple vertical stereo captures at positions $C = \{C_0, \dots C_i, \dots\}$ and corresponding reconstructed mesh models for each pair $M_i = \{V_i, N_i, T_i\}$ composed of vertices $V_i = \{v_0^i, \dots, v_s^i, \dots\}$, vertex normal vectors $N_i = \{n_0^i, \dots, n_s^i, \dots\}$ and faces $T_i = \{t_0^i, \dots, t_s^i, \dots\}$. The vertex normal vector n_s^i of the s-th vertex on the i-th reconstructed mesh is calculated by averaging the normal vectors of neighbouring faces within the range of R from the vertex v_s^i to get a global surface normal direction regardless of the fine details on the surface. Then the projection vector τ_s^i and facing angle θ_s^i are expressed as follows:

$$\tau_s^i = C_i - v_s^i \tag{14}$$

$$\theta_s^i = arccos\left(\frac{c_s^i \cdot n_s^i}{|c_s^i||n_s^i|}\right) \tag{15}$$

In order to find overlapping surfaces, the visibility constraint (Hilton, 2005; Merrell et al., 2007; Furukawa et al., 2009) is applied on all vertices. The visibility test is generally used for occlusion detection, but we use it to identify surface overlaps. Conflicted surfaces from the vertex v_s^i are searched along the normal vector n_s^i in the range of Th_R. If we assume a set of overlapping vertices $V_o = \{v_s^i\}$ from meshes $M = \{M_i\}$ is detected by the test, we calculate $\tau_o = \{\tau_s^i\}$ and $\theta_o = \{\theta_s^i\}$ for the set V_o. A measurement of reliability $U(v)$ of each

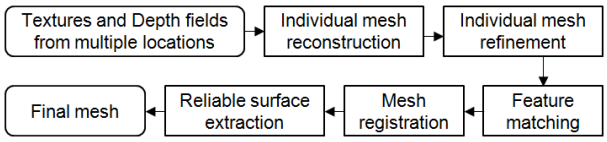
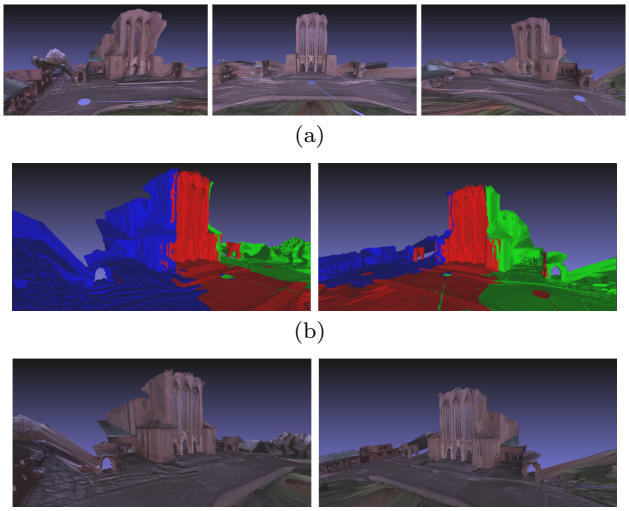


Fig. 11 Block diagram for mesh registration



(a)

(b)

(c)

Fig. 12 Results of mesh fusion: (a)Single view reconstructions from three different points (left, centre, right); (b)Integrated mesh (Blue: left, Red: centre, Green: right); (c)Rendering with textures

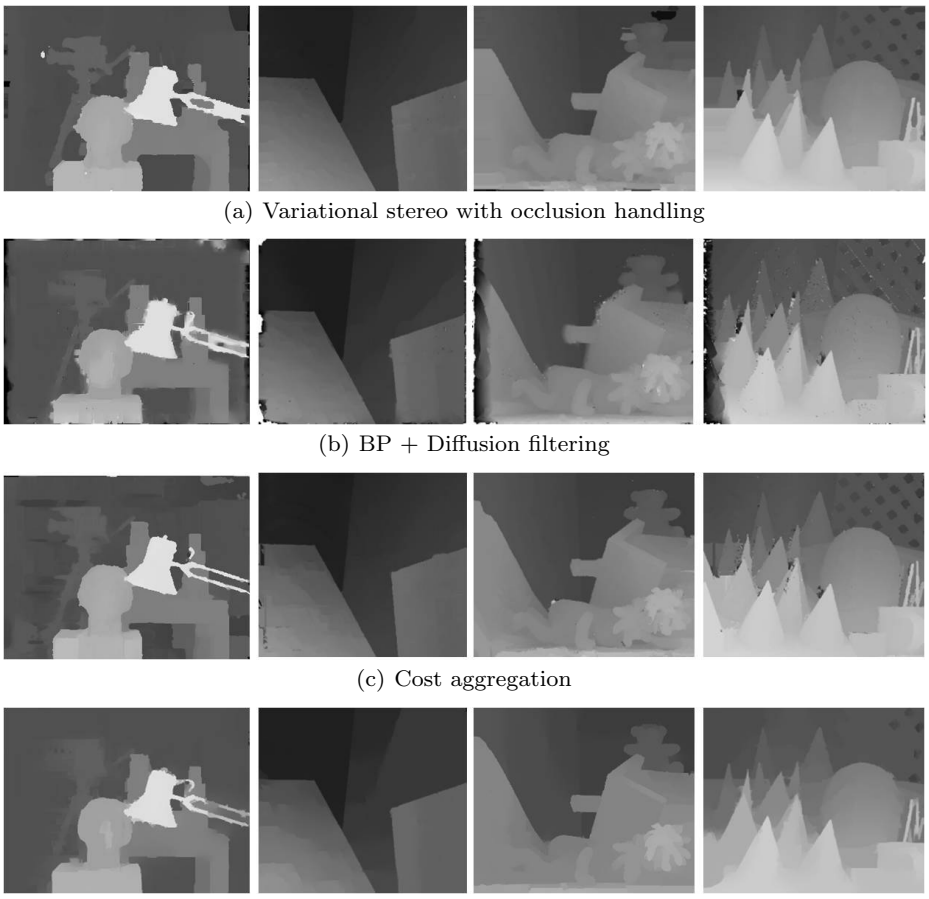
vertex is calculated based on the facing angle and the distance from the camera centre using Eq. (16), and the most reliable vertices are selected.

$$\begin{aligned}\overline{V} &= \underset{v_s^i \in V_o}{\text{argmax}}\, U(v_s^i) \\ &= \underset{v_s^i \in V_o}{\text{argmax}}\, \left(U_d(v_s^i) + \lambda_2 U_\theta(v_s^i)\right)\end{aligned} \tag{16}$$

$$U_d(v_s^i) = \begin{cases} d_{max}/|\tau_s^i|, & \text{if } |\tau_s^i| < d_{max} \\ 0, & \text{else} \end{cases} \tag{17}$$

$$U_\theta(v_s^i) = \left|\frac{\pi/2}{\theta_s^i}\right|, \quad \left(-\pi/2 < \theta_s^i < \pi/2\right) \tag{18}$$

Application of this algorithm for all individual vertices is a time-/memory-consuming process and produces erroneous results with small isolated surface details. Therefore we down-sample the vertices with averaging normal vectors for neighbours in the radius of R and perform the above reliable vertex selection. Once all false and secondary vertices are removed, the removed vertices at the border with reliable surfaces are recovered as an morphological dilation process in order to cover visible cracks as shown in Fig. 10 (c). Finally the original dense mesh is recovered from the remaining down-sampled vertices. We do not merge or



(a) Variational stereo with occlusion handling

(b) BP + Diffusion filtering

(c) Cost aggregation

(d) Proposed algorithm

Fig. 13 Estimated disparity maps (from left to right: Tsukuba, Venus, Teddy and Cones)

re-triangulate vertices from different capture points to keep surface details. Mesh zipping method (Turk and Levoy, 1994) can be applied for overlapped vertices if a manifold mesh structure is required.

