Toward Flexible 3D Modeling using a Catadioptric Camera

Maxime Lhuillier
LASMEA UMR 6602 CNRS, Université Blaise Pascal
24 avenue des Landais, 63177 Aubiere Cedex, France.
maxime.lhuillier.free.fr

Abstract

Fully automatic 3D modeling from a catadioptric image sequence has rarely been addressed until now, although this is a long-standing problem for perspective images. All previous catadioptric approaches have been limited to dense reconstruction for a few view points, and the majority of them require calibration of the camera. This paper presents a method which deals with hundreds of images, and does not require precise calibration knowledge. In this context, the same 3D point of the scene may be visible and reconstructed in a large number of images at very different accuracies. So the main part of this paper concerns the selection of reconstructed points, a problem largely ignored in previous works. Summaries of the structure from motion and dense stereo steps are also given. Experiments include the 3D model reconstruction of indoor and outdoor scenes, and a walkthrough in a city.

1. Introduction

Producing photo-realistic 3D models for walkthroughs in a complex scene given an image sequence is a long-term research problem in Computer Vision and Graphics. A minimal requirement for interactive walkthrough is the scene rendering in any view direction around the horizontal plane, when the viewer moves along the ground. This suggests a wide field of view for the given images, for which many kinds of cameras are possible [3]: catadioptric cameras, fish-eyes, or systems of multi-cameras pointing in many directions. Since we would like to capture any scene (indoor and outdoor) where a pedestrian can go, the hardware involved should be hand-held/head-held and not cumbersome. A catadioptric camera is a good candidate for all these constraints, and it has been adopted in this work.

The main drawback of this choice is the low resolution compared with a standard (perspective) camera for a given field of view. We would like to compensate for this problem using still image sequence taken with an equiangular catadioptric camera. Still images are preferred to video images thanks to their better quality (resolution, noise). An equiangular camera has also been selected from among other catadioptric cameras, since it is designed to spread the resolution of view field well in the whole image. These two choices mainly have two consequences. First, a still image sequence requires some effort and patience: the user should alternate a step forward and a (non blurred) shot. Second, an equiangular catadioptric camera is not a central camera. A non-central model complicates all involved methods, since the back-projected rays do not intersect a single point in space as the perspective model. This paper shows that the results obtained are worthwhile with a such setup, and that a central approximation for the camera is sufficient in many cases.

The presented approach is fully automatic given an image sequence acquired by the camera moving in the scene: (1) estimate the full geometry using a Structure from Motion (**SfM**) method, and (2) build a 3D model using multi-view dense stereo and stereo fusion. By SfM, we mean the automatic, robust and optimal estimation of the camera motion (all extrinsic and some intrinsic parameters) and the scene structure (a sparse set of points in 3D) from images. The 3D model is a list of textured triangles in 3D which approximate the visible part of the scene, and is visualized using standard tools. Although this approach is now “standard” for a monocular, hand-held and perspective camera (e.g. [14, 16, 11]), no such 3D reconstruction results have been obtained before with a catadioptric camera.

1.1. Previous Works

During the past decade many researchers have worked on the 3D model reconstruction given an image sequence. A lot of work has been done in different contexts involving active or passive sensors, aerial or pedestrian view points, general or specific camera motions (e.g. turntable), general or specific objects to modelize (e.g. faces), manual or automatic methods. A complete survey is outside the scope of this paper, and we only focus on the most related work.

Previous authors [7, 1, 5] reproject their original images onto virtual cylinder (panoramic) images to apply dense stereo with the epipolar constraint. The first attempt [7] merges images taken by a rotated perspective camera, and the others [1, 5] reproject calibrated (central) catadioptric images. Point tracking and the 8-point algorithm [6] are used by [7, 1] to estimate the geometry of pair of successive cylinder images. Then, the relative scales between geometry pairs are recovered by measuring the distance between camera poses or odometry [7, 1], and the global geometry of the sequence is obtained. Finally, these two approaches use multi-baseline stereo methods inspired by [15]. The third work [5] uses active methods to recover the sequence geometry, and a graph-cut method [8] followed by a lot of postprocessing for dense stereo. A simple manual method is also proposed by [4], but the important point here is the introduction of an image pair selection scheme to reconstruct a point with the best "reliability". This is a key issue (neglected by other authors) for 3D modeling from catadioptric images, since the same 3D point may be visible and reconstructed in a great number of images at very different resolutions and baselines.

