PLANAR URBAN SCENE RECONSTRUCTION FROM SPHERICAL IMAGES USING FACADE ALIGNMENT

Hansung Kim and Adrian Hilton

Centre for Vision, Speech and Signal Processing, University of Surrey, GU2 7XH, UK
h.kim@surrey.ac.uk, a.hilton@surrey.ac.uk

ABSTRACT

We propose a plane-based urban scene reconstruction method using spherical stereo image pairs. We assume that the urban scene consists of axis-aligned approximately planar structures (Manhattan world). Captured spherical stereo images are converted into six central-point perspective images by cubic projection and facade alignment. Facade alignment automatically identifies the principal planes direction in the scene allowing the cubic projection to preserve the plane structure. Depth information is recovered by stereo matching between images and independent 3D rectangular planes are constructed by plane fitting aligned with the principal axes. Finally planar regions are refined by expanding, detecting intersections and cropping based on visibility. The reconstructed model efficiently represents the structure of the scene and texture mapping allows natural walk-through rendering.

Index Terms— 3D Reconstruction, Spherical Imaging

1. INTRODUCTION

Simplified urban scene modelling has been a long-standing area of research. The FAÇADE system introduced an approach for modelling and photo-realistic rendering of architecture from a number of photographs [1]. Sinha et al. used feature matching and Structure-from-Motion methods with line and vanishing point detection algorithms for interactive 3D architectural modelling from photo collections [2]. Google Sketchup\(^1\) provides a simple tool for 3D reconstruction from multiple photos. This is useful for representations of simple scenes but is limited for complex scenes because it requires manual matchings for each primitives.

To achieve fully automatic reconstruction, Sinha et al. suggested extracting vanishing directions and fitting point clouds into 3D planes reconstructed based on the vanishing directions [3]. Toldo et al. proposed planar patch extraction using the J-linkage algorithm and reconstructed the scene with view clustering tree and hierarchical reconstruction [4]. Some approaches have invoked the stronger Manhattan world assumption which states the world is piecewise planar and aligned to orthogonal axes. Micusik et al. proposed super-pixel stereo on a Markov Random Field and aligned surfaces to three dominant directions [5]. Furukawa et al. also built various scenes by axis aligned depth map integration relying on the Manhattan world assumption [6].

In this paper, we propose a simple and efficient method to reconstruct a simplified planar structured environment model with spherical images. A line scan camera captures the full surrounding scene as vertical stereo image pairs. The captured spherical images are projected into a unit cube with a novel facade alignment algorithm based on the Hough transform. Each face image of the cubic projection is a distortion-free central-point perspective image whose principal axes are aligned to vertical/horizontal directions and the image centre. From the pair of cubic projection images, independent 3D rectangular planes are constructed by stereo matching and plane fitting. The planes are then connected or refined to more meaningful structure by expanding and cropping based on visibility. The result efficiently represents the structure as a set of planar billboards which can be used to render the scene for applications such as virtual walkthrough and pre-visualisation.

2. CUBIC PROJECTION IMAGING

2.1. Spherical image capture

In this work, we use a commercial off-the-shelf line-scan camera\(^2\) with a fisheye lens in order to capture the full environment as a high resolution spherical image. The camera rotates about an axis passing through the optical centre. The rays are sampled on a hemisphere at its centre of projection and stitched from the rotating slits together into a new image. Therefore the imaging geometry of the line-scan capture can be regarded as conventional perspective projection, and the result is a latitude-longitude image like a world map. Mapping errors from misalignment and lens distortion can be corrected by rectification.

\(^1\)Google Sketchup, http://sketchup.google.com/

To recover depth information from the images, the scene is captured with the camera at two different heights. This vertical stereo line-scan camera capture has two advantages: 1) the full surrounding scene is captured; 2) stereo matching can be simplified to a 1D search along the vertical scan line in the latitude-longitude geometry if the pair of images are vertically aligned.