Figure 11 shows the whole process for multiple mesh registration and Fig. 12 shows an example of the mesh registration with the surface selection. Figure 12 (a) shows single viewpoint reconstructions from three different viewpoints for the same scene. Surface distortions are observed for the surface parallel to the viewing direction as well as errors from occlusions. Mesh integration is performed for the three incomplete meshes with the heuristically decided parameter sets: M=8, λ_2=1.0 R=10cm, d_{max}=30m and Th_R=70cm. The contribution of each mesh to the final merged mesh is shown in Fig. 12 (b). The blue part is from the left viewpoint, the red is from the central view, and the green is from the right view. Figure 12 (c) show the final rendering with texture from corresponding images.

6 Experimental Results

All test scenes presented in this section were captured with a Spheron commercial line scan camera introduced in Section 3. We attached a Nikon 16mm f/2.8 AF fisheye lenses to the system and captured vertical stereo pairs with a baseline of 60cm. The maximum resolution of images is 12574×5658, and we use full resolution images for single stereo pair reconstruction, and half resolution pairs for multiple stereo reconstructions due to restrictions of memory. In order to demonstrate the general performance of the proposed disparity estimation algorithm, we first show the results with narrow baseline stereo images from the Middlebury benchmarking site. Then we present mesh models reconstructed from stereo image pairs captured with the line scan camera. Reconstructions are evaluated against ground-truth range data scanned with a LIDAR sensor. We also analyse the characteristics and behaviour of errors for spherical stereo imaging. Finally we compare the results of mesh fusion with other MVS and SfM-based methods, and show virtual view rendering results of reconstructed models from arbitrary viewpoints.

Algorithm	Avg. Rank	Tsukuba ground truth nonocc	Tsukuba all	Tsukuba disc	Venus ground truth nonocc	Venus all	Venus disc	Teddy ground truth nonocc	Teddy all	Teddy disc	Cones ground truth nonocc	Cones all	Cones disc	Average Percent Bad Pixels
AdaptingBP [17]	4.7	1.11 9	1.37 5	5.79 11	**0.10** 1	0.21 3	1.44 2	4.22 4	7.06 5	11.8 5	2.48 2	7.92 8	7.32 3	4.23
CoopRegion [41]	4.8	**0.87** 1	**1.16** 1	**4.61** 1	0.11 2	0.21 2	1.54 4	5.16 11	8.31 7	13.0 8	2.79 7	7.18 3	8.01 10	4.41
DoubleBP [35]	6.7	0.88 3	1.29 2	4.76 3	0.13 8	0.45 14	1.87 9	3.53 3	8.30 6	9.63 2	2.90 9	8.78 17	7.79 7	4.19
.....														
GradAdaptWgt [60]	32.5	2.26 49	2.63 37	8.99 42	0.99 47	1.39 48	4.92 47	8.00 41	13.1 31	18.6 46	2.61 5	7.67 4	7.43 4	6.55
AdaptWeight [12]	33.8	1.38 26	1.85 24	6.90 28	0.71 41	1.19 44	6.13 43	7.88 38	13.3 36	18.6 47	3.97 36	9.79 31	8.26 13	6.67
YOUR METHOD	34.0	2.85 66	3.71 61	15.1 70	0.14 9	0.35 13	2.00 12	8.48 60	13.2 33	17.3 38	3.65 29	9.07 22	10.2 31	7.18
SegTreeDP [22]	34.3	2.21 47	2.76 39	10.3 51	0.46 34	0.60 22	2.44 17	9.58 68	15.2 58	18.4 43	3.23 19	7.86 6	8.83 28	6.82

Fig. 14 Middlebury ranking with 1.0 pixel threshold

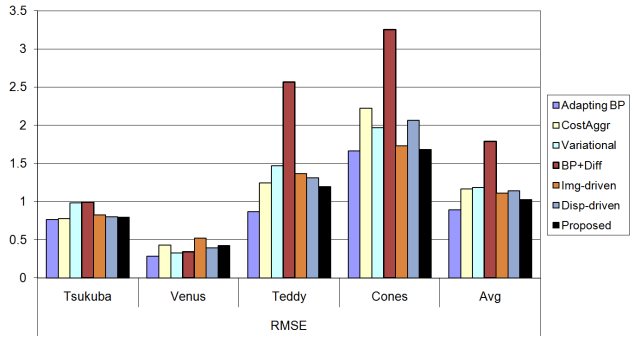


Fig. 15 Comparison of RMSE

6.1 Evaluation of disparity estimation

For evaluation of the proposed PDE-based disparity estimation algorithm, we used the Middlebury stereo benchmarking test bed[5]. Figure 13 shows subjective comparison of estimated disparity maps with state-of-the-art methods. We present comparison with the following methods: Ben-Ari and Sochen (2007)'s variational method which was introduced in Section 4.3; BP+diffusion (Banno and Ikeuchi, 2009) approach which is not a variational method but uses anisotropic diffusion filtering after belief-propagation to make smooth 3D surface reconstructions; and Min and Sohn (2008)'s local cost aggregation method using anisotropic diffusion and occlusion handling. The test bed does not include many variational methods because most variational methods cannot estimate large displacements. Multi-resolution scheme would allow to overcome this problem, but such methods are not available in the test bed. Comparison to the proposed method demonstrates that the proposed method produces smoother maps with sharp object boundaries even in occluded regions. The proposed algorithm is ranked at 34 in the Bad Pixel Percentage (BPP) test among 90 algorithms on the test bed as shown in Fig. 14. The approach shows a relatively good performance for simple scenes like Venus and Cones. The variational stereo, BP+diffusion and cost aggregation are ranked at 47, 39, and 28, respectively.

[5] Middlebury stereo, `http://vision.middlebury.edu/stereo/`

The proposed method is not ranked high in the BPP test. However, good performance in the BPP test does not guarantee good surface reconstruction, because the BPP test calculates only the ratio of erroneous pixels and ignores the magnitude of errors which can produce large errors in model reconstruction. Variational methods tend to spread errors into neighbouring pixels to suppress prominent errors which result in a relatively low ranking in the BPP test despite good subjective performance indicated in Fig. 13. Many GC and BP-based methods are ranked higher in the BPP test, but they produce discrete disparity maps which can cause quantisation artifacts as already shown in Fig. 3. The Adapting BP algorithm (Klaus et al., 2006) which is ranked top of the BPP test fits the disparity fields into segmented planes so it loses all surface detail.

We also compared root mean square error (RMSE) of the disparity map to the ground truth in Fig. 15. The image-driven (Kim and Sohn, 2003b) and disparity-driven (Zimmer et al., 2008) approaches are variations of the tensor types introduced in Section 4.2. The results were produced from the same initial disparity maps for the proposed algorithm. The errors are measured except at the boundary 25 pixels of the images because some algorithms do not have boundary processing. The comparison of RMSE shows that the proposed method competes with state-of-the-art methods.