As mentioned before, SfM [6] is the first problem to solve. Although the principles are well known, optimal (including global bundle adjustment) and robust SfM systems are not so common if the only given data are a long image sequence acquired by a general catadioptric camera and some knowledge about the calibration. This contrasts with the situation regarding perspective cameras. The most advanced research into this subject have been [18] and [13]. A bundle adjustment (BA) is applied once at the very end of the geometry estimation process given an accurate camera calibration [18], although it is recognized that a hierarchical scheme for BA [6] greatly improves the accuracy and robustness simultaneously. The uncalibrated approach [13] is a two step camera modeling method: (1) recover the sequence geometry using an approximate central model for the camera and (2) upgrade the sequence geometry using a non-central model enforcing the mirror knowledge.

1.2. Contributions

Section 2 presents a central camera model and summarizes the SfM method. The differences with [13] are: long sequences thanks to a hierarchical BA, general radial function with some rough parameter knowledge (instead of the knowledge of model type with unknown intrinsic parameters). Section 3 presents a method to obtain a "local" 3D model of the scene given three images of the sequence, using dense stereo. We don't use cylinder images like [7, 1, 5] in the dense stereo step to facilitate the 3D reconstruction of the scene ground. Section 4 describes a generalization of the image pair selection scheme to reconstruct a point: we replace the heuristic two-view criterion [4] (using angle between rays) by a general multi-view criterion (using uncertainty ellipsoids). The most accurate parts of all local models are selected by this criterion, and they are merged into the global 3D model of the scene. This subject was neglected by other authors, although the same 3D point may be estimated with very different accuracies (in our catadioptric context) depending on the images selected for reconstruction. Note that selection is a preprocessing step for merging methods like [2, 17], it is not a concurrent method. Experiments in Section 5 include both indoor and outdoor scene reconstructions (only indoor examples are provided in previous works [7, 1, 4, 5]).

2. Geometry Estimation (Summary)

In this Section, the automatic SfM approach is briefly summarized (more technical details in [9]). The following assumptions are required: (1) a surface-of-revolution mirror whose lower and upper circular cross sections are visible (2) a perspective camera with zero skew and aspect ratio set to 1 (3) mirror and perspective camera with the same symmetry axis (4) constant calibration and (5) approximate knowledge of the two field of view angles.

2.1. Central Camera Model

The camera model is defined by its orientation R (a rotation), the center $t \in \Re^3$ ($\Re$ is the real numbers), both expressed in the world coordinate system, and a central projection $p : \Re^3 \setminus \{0\} \rightarrow \Re^2$. Using the same notation X for a finite 3D point and its world coordinates, the direction of the ray from t to X in the camera coordinate system is given by $d = R^\top (X - t)$. The image projection of X is $p(d)$. The model has a symmetry around the z-axis of the camera coordinate system: the omnidirectional image is between two concentric circles of radii r_{up} and r_{down}, and there is a positive and decreasing function r such that

$$p(x, y, z) = r(\alpha(x, y, z)) \begin{pmatrix} \frac{x}{\sqrt{x^2+y^2}} & \frac{y}{\sqrt{x^2+y^2}} \end{pmatrix}^\top$$

with $\alpha(x, y, z)$ the angle between the z-axis and the ray direction $d = \begin{pmatrix} x & y & z \end{pmatrix}$. Let α_{up} and α_{down} be the two angles which define the field of view. We have $\alpha_{up} \leq \alpha \leq \alpha_{down}$ and $r_{down} \leq r(\alpha) \leq r_{up}$. An exact equiangular camera is obtained if r is a linear function.

2.2. Calibration Initialization

Thanks to the assumptions, the projections of the mirror circular cross sections are concentric circles. First these circles are detected and estimated in each catadioptric image using RANSAC and Levenberg-Marquardt methods applied on regularly polygonized contours (in this paper, the detection stability is improved assuming that the circles are fixed in the whole sequence). Then, the initial $r(\alpha)$ is defined by the linear function such that $r(\alpha_{up}) = r_{up}$ and $r(\alpha_{down}) = r_{down}$. The image circle radii r_{down}, r_{up}

are obtained in the first step, and the field of view angles $\alpha_{down}, \alpha_{up}$ are given by the mirror manufacturer. These angles are not exactly known since they depend on the relative position between the mirror and the pinhole camera.

2.3. Geometry Initialization

Harris points are detected and matched for each pair of consecutive images in the sequence using ZNCC correlation, without any epipolar constraint. The corresponding ray directions are also obtained from the calibration initialization. Then, the essential matrices for these pairs are estimated by RANSAC (using the 7-point algorithm [6]) and refined by Levenberg-Marquardt. 3D points are also reconstructed for each pair. Many of these points are tracked in three images, and are used to initialize the relative 3D scales between two consecutive image pairs. Finally, the full sequence geometry is obtained by many bundle adjustments (BA) applied in a hierarchical framework to merge all partial geometries [6]. These BA simultaneously refine the parameters of 3D points and camera poses by minimizing the sum of squared reprojection errors.