2.2. Cubic Projection

The latitude-longitude images can be directly used for disparity estimation and 3D reconstruction. However, we propose to convert the image into conventional perspective images via projection of the spherical image to a cube, referred to here as cubic projection. Conversion from spherical to perspective image sampling aids subsequent scene analysis as in the spherical image straight scene lines are curved and remain straight in a perspective image. The cubic projection projects the latitude-longitude images can be directly used for disparity estimation and 3D reconstruction. However, we propose to convert the image into conventional perspective images via cubic projection. Conversion from spherical to perspective image sampling aids subsequent scene analysis as in the spherical image straight scene lines are curved and remain straight in a perspective image. The cubic projection projects all pixels on the unit sphere to the unit cube in the range of [-1, 1] in each axis. The converted image is decomposed into six face images as Fig. 1 (a) and (b). We arbitrarily set 0° of longitude as the x-axis in the initial cubic projection.

As we can see in Fig. 1 (b), each side face image of the cubic projection has two vanishing points. We can make these images single vanishing point images if we set the x-axis of the cubic projection to be aligned to the Manhattan world axis, which we refer to as facade alignment.

2.3. Facade Alignment

Facade alignment is the process of matching the main facades in the scene to be perpendicular to the main axes of the cubic projection by rotating the spherical image around the vertical axis. Figure 1 (c) is the projection result when 126° of the longitude was set to the x-axis as a result of the facade alignment. We can observe that the horizontal and vertical lines in the scene are aligned to horizontal and vertical directions in each image, respectively, and the lines aligned to the direction converge to the image centre. Therefore we can consider these images as central-point perspective images. This central-point perspective image has significant advantages for axis-aligned plane reconstruction. Most of the current plane based reconstruction algorithms use vanishing point and principal directions detections in 3D space [3, 5, 6], but this cubic projection with the facade alignment can detect 3D principal directions in 2D images. Therefore, aligned 3D planes can be easily built from stereo matching.

In order to find the most reliable angle shift \( t \) to set the x-axis, we consider the number, sparseness, average length and average angle errors of scene image lines resulting from the probabilistic Hough transform [7]. The angle shift is equivalent to the horizontal pixel shift in the line-scan image. We detect the following three kinds of Hough lines as aligned lines among all detected lines \( H \): horizontal Hough lines \( H_h \), vertical Hough lines \( H_v \) and perspective Hough lines \( H_p \).

\[
\arg\max_{0^\circ \leq \theta < 360^\circ} E_F(t) = \lambda_R E_R(t) + \lambda_S E_S(t) + \lambda_L E_L(t) \tag{1}
\]

\[
E_R = \frac{|H_h \cup H_v \cup H_p|}{|H|}
\]

\[
E_S = \sigma_{H_h} / |H|
\]

\[
E_L = \frac{1}{|H_h|} \sum_{H_h} l(h) / (|I_w| \cdot \theta(h))
\]

In Eq.(1), \( E_R \) represents the ratio of the number of aligned Hough lines to all Hough lines. \( E_S \) is the standard deviation of y-position of \( H_h \) which shows sparseness of the horizontal Hough lines. We give higher priority to sparse features in the scene because dense Hough lines can be detected from small areas of the scene which have complex line structures and may bias the optimisation. Finally, \( E_L \) shows the magnitude and accuracy of the detected horizontal lines \( H_h \) where \( l(h) \) is length of the line. We experimentally set \( \lambda_R=1.0, \lambda_S=1.0, \lambda_L=0.3 \).

According to the Manhattan world assumption, scenes should show four distinct peaks at around 90° intervals in
$E_F(t)$. However, some scenes have ambiguous peaks if they include slanted surfaces or have high frequency textures which induce outliers in the Hough transform. Therefore, we detect the optimal shift by maximising the following energy sum which shows the distinct peak point.

$$\arg\max_{0^\circ \leq t < 90^\circ} \sum_{k=0}^{3} E_F(t + k \cdot 90^\circ)$$  \hspace{1cm} (2)

3. PLANE-BASED SCENE RECONSTRUCTION

3.1. Depth Reconstruction

Stereo matching can be carried out for spherical image pairs. Any disparity estimation algorithm can be used as long as it does not produce too many outliers. In our experiments, we used a block matching method with region-diving technique which produces reliable disparity fields by detecting occlusion regions and ambiguous regions based on bi-directional matching and ordering constraint [8].