The tests in this section demonstrate that the proposed algorithm shows comparable performance in the general cases of stereo disparity estimation. The advantage of the proposed approach is to generate continuous depth fields while preserving surface details. The performance of the approach is evaluated further in Section 6.3.

6.2 Error analysis of spherical stereo

Before moving to 3D reconstruction using spherical stereo, we consider the characteristic of errors in spherical stereo because it has serious radial distortion and works in a spherical coordinate system. In spherical stereo, depth error $\varepsilon = (\hat{r} - r)$ from disparity estimation error e varies according not only to the distance but also to the elevation angle for spherical stereo imaging as we can derive from Eq. (2), when the estimated depth with error is given as Eq. (19).

$$\hat{r}_t = r_t + \varepsilon_t = B / \left(\frac{\sin\theta_t}{\tan(\theta_t + d_t + e_t)} - \cos\theta_t \right) \qquad (19)$$

Figure 16 (a) shows the behaviour of depth errors according to the elevation angle θ_t and baseline distance B for 1° error in disparity at the same dispar-

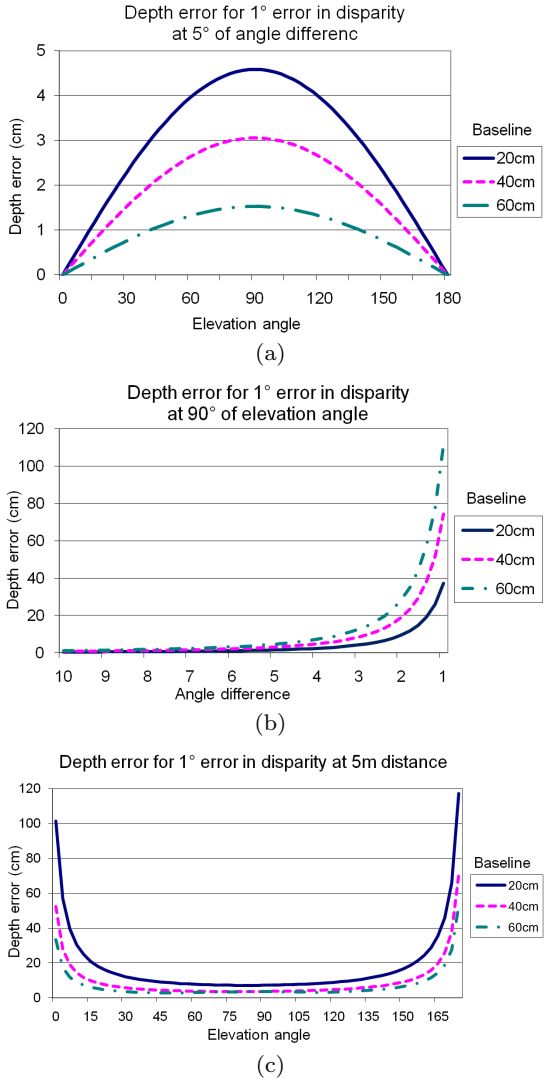


Fig. 16 Behaviour of depth errors in spherical stereo: (a)Depth error at the same angle disparity; (b)Depth error at the same elevation angle; (c)Depth error at the same distance

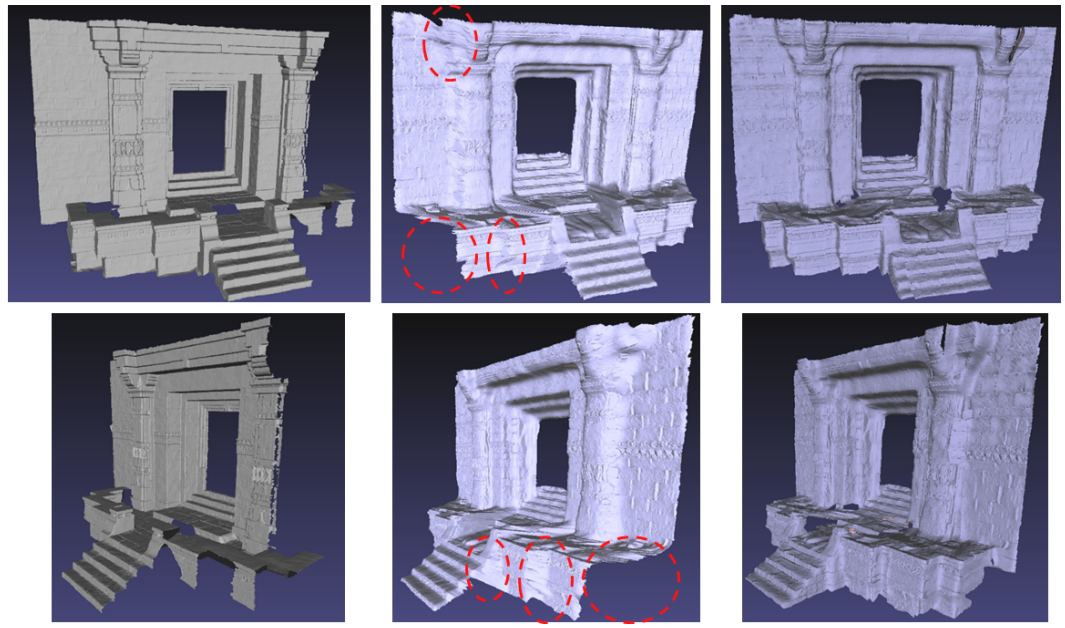
ity d. Figure 16 (b) shows the depth error according to the disparity d at 90° of elevation angle (e.g. according to the distance from the camera). From Fig. 16 (a), we observe that the disparity error around the centre of the image is more sensitive to depth error because the error increases as it comes to the centre area. However, the distance for the same angle difference increases as it approaches the poles (0° and 180°). This is shown in the spherical stereo geometry of Fig. 2 (b), and the depth error also increases as the distance increases as seen in Fig. 16 (b). Therefore they counterbalance each other to some extent. Figure 16 (c) shows the relation between the depth error and elevation angle at the fixed distance. If we assume that the target points are at 5m distance from the camera, the depth errors for the same disparity error vary by 5∼10% in the range 30°∼150° of elevation angle and rapidly increase towards the poles. The depth in this experiment is not z-depth in the Cartesian coordinate but radial distance in the spherical coordinate.

In the following experiments, we reconstruct models only for the range of 30°∼150°. In practice, the regions around the poles are not normally meaningful because 0°∼30° is the sky region in outdoor scene and the tripod is captured in 150°∼180°.