2.4. Geometry Refinement

Once the full geometry is obtained for the approximate function $r(\alpha)$ defined above, $r(\alpha)$ is redefined as a cubic polynomial whose the 4 coefficients should be estimated. An additional BA is applied to estimate the $4 + 6c + 3p$ parameters of the sequence (c is the number of cameras, p is the number of 3D points) and increase the numbers of 3D and 2D inliers.

3. Local 3D Models

Assume that the geometry of the catadioptric image sequence is estimated with the method described in Section 2. Since a central camera model is used, the epipolar constraint and reprojections to virtual surfaces (like cylinder or cube) are easy. This is not the case with a non-central model.

A local model is reconstructed from three images: one reference image ref, one secondary image sec_1 before and one secondary image sec_2 after the reference image in the sequence. The three images do not have to be consecutive to deal with many baselines and reconstruct parts of the scene at various depths. Obviously, large baselines increase the accuracy for distant parts, but small baselines are also necessary for the closest parts (e.g. ground) to increase the common field of view and simplify the matching problem.

Previous authors [7, 1, 5] reproject their images onto virtual cylinder images to apply dense stereo with the epipolar constraint. The resulting epipolar curves are sinusoids [12]. The main disadvantage of cylindrical images is the high image distortion for top and bottom parts, corresponding to the sky and the ground. This increases the difficulty of the matching problem for these scene components. For this reason, we prefer to reproject a catadioptric image onto the 6 faces of a virtual cube and apply a classical two-view dense stereo (for convenience [10]) to two parallel faces of two cubes. The resulting epipolar curves become parallel lines, except for the faces which contain the epipoles: the epipolar lines intersect the epipole at the face center.

Thus, the method for a catadioptric image triple (ref, sec_1, sec_2) is as follows. First, two-view dense stereo is applied for a cube of ref and a cube of sec_1, and for a cube of ref and a cube of sec_2. Second, the stereo results are combined in the original catadioptric image ref. For each pixel of ref, the corresponding points in sec_1 and sec_2 are obtained from the mappings between cubes and the mappings between cubes and catadioptric images. A 3D point is obtained for the current pixel of ref by intersection of three rays using Levenberg-Marquardt. If the reprojection error is greater than a threshold (or if one of the three rays is not available), we can legitimately doubt the matching quality and no 3D point is retained. Last, many gaps (pixels of ref without 3D points) are filled with 3D points by interpolation.

4. Global 3D Model

Let L be the list of all local models that we have computed with the methods in Sections 2 and 3. Some parts of these models should be selected and merged into a global 3D model. First, Section 4.1 defines the virtual uncertainty $U_l(P)$ as a function of any 3D points P and any local models l of L. The virtual uncertainty is the usual uncertainty if P is reconstructed by l, and the virtual uncertainty extends it for other local models. Second, the local 3D model selection for reconstruction is presented in Section 4.2. Given a point P reconstructed by a local model l_0, we wish to know if l_0 is one of the local models in L with the smallest virtual uncertainties for P. If this is not the case, P should not be retained in the global model since a better model is available for P reconstruction. Even if l_0 minimizes the virtual uncertainty for P, the quality of P may be too bad for a global model (e.g. if P and all camera centers of l_0 are collinear points). For this reason, some "reliability" conditions for P are described in Section 4.3. Last, Section 4.4 describes how to use the selection criterion efficiently and obtain a global model from all local models of L.

4.1. Virtual Uncertainty Definition

The virtual uncertainty $U_l(P)$ for a 3D point P and a local model l is defined as follows. Let J_l be the Jacobian of the function $\Re^3 \rightarrow \Re^{2k}$ projecting P on all images $i_1 \cdots i_k$ from which l is reconstructed. We note that $C_l(P) = \sigma^2(J_l^\top(P)J_l(P))^{-1}$ is the covariance matrix associated to the reconstruction problem of P from images $i_1 \cdots i_k$, assuming independent and identical Gaussian noise of σ pixels for reprojection errors and zero uncertainties for cameras. An estimate of σ is provided by the reprojection errors of ray intersections (see Section 3). We define the virtual uncertainty $U_l(P)$ as the length of the major semi-axis of the uncertainty ellipsoid defined by covariance $C_l(P)$ and a probability p.

This definition of $U_l(P)$ assumes that P is "visible" in all images of l. If this is not the case, we consider that l can not reconstruct P accurately and choose $U_l(P) = +\infty$. The visibility definition depends on the knowledge we have of the global surface to be reconstructed. Obviously, a visible point P should be in the view fields of all images of l. Furthermore, the surface normal at P defines a half-space which should contain each camera location of l. We may also consider the global surface to be reconstructed itself as a possible occluder for P in each image of l, but this problem is not integrated in this paper.