The disparity map is converted to the cubic images and aligned to the original textures. Then all 2D points with disparity information can be projected into 3D space to generate a 3D point cloud with spherical stereo geometry [9].

3.2. Plane Reconstruction

Rectangular planes are constructed from the point clouds by the total least squares (orthogonal regression) fitting algorithm [10] and bounding box extraction. They are fit into three classes ($x$, $y$ and $z$ planes) according to the constraints $n_i \sim N(0, \sigma_n^2) (i \in x, y, z)$ for their normal vectors $n_i$. $\sigma_n^2$ is the variation from the ideal normal vector for a plane in the $i$-direction. We set $\sigma_n^2$ as 0.15 for our experiments. All other planes which do not belong to any categories are removed. The result can be biased by outliers from disparity estimation errors. Therefore we iteratively refine the results with RANSAC [11].

After categorisation, we eliminate small planes and refine plane intersections. All planes are checked for intersection with other planes in the extension range of $T_e$ which is experimentally set to 80cm. If any intersection is found, we check the length of intersection and visibility to determine the type of intersection. If the intersection is larger than half of the bigger plain width, two planes are welded at the intersection and a corner is generated as illustrated in Fig. 2 (a). Otherwise only the smaller plane stops growing at the intersection to generate a T-junction, as illustrated in Fig. 2 (b). If two planes already have an intersection the residual parts are eliminated based on a visibility constraint [9] as illustrated in Fig. 2 (c). If the plane does not meet any intersection during the extension to any direction, we keep the original boundary.

Finally we fill in gaps from self-occlusion. The scene captured by the spherical cameras inevitably has self-occlusions in the scene as illustrated in Fig. 3 (a). We adopt the minimum volume solution [6] based on the Manhattan world assumption. The plane occluded at the rear of the front plane is extended to the boundary of the orthogonal line-of-sight (LOS) from the surface normal direction as Fig. 3 (b). Occluded regions perpendicular to the orthogonal LOS are reconstructed from other viewpoints.

4. EXPERIMENTAL RESULTS

We tested the proposed algorithm on 3 different multiple spherical stereo data sets as shown in Fig. 4. We captured the scene with a resolution of $3143 \times 1414$ pixels. Reconstructions from multiple viewpoints were easily registered by defining correspondence manually. Once all viewpoints are registered into a unified coordinate system, close planes are merged before the plane refinement stage.

Figure 5 shows the reconstructed plane structure of the scenes and their texture mapping results. We can see that the proposed method generates a coarse approximation of the scene structure as a set of 3D plane proxies which are sufficient to provide a visual walkthrough of the scene using texture mapping. Texture mapping produces reasonably natural rendering results. We did not use occlusion handling in texture mapping so occluded regions in the scene show wrong texture from the occluders. This can be compensated from textures in other viewpoints. Free-viewpoint rendering video of the scenes is available from: http://youtu.be/6sJ5sroQChg
5. CONCLUSION

In this paper, a plane-based simplified urban scene reconstruction method from spherical stereo image pairs was proposed. The planar representation provides a compact representation of the scene as a set of 3D planar billboard proxies for rendering. Detailed geometry is not represented in the reconstruction, but texture mapping allows rendering of the appearance of detailed geometry suitable for scene visualisation.

Future extensions of this research include: 1) Plane-based reconstruction and refinement with arbitrary planes which are not aligned to the principal axes; 2) Texture mapping from multiple viewpoints considering visibility; 3) Volumetric reconstruction by fitting reconstructed structure into basic 3D elements such as cuboid and triangular prism to reconstruct a complete structured scene approximation.

Acknowledgements

This research was supported by the UK TSB project SyMMM and the European Commission, FP7 IMPART project (grant agreement No 316564).

6. REFERENCES