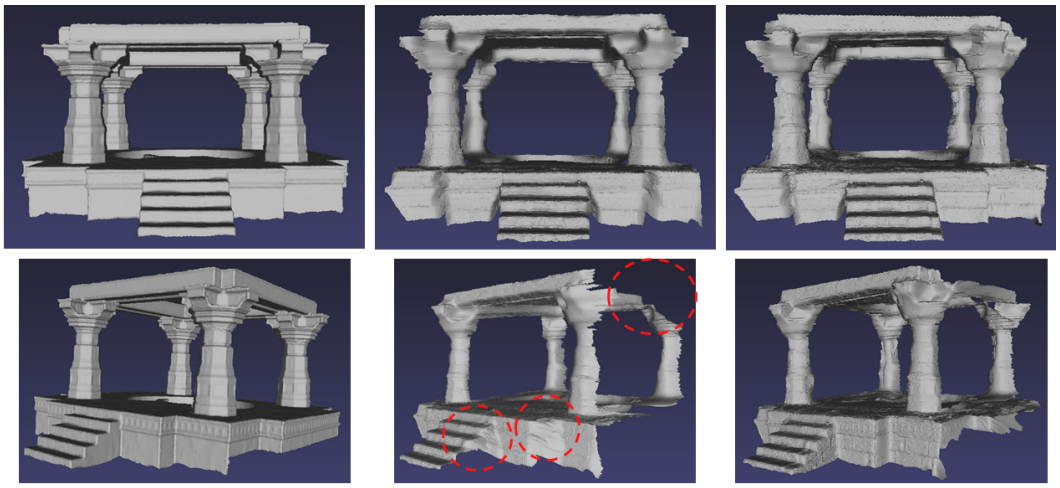
6.3 Evaluation of scene reconstruction against ground-truth LIDAR data

For objective evaluation of 3D reconstruction from spherical image pairs, we use two scenes reconstructed from image pairs captured at three different locations and compare the models with ground-truth models scanned by a LIDAR sensor. Figure 17 shows the ground-truth models and the reconstructed models from single/multiple viewpoints by the proposed algorithm. The "Gate" has width of 9m and height of 6m, and the "Cupola" is 6.2m×3.8m. Both objects are around 6m apart from the central capture point and stereo pairs are captured with a baseline of 60cm and resolution of 12574×5658. The "Gate" model was captured from 3 different points and the "Cupola" was from 2 points. The reconstructed model shows fine structure with details of the surface relief pattern. The multiple stereo reconstruction recovers self-occluded regions while maintaining the surface details of individual captures.

Direct comparison of the accuracy and completeness of reconstructed meshes is difficult because the reconstructed regions and areas are different as shown in Fig. 17 (even the model from the LIDAR scan does not have complete structure.). Therefore we produced Z-depth maps from arbitrarily chosen viewpoints and measure an average depth error for common regions. Table 3 shows evaluation results from two different viewpoints for each model and Fig. 18 shows examples of errors mapped into gray scale. In Table 3, we can see that the multiple stereo reconstructions have slightly better results than single stereo reconstruction. This is a bit more obvious in the slanted view (Viewpoints 2) than the frontal view (Viewpoints 1). However, the differences are not remarkable because most of the errors in Table 3 are from the vertical self-occlusions as seen in Fig. 18. The vertical self-occlusions could not be recovered in this experiment because: (1) the test models were captured only from horizontally scattered locations, (2) the ground-truth LIDAR data was scanned at 23m from the main objects while the spherical view was captured at 6m from the objects. Therefore, there



(a) Gate



(b) Cupola

Fig. 17 Reconstructed models (Left: Ground-truth by LIDAR scan, Middle: Reconstruction from a single stereo pair, Right: Reconstruction from multiple spherical stereo pairs)

is a limitation in producing sharp planes parallel to the depth-direction from spherical images because of relatively large FOV. Another point to be considered is that this comparison was performed only for commonly reconstructed regions. As seen in Fig. 17 (a), the multi-view reconstruction could complete lower walls where single-view reconstruction could not due to occlusion.

6.4 Multi-viewpoint reconstruction and free-viewpoint rendering

For full outdoor scene reconstruction, we captured multiple spherical stereo pairs of scenes and reconstructed 3D models using the proposed algorithms. Fig. 19 shows capture points and images of four scenes “Cathedral”, “Carpark”, “Highstreet” and “Quarry” for experiments in this section. Red points in the first row show spherical stereo capture points and the blue lines are normal still image capture paths for SfM-based reconstruction methods. The Cathedral scene is composed of one main building and surrounding open areas. The main building has a complex structure with many self-occlusions and complicated details such as sculptures. The Carpark scene has more occlusions by cars and there is relatively few overlapping regions between image 1 and 3. The Highstreet scene covers a street of 150m with 7 image pairs and includes reflections on windows and change of

(a) Depth from the viewpoint 2 of "Gate"

(b) Depth from the viewpoint 2 of "Cupola"

Fig. 18 Depth errors (Left: Ground-truth depth in common regions, Middle: Depth error of single view reconstruction, Right: Depth error of multi view reconstruction)

Table 3 Depth error evaluation (unit: cm, σ: standard deviation)

	Gate				Cupola			
	Viewpoint 1 (frontal)		Viewpoint 1 (Fig.18)		Viewpoint 1 (frontal)		Viewpoint 1 (Fig.18)	
	mean		σ		mean		σ	
Single-view	5.32	14.15	1.81	21.52	3.75	19.85	5.27	28.30
Multi-view	4.78	16.02	1.10	19.72	3.21	16.52	4.45	23.67

Table 4 Comparison with SfM-based methods

	Resolution	Carpark			Cathedral		
		# of input images	# of output vertices	Running time (min)	# of input images	# of output vertices	Running time (min)
Arc3D	2272×1704	50	5,995,192	372	92	6,389,091	855
Bundler	2272×1704	50	50,888	47	92	155,295	115
PMVS			113,107	72		221,076	152
Poisson			119,838	80		238,829	168
Proposed	6284×2794	4	443,544	225	6	747,157	287

lightings. The Quarry scene was captured in more devastated area with less features and covers 30m×30m with 4 image pairs.

One of the most serious problems in stereo matching for building scenes is reflection or transparency of windows. Real scenes include non-Lambertian surfaces and different specular reflections on the surface in the stereo image pair which induce errors in disparity estimation. Moreover, the scenes reflected on the glass come from farther away than the real position of the windows and result in false depth for the glass. A grammar-driven approach can be a solution for window detection (Simon et al., 2011; Mathias et al., 2011). However, the grammar-based approach has a problem that semantic segmentation is not always stable, and this approach works only within the given rule and categories. Any object or building outside of the given categories may induce errors in reconstruction. Therefore we manually corrected the initial disparity values for windows, reflection, lens flare regions and marked as occlusion. We also removed pedestrians and marked corresponding regions as occlusions. For the marked occlusion regions, we set the weighting term $h(\cdot)$ in Eq. (10) to zero so that only disparity field smoothing is performed for the regions in disparity estimation.

We first compared the reconstruction results with other SfM and MVS-based methods. We captured the Carpark and Cathedral scenes with a normal camera

(a) Cathedral (b) Carpark (c) Highstreet (d) Quarry

Fig. 19 Multiple outdoor capture with spherical camera (Top: capture points on maps (from http://maps.google.com), Bottom: Top images of captured stereo pairs)

and tried to reconstruct 3D models using ARC3D (Vergauwen and Gool, 2006), Bundler (Snavely et al., 2008), PMVS (Furukawa and Ponce, 2010), and Poisson reconstruction (Kazhdan et al., 2006) algorithms. The ARC3D is a web-based 3D reconstruction service running on a server connected to a cluster of computers. This estimates the camera parameters, as well as dense depth maps for the uploaded images. Bundler produces camera parameters using feature matching and 3D points clouds of scene geometry by triangulation. PMVS uses the camera parameters from Bundler and produces a more dense point cloud using multi-view stereo matching. Finally, the Poisson reconstruction builds a mesh model from the oriented points clouds.