The virtual uncertainty extends the "reliability" proposed by [4]: only the case $k = 2$ was considered and the reliability was defined by $\frac{\pi}{2} - \arccos(|d_1.d_2|)$, with d_1 and d_2 the directions of the rays which go across P and camera centers i_1 and i_2 respectively. Although intuitive (best/worst reliabilities are for two perpendicular/parallel rays), the reliability does not depend on the distance between camera centers. Uncertainty does.

4.2. Local 3D Model Selection for Reconstruction

The selection criterion is defined using the virtual uncertainty. Assume that a point P is reconstructed by a local model l_0. We would like to know if l_0 is one of the best local models in L to reconstruct P.

One can estimate $U_l(P), \forall l \in L$ and sort them in increasing order. If $U_{l_0}(P)$ is ranked at top n, l_0 is one of the best local models. A threshold $n > 1$ is useful to increase the density of points retained in the global model, and also to tolerate certain matching failures (false negatives). However, the virtual uncertainties at top n may have different magnitude orders. We avoid this drawback by estimating the relative uncertainty $U^r_{l_0}(P) = \frac{U_{l_0}(P)}{min_{l \in L} U_l(P)}$ and retain l_0 as one of the best models for P if $U^r_{l_0}(P) \leq 1 + \epsilon$. As has been previously stated, a threshold $\epsilon > 0$ is useful to increase the point density in the global model.

We note that these criterion selections are independent of the choice of probability p and noise σ used to define the virtual uncertainties: changing p or σ is just a multiplication of all $U_l(P)$ by a same scale factor.

4.3. Reliability Conditions

Assume that a point P is reconstructed by a local model l_0. Point P may be considered as unreliable and rejected if it is too far away from the camera centers of l_0, or if these centers and P are collinear points. These cases are possible even if l_0 is known to be one of the best local models. So, a reliability condition is useful to decide whether P should be included in the global model or not.

At first glance, the reliability condition might be "P is reliable if $U_{l_0}(P) < U_0$" with U_0 a threshold. However, the rendering of the resulting global model will suffer from two drawbacks in our visualization purpose (walkthrough in the scene). First, points at the foreground will have greater reprojection errors than points in the background. Second, too many points in the background will be discarded since point uncertainty increases with depth. So, a second definition for the reliability condition might be "P is reliable if $U_{l_0}(P) < U_0 d^\gamma(P, l_0)$" with $\gamma > 0$ and d the mean distance between P and the camera centers of l_0. Now, the two drawbacks are reduced, but this definition is heuristic and depends on thresholds p (a probability defined in Section 4.1), U_0 and γ.

A third definition is given by [4]: P is reliable if there is a camera pair (i, j) such that $\frac{\pi}{2} - \arccos(|d_i.d_j|) < \frac{\pi}{2} - \theta_0$ with θ_0 a threshold and d_i the direction of ray which goes across P and the i-th camera center of l_0. This definition requires only one threshold θ_0 (a lower bound for angle between two camera rays). It deals with both cases "P is too far away" and "P and camera centers of l_0 are collinear points".

4.4. From Local to Global Model

A straightforward use of the selection criterion for the global model generation has a time complexity proportional to $(\#L)^2(\#P)$ with $\#L$ being the number of local models in L and $\#P$ the (average) number of 3D points in a local model. This complexity may be high since $\#L$ is at least proportional to the image sequence length (sometimes many hundreds) and $\#P$ has the same magnitude order as the number of pixels in a catadioptric image (hundred of thousands).

The time calculation is reduced by subdividing each local model into small patches, and by applying the selection test (and the reliability test) only one time for each patch given a representative 3D point. Once a global set of patches is selected from all local models, usual merging/fusion methods [2, 17] may be used to reduce the redundancy of overlapped patches in space. In this paper, each patch is a square (assembled in rings of catadioptric images) and its representative point has the median depth of the patch. Although the reconstruction system does not yet include a merging method, the recovered 3D models are convincing as shown in the next Section. Two textured triangles are used for each patch.

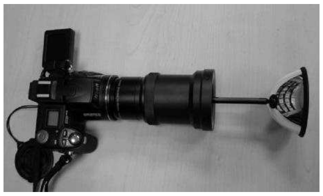

Figure 1. The “0-360” mirror with the Nikon Coolpix 8700.