Bundler was run on a Intel Core 2 Duo 3.0GHz Linux machine with 4G RAM, PMVS and Poisson reconstruction on a Intel Core i7 2.93GHz Windows machine with 16G RAM, and the proposed algorithm on a Intel Xeon 3.0GHz Linux machine with 32G RAM. We could not run the proposed algorithm with the maximum resolution (12574×5658) because of memory overflow. We used half resolution images for this experiment. Table 4 shows comparison of the SfM and MVS methods and the proposed method, and Fig. 20 and 21 show reconstructed geometry. In Table 4, running times for Bundler, PMVS and Poisson are accumulated because the PMVS requires output of the Bundler as input, and the Poisson reconstruction also requires point clouds from the PMVS. We could not measure running time for the ARC3D so the figure in Table 4 show the time for getting response from the server. Most time in running the proposed algorithm is used to extract floating-point disparity fields for high resolution images because the computational load is proportional to $O(N^2)$. Though it took more time than the SfM-based methods, we can see that the proposed method generated much denser points from a smaller dataset. The ARC3D generated more vertices, but we can see that most of them are concentrated on specific regions or outliers. The numbers of vertices by the proposed algorithm shown in Table 4 are only for parts of the scene which the SfM-based methods could reconstruct. The whole scene reconstructed by the proposed method covers the full 360° and produced 856,549 and 1,203,990 vertices for the Carpark and Cathedral, respectively. We tried to reconstruct a full 360° scene with the SfM-based methods. However, this failed with 50 input images due to the lack of overlap and feature matching between images, and also failed with 200 input images of the same scene because of the required memory for computation.

Figure 20 and 21 show the performance of the proposed system over the SfM and MVS methods. The first row shows reconstructed 3D point clouds for each method and the second row is mesh structure from the points. ARC3D failed to reconstruct the Carpark scene

except the right wall. It shows better result for the Cathedral but it still includes many outliers. Bundler produced very sparse point clouds and it is hard to infer the geometry of the Carpark. PMVS produced better results but included many outliers and results in a bumpy surface in Poisson reconstruction. We can see that the proposed method generated full geometry with accurate details except for occluded regions. Texture mapping can be performed with UV mapping because the estimated depth is aligned with the texture images.

Fig. 22 shows snapshots of the rendered scenes from the reconstructed models with a virtual camera. The results show natural-looking geometry and textures of the environments. The Highstreet scene in Fig. 22 (c) was reconstructed from seven pairs of extremely sparse camera captures. It covers a street of 150m and the average distance between cameras is 22m. Therefore some geometry and texture distortions are observed around the joins between partial models. The quarry scene in Fig. 22 (d) has less features but the reconstructed model shows good geometrical structure of the scene. Free-viewpoint video of the reconstructed models is available from: `http://www.youtube.com/watch?v=x3KdI8ZWZiQ`

7 Conclusions

A system for 3D environment modelling using multiple pairs of spherical stereo images has been presented. The environment is captured by a line scan camera as vertical stereo pairs of spherical images at multiple locations. 3D mesh models for each stereo pair are reconstructed using spherical stereo geometry. We proposed a novel PDE-based disparity estimation algorithm for reconstructing continuous depth fields with sharp object boundaries even in occluded and highly textured regions. A hierarchical PDE solver has been introduced to avoid the problem of convergence to local minimum in the PDE solution and reduce computational complexity for high-resolution images. Instead of an additional camera calibration for all camera locations, 3D rigid transforms between reconstructions for different spherical stereo pairs are estimated by feature matching and transform estimation between views. Finally a complete 3D model of the environment is generated by selecting the most reliable overlapping surface regions taking into account surface visibility, surface orientations and distance from the camera centre of projection. The principal advantage of the proposed surface selection algorithm against other surface merging algorithms is to preserve surface detail in the individual stereo reconstruction and eliminate outlier surfaces resulting from occlusion.

The performance of the proposed scene reconstruction algorithms was evaluated against ground-truth from the Middlebury stereo test bed and LIDAR scans. The proposed algorithms show comparable performance to the state-of-the-art methods in the general stereo cases and produce accurate surface details with sharp depth discontinuities in 3D reconstruction. Compared against the ground truth models captured using LIDAR scans, the reconstructed geometry reproduces fine details on the surfaces and gives an average depth errors within 10cm for the whole surface at distance of 10m. Analysis of the errors in spherical stereo reconstruction shows that the depth errors are stable in the range of 30°∼150° vertical angle. The models for various outdoor scenes were reconstructed and compared with the results from state-of-the-art SfM and MVS-based approaches using relatively large image sets. The proposed approach generates more complete mesh models from a relatively small set of input images. Comparison also shows that the reconstructed models can be rendered from a wide range of viewpoints with high quality textures.

At the current stage, the biggest problem of the proposed system is memory handling because of the high resolution stereo image pairs which generate a large number of vertices. The resulting model size is also proportional to the number of capture points. Future research is investigating extraction of structured mesh model representations which have a hierarchical mesh structure to efficiently approximate the scene surfaces to a required level of geometric accuracy.

Acknowledgements This research was executed with the financial support of the EU FP7 project i3DPost, UK TSB project SyMMM and EU ICT FP7 project IMPART.

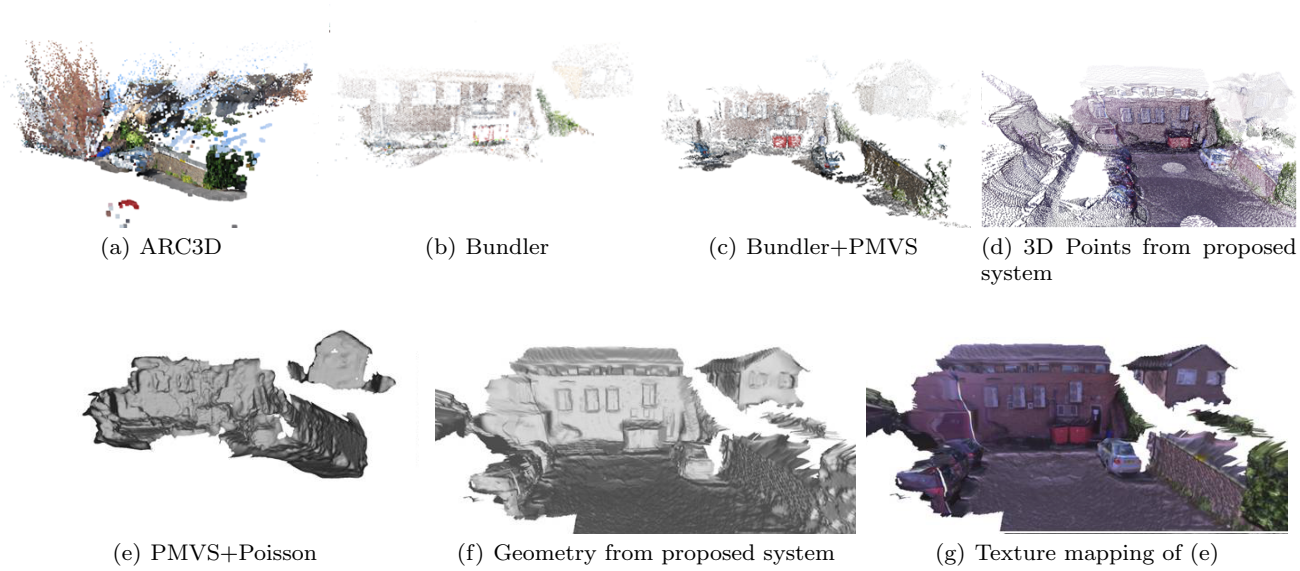
References

Agarwal, S., Snavely, N., Simon, I., Seitz, S., & Szeliski, R. (2009). Building rome in a day. In *Proc. ICCV*, pp. 72–79.

Aiger, D., Mitra, N., & Cohen-Or, D. (2008). 4-points congruent sets for robust surface registration. In *Proc. SIGGRAPH*, pp. 1–10.

Akbarzadeh, A., Frahm, J.-M., Mordohai, P., Clipp, B., Engels, C., Gallup, D., Merrell, P., Phelps, M., Sinha, S., Talton, B., Wang, L., Yang, Q., Stewenius, H., Yang, R., Welch, G., Towles, H., Nister, D., & Pollefeys, M. (2006). Towards urban 3d reconstruction from video. In *Proc. 3DPVT*, pp. 1–8.

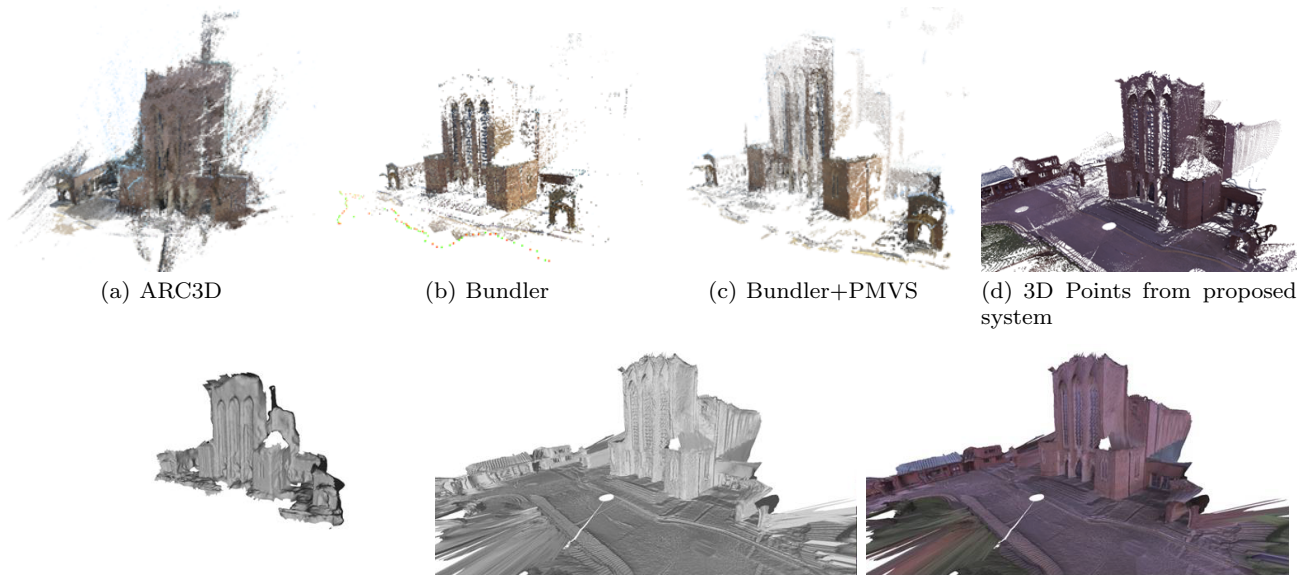
Alvarez, L., Deriche, R., Papadopoulo, T., & Sánchez, J. (2007). Symmetrical dense optical flow estimation with oclussions detection. *International Journal of Computer Vision 75*(3), 371–385.



(a) ARC3D (b) Bundler (c) Bundler+PMVS (d) 3D Points from proposed system

(e) PMVS+Poisson (f) Geometry from proposed system (g) Texture mapping of (e)

Fig. 20 Reconstruction of "Carpark"



(a) ARC3D (b) Bundler (c) Bundler+PMVS (d) 3D Points from proposed system

(e) PMVS+Poisson (f) Geometry from proposed system (g) Texture mapping of (e)

Fig. 21 Reconstruction of "Cathedral"

Alvarez, L., Deriche, R., Sánchez, J., & Weickert, J. (2002). Dense disparity map estimation respecting image discontinuities: A pde and scale-space based approach. *Journal of Visual Communication and Image Representation 13*(1), 3–21.

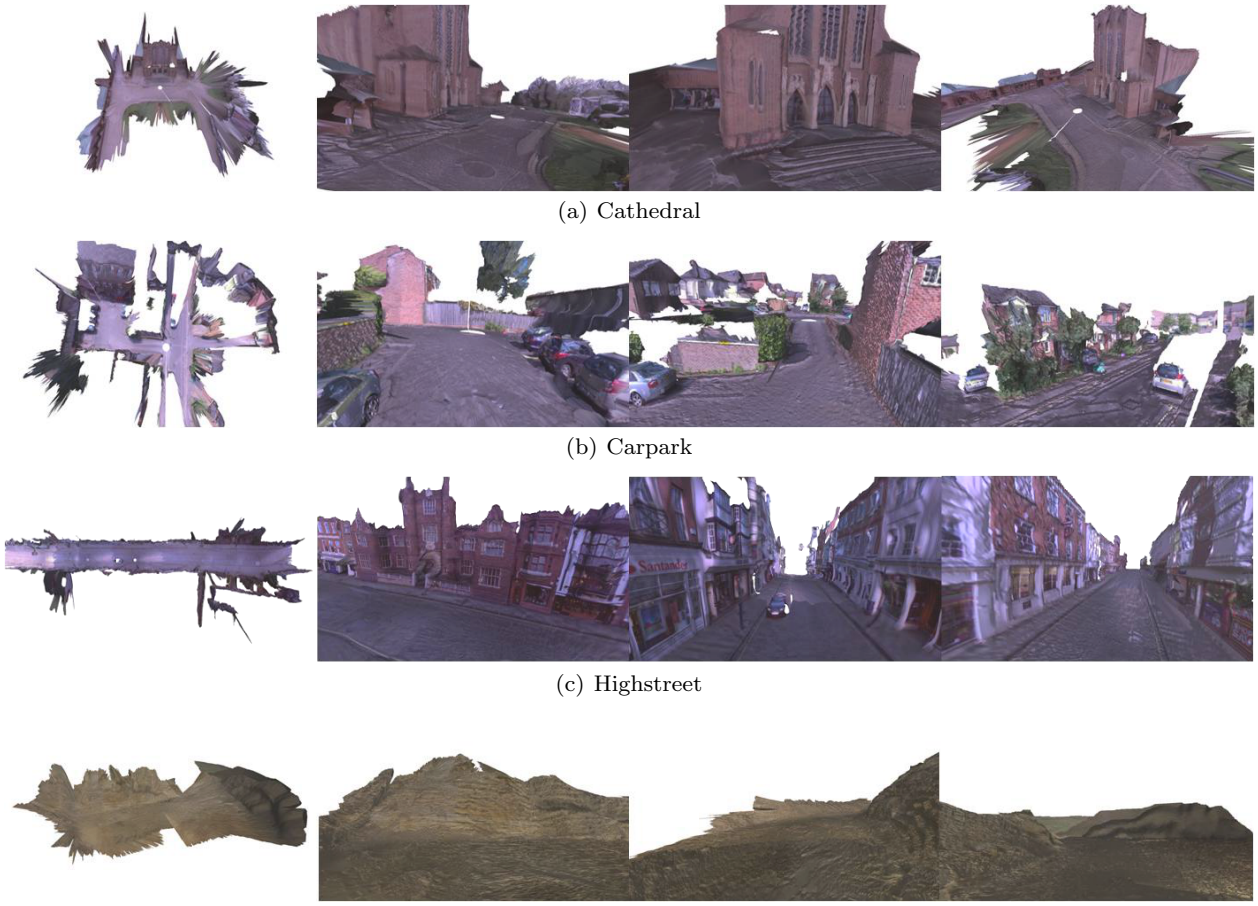
Anguelov, D., Dulong, C., Filip, D., Frueh, C., Lafon, S., Lyon, R., Ogale, A., Vincent, L., & Weaver, J. (2010). Google street view: Capturing the world at street level. *IEEE Computer 43*(6), 32–38.