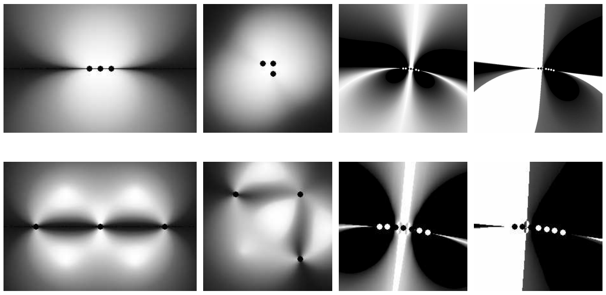

Figure 2. Uncertainty maps $P \mapsto U_l(P)$ or $U_l^r(P)$ in the horizontal plane (smaller values are white). The camera centers of local model l are black points in this plane. From left to right: $U_l(P)$ for 3 aligned cameras, $U_l(P)$ for 3 non-aligned cameras, $U_l^r(P)$ for a 3-view local model in the middle of a 7-view sequence, $U_l^r(P)$ for a 3-view local model at the beginning of a 7-view sequence. The neighborhoods of camera centers are zoomed in the second row using histogram normalization. Right: the black color is used if $1.5 < U_l^r(P)$, and camera centers are white if not in l.

5. Experiments

After the presentation of the experimental context in Section 5.1, synthetic experiments on local 3D model selection are presented in Section 5.2. Sections 5.3 and 5.4 give and discuss results obtained with the 3D scene reconstruction system for real sequences.

5.1. Context

The user moves along a trajectory on the ground with the catadioptric camera mounted on a monopod, alternating a step forward and a shot. A roughly equiangular catadioptric camera is used (the “0-360” mirror with the Nikon Coolpix 8700, shown in Figure 1). This is a non-central camera, such that the symmetry axes of camera and mirror are assumed to be the same. The mirror manufacturer gives the view field angles $\alpha_{up} = 37.5$ and $\alpha_{down} = 152.5$ degrees. These angles are not exactly known since they depend on the relative position between the mirror and the pinhole camera. Image dimensions are 1632×1224 pixels.

5.2. Local 3D Model Selection for Reconstruction

This part presents the experiment of the local 3D model selection (Section 4) with synthetic camera motions.

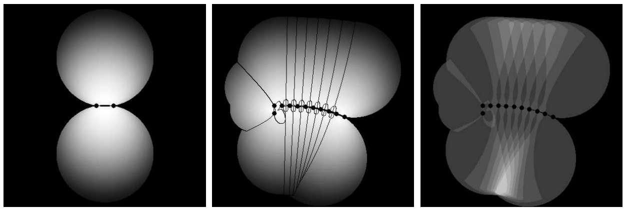

Figure 3. Left: the uncertainty map $P \mapsto U_l(P)$ of a 2-view local model l. Middle and right: all local models of 3 consecutive views are considered in a 11-view sequence. Middle: the uncertainty map $P \mapsto U_{l(P)}(P)$ with $l(P)$ the best local model (the local model minimizing $l \mapsto U_l(P)$). The regions of points P which have the same $l(P)$ are bordered by black lines. Right: the number of local models l accepted by the local model selection ($\epsilon = 0.1$) is encoded by a gray level (darkest gray: 1 local model, white: 9 local models). In all cases, pixels are black if the corresponding P does not satisfy the reliability condition with $\theta_0 = 10^o$.

First, the uncertainty map $U_l(P)$ is shown on the left of Figure 2 for two local models. As expected, $U_l(P)$ increases if point P goes near the line where all camera centers of l lies (if any) or if P goes away from camera centers of l. The first case (local model with 3 collinear view points) often occurs if the given camera trajectory is a smooth curve on the ground. Since this local model reconstructs many points close the given trajectory with a great deal of uncertainty, these points will appear very noisy in a rendered image even in the favorable context of a rendering view point in the neighborhood of the given camera trajectory. Selecting points in the global model with low uncertainty is a poor solution since distant points will also be rejected. High uncertainty is more acceptable for distant points than close points if the rendering view points is in the neighborhood of the given trajectory. The second local model (with 3 non-collinear view points) is a more favorable case since P is never aligned with all camera centers of l.

The relative uncertainty $U_l^r(P)$ is shown on the right of Figure 2 for 3-view local models in a 7-view sequence with nearly collinear camera centers. As described in Section 4, a point P reconstructed by l should have a small $U_l^r(P)$ to be accepted in the global model: $U_l^r(P)$ should be in $[1, 1+\epsilon]$. In the first case, the local model is not at the end of the whole sequence. We note that a kind of planar slice of the 3D space contains small values of $U_l^r(P)$, with $U_l^r(P) = 1$ at the central component. The planar slice goes across the middle camera of the local model, its thickness increases with the distance to the middle camera, and it is connected to both ends of the whole camera sequence. In the second case, the local model is at the end of the whole sequence and the planar slice is replaced by a large section of a half space. In both cases, we see that the local 3D model selection (i.e. thresholding $U_l^r(P)$) is not sufficient to decide whether the point of a local model should be accepted in the

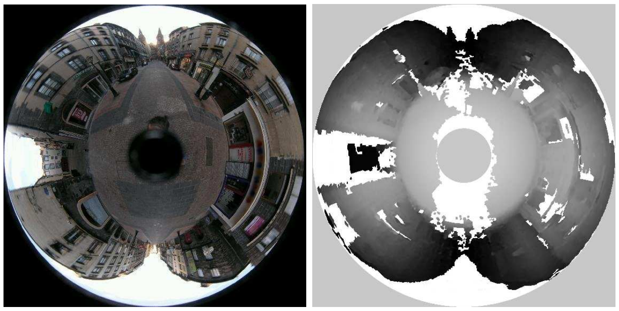

Figure 4. Reference image 211 and its depth map (white pixels have no depth).

global model.