Asai, T., Kanbara, M., & Yokoya, N. (2005). 3d modeling of outdoor environments by integrating omnidirectional range and color images. In *Proc. 3DIM*, pp. 447–454.

Banno, A. & Ikeuchi, K. (2009). Disparity map refinement and 3d surface smoothing via directed anisotropic diffusion. In *Proc. 3DIM*.

Banno, A. & Ikeuchi, K. (2010). Omnidirectional texturing based on robust 3d registration through euclidean reconstruction from two spherical images. *Computer Vision and Image Understanding 114*(4), 491–499.

Bay, H., Ess, A., Tuytelaars, T., & Gool, L. (2008). Surf: Speeded up robust features. *Computer Vision and Image Understanding 110*, 346–359.



(a) Cathedral

(b) Carpark

(c) Highstreet

(d) Quarry

Fig. 22 Free viewpoint rendering of reconstructed models (First column: aerial view, 2∼4th columns: free-navigation in the scene)

Ben-Ari, R. & Sochen, N. (2007). Variational stereo vision with sharp discontinuities and occlusion handling. In *Proc. ICCV*, pp. 1–7.

Benosman, R. & Devars, J. (1998). Panoramic stereovision sensor. In *Proc. ICPR*, pp. 767–769.

Besl, P. & McKay, N. (1992). A method for registration of 3-d shapes. *IEEE Trans. Pattern Analysis and Machine Intelligence 14*(2), 239–256.

Brox, T., Bruhn, A., Papenberg, N., & Weickert, J. (2004). High accuracy optical flow estimation based on a theory for warping. In *Proc. ECCV*, pp. 25–36.

Burt, P. J. (1981). Fast filter transforms for image processing. *Computer Vision, Graphics and Image Processing 6*, 20–51.

Chen, S. (1995). Quicktime vr - an image based approach to virtual environment navigation. In *Proc. SIGGRAPH*, pp. 29–38.

Chen, Y. & Medioni, G. (1992). Object modeling by registration of multiple range images. *Image and Vision Computing 10*(3), 145–155.

Cornelis, N., Leibe, B., Cornelis, K., & Gool, L. (2008). 3d urban scene modeling integrating recognition and reconstruction. *International Journal of Computer Vision 78*(2), 121–141.

Dellaert, F., Seitz, S., Thorpe, C., & Thrun, S. (2000). Structure from motion without correspondence. In *Proc. CVPR*.

Desouza, G. & Kak, A. (2002). Vision for mobile robot navigation: a survey. *IEEE Trans. Pattern Analysis and Machine Intelligence 24*(2), 237–267.

Feldman, D. & Weinshall, D. (2005). Realtime ibr with omnidirectional crossed-slits projection. In *Proc. ICCV*, pp. 839–845.

Fischler, M. & Bolles, R. (1982). Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communication of the ACM 24*, 381–395.

Fisher, R. (2007). Cvonline. *http://homepages.inf.ed.ac.uk/rbf/CVonline/LOCAL_COPIES/FISHER/REGIS/regis.html*.

Frahm, J.-M., Fite-Georgel, P., Gallup, D., Johnson, T., Raguram, R., Wu, C., Jen, Y.-H., Dunn, E., Clipp, B., Lazebnik, S., & Pollefeys, M. (2010). Building rome on a cloudless day. In *Proc. ECCV*, pp. 368–381.

Furukawa, Y., Curless, B., Seitz, S., & Szeliski, R. (2009). Manhattan-world stereo. In *Proc. CVPR*.

Furukawa, Y., Curless, B., Seitz, S., & Szeliski, R. (2010). Towards internet-scale multi-view stereo. In *Proc. CVPR*.

Furukawa, Y. & Ponce, J. (2010). Accurate, dense, and robust multiview stereopsis. *IEEE Trans. Pattern Analysis and Machine Intelligence 32*(8), 1362–1376.

Gargallo, P. & Sturm, P. (1988). Bayesian 3d modeling from images using multiple depth maps. In *proc. CVPR*, pp. 885–891.

Geman, S. & McClure, D. (1985). Bayesian image analysis: An application to single photon emission tomography. In *Proc. Statistical Computation Section*, pp. 12–18.

Goesele, M., Snavely, N., Curless, B., Hoppe, H., & Seitz, S. M. (2007). Multi-view stereo for community photo collections. In *Proc. ICCV*, pp. 368–381.

Granger, S., Pennec, X., & Roche, X. (2001). Rigid point-surface registration using oriented points and an em variant of icp for computer guided oral implantology. In *Proc. MICCAI*, pp. 752–761.

Haala, N. & Kada, M. (2005). Panoramic scenes for texture mapping of 3d city models. In *Proc. PanoPhot.*

Hilton, A. (2005). Scene modelling from sparse 3d data. *Image and Vision Computing 23*(10), 900–920.

Hilton, A., Stoddart, A., Illingworth, J., & Windeatt, T. (1998). Implicit surface based geometric fusion. *Computer Vision and Image Understanding 69*(3), 273–291.

Hirschmüller, H. & Scharstein, D. (2008). Evaluation of stereo matching costs on images with radiometric differences. *IEEE Trans. Pattern Analysis and Machine Intelligence 31*(9), 1582–1599.

Ince, S. & Konrad, J. (2008). Occlusion-aware optical flow estimation. *IEEE Trans. Pattern Analysis and Machine Intelligence 17*, 1443–1451.

Johnson, C. (1988). *Numerical Solution of Partial Differential Equations by the Finite Element Method.* Cambridge: Cambridge University Press.

Kang, S. & Szeliski, R. (1997). 3-d scene data recovery using omnidirectional multibaseline stereo. *International Journal of Computer Vision 25*(2), 167–183.

Kazhdan, M., Bolitho, M., & Hoppe, H. (2006). Poisson surface reconstruction. In *Proc. SGP*, pp. 61–70.

Kim, H. & Hilton, A. (2009). 3d environment modelling using spherical stereo imaging. In *Proc. 3DIM.*

Kim, H. & Hilton, A. (2010). 3d modelling of static environments using multiple spherical stereo. In *Proc. RMLE workshop in ECCV.*

Kim, H. & Sohn, K. (2003a). Hierarchical depth estimation for image synthesis in mixed reality. In *Proc. SPIE Electronic Imaging*, pp. 544–553.

Kim, H. & Sohn, K. (2003b). Hierarchical disparity estimation with energy-based regularization. In *Proc. ICIP*, pp. 373–376.

Klaus, A., Sormann, M., & Karner, K. (2006). Segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure. In *Proc. ICPR.*

Kolmogorov, V. & Zabih, R. (2001). Computing visual correspondence with occlusions using graph cuts. In *Proc. ICCV.*

Lemmens, M. (2007). Airborne lidar sensor. *GIM International 21*(2).

Lhuillier, M. (2008). Automatic scene structure and camera motion using a catadioptric system. *Computer Vision and Image Understanding 109*(2), 186–203.

Li, S. (2006). Real-time spherical stereo. In *Proc. ICPR*, pp. 1046–1049.

Mathias, M., Martinovic, A., Weissenberg, J., & Gool, L. J. V. (2011). Procedural 3d building reconstruction using shape grammars and detectors. In *Proc. 3DIMPVT*, pp. 304–311.

Merrell, P., Akbarzadeh, A., Wang, L., Mordohai, P., Frahm, J.-M., Yang, R., Nister, D., & Pollefeys, M. (2007). Real-time visibility-based fusion of depth maps. In *Proc. ICCV.*

Micusik, B. & Kosecka, J. (2009). Piecewise planar city 3d modeling from street view panoramic sequences. In *Proc. CVPR*, pp. 2906–2912.

Micusik, B., Martinec, D., & Pajdla, T. (2004). 3d metric reconstruction from uncalibrated omnidirectional images. In *Proc. ACCV.*

Min, D. & Sohn, K. (2008). Cost aggregation and occlusion handling with wls in stereo matching. *IEEE Trans. Image Processing 17*(8), 1431–1442.

Nagel, H. & Enkelmann, W. (1986). An investigation of smoothness constraints for the estimation of displacements vector fields from image sequences. *IEEE Trans. Pattern Analysis and Machine Intelligence 8*, 565–593.

Nayar, S. K. & Karmarkar, A. (2000). 360 x 360 mosaics. In *Proc. CVPR*, pp. 2388–2388.

Pollefeys, M., Koch, R., Vergauwen, M., & Gool, L. (2000). Automated reconstruction of 3d scenes from sequences of images. *ISPRS Journal of Photogrammetry and Remote Sensing 55*(4), 251–267.

Pollefeys, M., Nistér, D., Frahm, J., Akbarzadeh, A., Mordohai, P., Clipp, B., Engels, C., Gallup, D., Kim,

S., Merrell, P., Salmi, C., Sinha, S., Talton, B., Wang, L., Yang, Q., Stewénius, H., Yang, R., Welch, G., & Towles, H. (2008). Detailed real-time urban 3d reconstruction from video. *International Journal of Computer Vision 78*(2), 143–167.

Rusinkiewicz, S. & Levoy, M. (2001). Efficient variants of the icp algorithm. In *Proc. 3DIM*, pp. 145–152.

Salman, N. & Yvinec, M. (2009). Surface reconstruction from multi-view stereo. In *Proc. ACCV*.

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

Simon, L., Teboul, O., Koutsourakis, P., & Paragios, N. Random exploration of the procedural space for single-view 3d modeling of buildings. *International Journal of Computer Vision*.

Sizintsev, M. (2008). Hierarchical stereo with thin structures and transparency. In *Proc. CRV*, pp. 97–104.

Slesareva, N., Bruhn, A., & Weickert, J. (2005). Optic flow goes stereo: A variational method for estimating discontinuity- preserving dense disparity maps. In *Proc. DAGM*, pp. 33–40.

Snavely, N., Seitz, S., & Szeliski, R. (2006). Photo tourism: Exploring photo collections in 3d. In *Proc. ACM SIGGRAPH*, pp. 835–846.

Snavely, N., Seitz, S., & Szeliski, R. (2008). Modeling the world from internet photo collections. *International Journal of Computer Vision 80*(2), 189–210.

Soucy, M. & Laurendeau, D. (1995). A general surface approach to the integration of a set of range views. *IEEE Trans. Pattern Analysis and Machine Intelligence 17*(4), 344–358.

Strecha, C., Fransens, R., & Gool, L. J. V. (2004). Wide-baseline stereo from multiple views: A probabilistic account. In *Proc. CVPR*, pp. 552–559.

Strecha, C., Hansen, W., Gool, L., Fua, P., & Thoennessen, U. (2008). On benchmarking camera calibration and multi-view stereo for high resolution imagery. In *Proc. CVPR*, pp. 1–8.

Sun, D., Roth, S., Lewis, J., & Black, M. (2008). Learning optical flow. In *Proc. ECCV*, pp. 83–97.

Sun, J., Zheng, N., & Shum, H. (2003). Stereo matching using belief propagation. *IEEE Trans. Pattern Analysis and Machine Intelligence 25*(7), 787–800.

Szeliski, R. & Scharstein, D. (2004). Sampling the disparity space image. *IEEE Trans. Pattern Anal. Mach. Intell. 26*(3), 419–425.

Teller, S., Antone, M., Bodnar, Z., Bosse, M., Coorg, S., Jethwa, M., & Master, N. (2003). Calibrated, registered images of an extended urban area. *International Journal of Computer Vision 53*(1), 93–107.

Tighe, J., Feldman, & Lazebnik, S. (2010). SuperParsing: Scalable Nonparametric Image Parsing with Superpixels. *Proc. ECCV*.

Triggs, B., McLauchlan, P., Hartley, R., & Fitzgibbon, A. (1999). Bundle adjustment - a modern synthesis. In *Proc. International Workshop on Vision Algorithms: Theory and Practice*, pp. 298–372.

Turk, G. & Levoy, M. (1994). Zippered polygon meshes from range images. In *Proc. SIGGRAPH*, pp. 311–318.

Vergauwen, M. & Gool, L. (2006). Web-based 3d reconstruction service. *Machine Vision Applications 17*, 411–426.

Vu, H., Keriven, R., Labatut, P., & Pons, J. (2009). Towards high-resolution large-scale multi-view stereo. In *Proc. CVPR*, pp. 1430–1437.

Weickert, J. (1997). A review of nonlinear diffusion filtering. *Lecture Notes in Computer Science 1252*, 3–28.

Williams, J. & Bennamoun, M. (2001). Simultaneous registration of multiple corresponding point sets. *Computer Vision and Image Understanding 81*(1), 117–142.

Woodford, O.J., Torr, P.H.S., Reid, I.D., & Fitzgibbon, A.W. (2008). Global stereo reconstruction under second order smoothness priors. In *Proc. CVPR*, pp. 1–8.

Yang, Q., Wang, L., Yang, R., Stewénius, H., & Nistér, D. (2008). Stereo matching with color-weighted correlation, hierarchical belief propagation and occlusion handling. *IEEE Trans. Pattern Analysis and Machine Intelligence 31*(3), 492–504.

Yuille, A. & Poggio, T. (1984). A generalized ordering constraint for stereo correspondence. *MIT A.I. Memo 777*.

Zimmer, H., Bruhn, A., Valgaerts, L., Breuß, M., Weickert, J., Rosenhahn, B., & Seidel, H. (2008). Pde-based anisotropic disparity-driven stereo vision. In *Proc. VMV*, pp. 263–272.

Zomet, A., Feldman, D., Peleg, S., & Weinshall, D. (2003). Mosaicing new views: the crossed-slits projection. *IEEE Trans. Pattern Analysis and Machine Intelligence 25*(6), 741–754.