Figure 3 gives results obtained by combining the local 3D model selection and the angle-based reliability condition defined by θ_0 (see Section 4.3). In all cases, black pixels outside the gray-white region are points which do not satisfy the reliability condition with $\theta_0 = 10^o$. On the left, two circles are the set of points P such that the angle between the two camera centers are equal to θ_0. In the middle and on the right, we consider all 3-view (consecutive) local models of an 11-view sequence. The space partition defined by the best local model is given in the middle. On the right we see the number of local models satisfying the local model selection $\epsilon = 0.1$ with $\theta_0 = 10^o$ for a given P (i.e. the number of possible reconstructions for P). We note that this number increases if P goes out of the camera trajectory.

5.3. Old Town Sequence

It requires about 52 minutes to take the 354 images of the Old Town sequence. Some images are shown in Figure 5. The trajectory length is about $(35 \pm 5cm) \times 353 = 122 \pm 17m$ (the exact step lengths between consecutive images are unknown). The radii of large and small circles of the catadioptric images are 570 and 102 pixels. A rectangular field of view of $\frac{\pi}{4} \times \frac{\pi}{4}$ has approximately 260×210 pixels in the horizontal and vertical directions in theses conditions.

SfM is the first step of the method as described in Section 2. 59859 3D points are automatically reconstructed with 380947 points in images satisfying the geometry. A top view of the result is proposed in Figure 5. The final RMS error is 0.74 pixels, and a 2D point is considered as an outlier if the reprojection error is greater than 2 pixels. All calculations are started with an inaccurate calibration ($\alpha_{up} = 40, \alpha_{down} = 140$ degrees) to experiment the robustness of the SfM and the behavior of the calibration refinement. The recovered radial distortion $r(\alpha)$ is very close to a linear function (as expected) and the view field angles are improved: $\alpha_{up} = 35.5, \alpha_{down} = 152.2$ degrees.

The second step is the calculation of all local 3D models (see Section 3). A local model is build for each triple of consecutive images of the sequence, and is defined by a list of about 10000 patches partitioning the reference image. Figure 4 shows the depth map and the reference image of a local 3D model. Last, a global 3D model is obtained by selecting patches of all local models using the local model selection scheme (Section 4.2) and a reliability condition (Section 4.3). A patch of a local model is accepted if its representative 3D point (introduced in Section 4.4) satisfies (1) the angle-based reliability condition with $\theta_0 = 5^o$ and (2) the local model selection with $\epsilon = 0.1$. Only 24% of

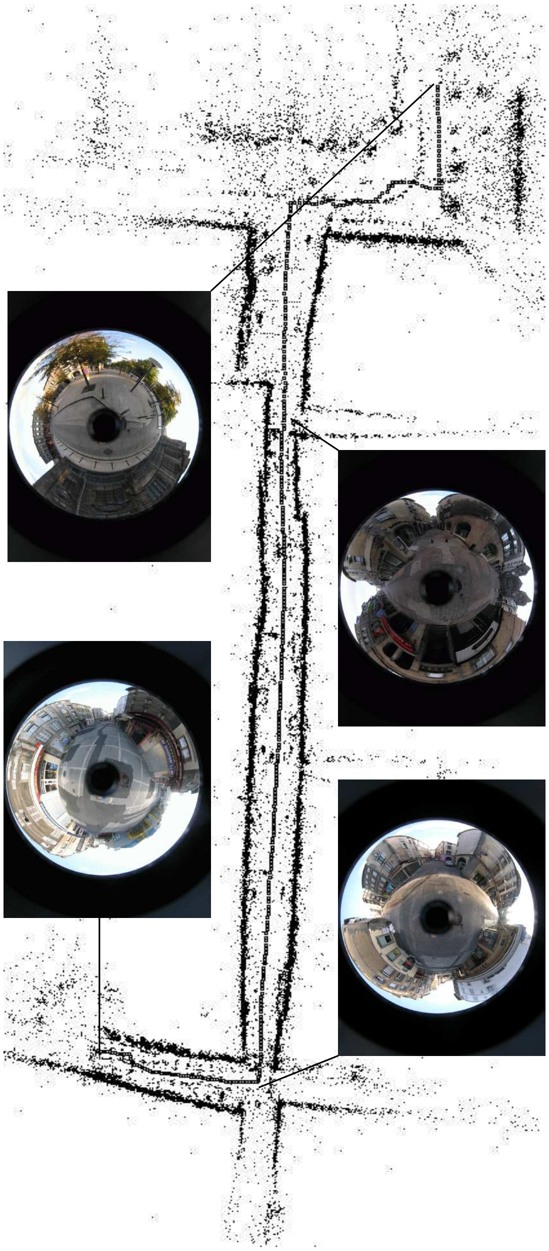

Figure 5. Some images of the Old Town sequence and a top view of the reconstruction using SfM, including 354 cameras (black squares) and 59859 points (black points).

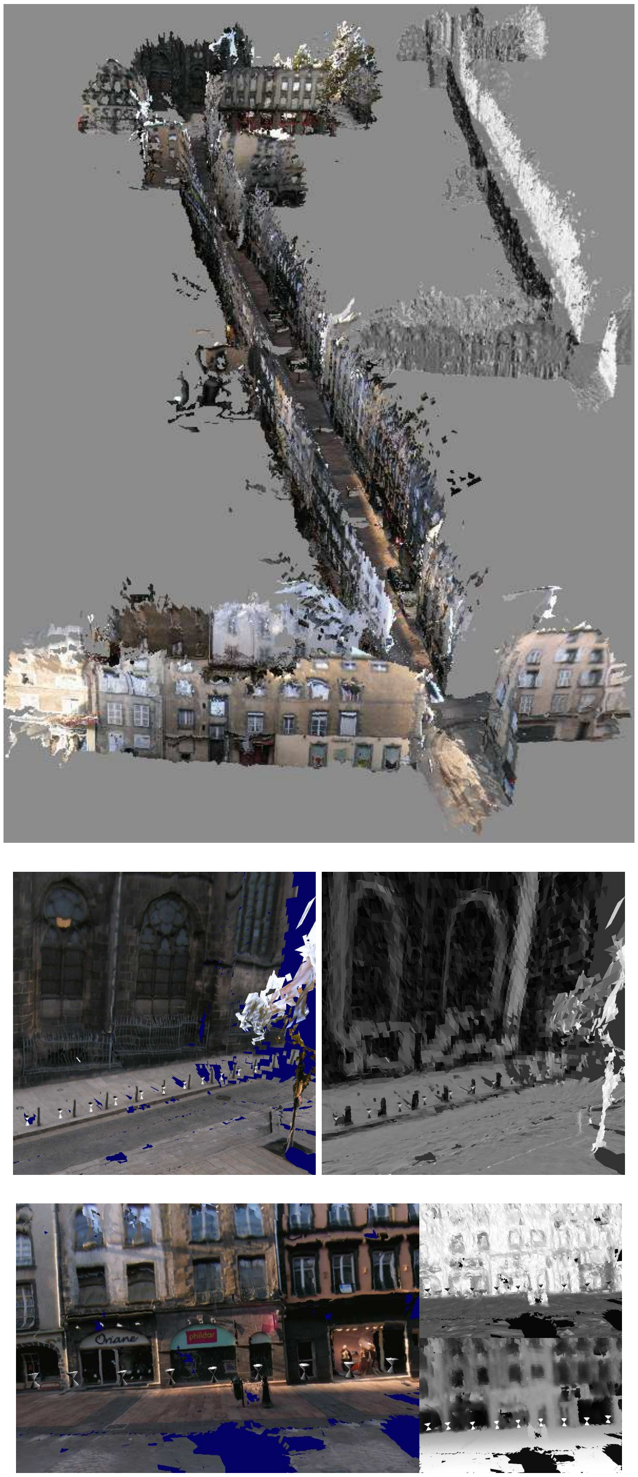

Figure 6. Three textured-mapped views of the Old Town global model (including cameras), obtained with local 3D model selection and reliability conditions. Patch orientations (or depth map) are also drawn using gray levels.

patches (819767 patches) are retained in the global model.

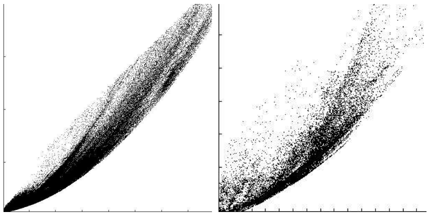

Figure 7. Quantitative results for uncertainty of global models (Old Town on the left, and Small Room on the right). Each patch has coordinates $(d_l(P), U_l(P))$ with P the representative point of the patch, $d_l(P)$ and $U_l(P)$ the mean depth and the uncertainty ($p = 90\%$) of P by the local model l which reconstructs P. The ranges (cm) are $[103, 890] \times [0.63, 39.9]$ (left) and $[55.6, 212] \times [0.72, 7.56]$ (right), and exclude 0.5% of patches.

Figure 6 shows three views of this model. It is not difficult for the reader to match the oblique view of the full model (top of Figure 6) with the top view of the SfM result in Figure 5. Furthermore, the joint video is composed of two parts. The first part shows all original images. Panoramic images are also given to help understand the scene, but they are not used by the method. A major lighting change occurs at frame number 295. The second part is a walkthrough in the whole scene in the neighborhood of the original view points. We consider that the results are acceptable in textured areas in spite of the low resolution (about 210-260 pixels for an angle in space of $\frac{\pi}{4}$). Currently, the main problem is due to many parts of walls and streets which are not textured enough to be matched and reconstructed. A second problem is due to the patches which have regular size (about 8×8 pixels) and locations in images. The resulting approximation of occluding contours is sometimes very crude.

Quantitative results for the 90% point uncertainties are given in Figure 7 for the global model. In $99, 5$ percent of cases, the uncertainty increases from 0.6 cm to 40 cm while point depth grows from 1 m to 9 m.

5.4. Small Room Sequence

The Small Room sequence is composed of 18 indoor images, such that the step length between two consecutive view points is about 15 cm (exact values are unknown). SfM results are: 2974 reconstructed points, 15096 image points satisfying the geometry with RMS=0.73 pixels, $\alpha_{up} = 37.8$ and $\alpha_{down} = 152.7$ degrees (with $\alpha_{up} = 40$ and $\alpha_{down} = 140$ as initial values).

Only 26% of patches (35709 patches) are retained in the global model after local model selection ($\epsilon = 0.1$) and reliability constraint ($\theta_0 = 5^o$). Figure 7 gives quantitative uncertainties for the global model and Figure 8 shows many views of the global model. The main objects are easy to recognize. As mentioned in the Old Town example, many parts of the room are not reconstructed due to the lack of texture, and the patches ignore the occluding contours.

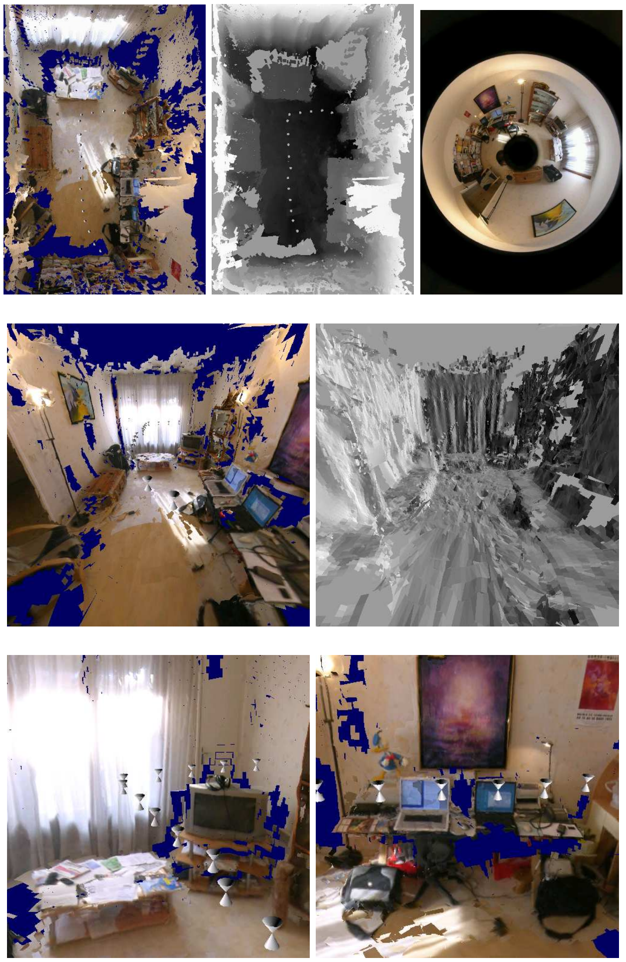

Figure 8. Four textured-mapped views of the Small Room global model, obtained with local 3D model selection and reliability conditions. Depth map and patch orientations are drawn on the top and in the middle. One image of the sequence is also shown.

We note that a single local model at the middle of the sequence is not sufficient to reconstruct all objects, and more local models are welcome to reduce the uncertainties.

6. Conclusion

A fully automatic 3D modeling method from a catadioptric image sequence is proposed. First, still images are acquired by a roughly equiangular camera. Second, the sequence geometry (including some intrinsic parameters) is successfully estimated with a central camera model. Third, many local 3D models along the sequence are reconstructed with 3 views. The image reprojection on a virtual cube (instead of a cylinder) allows the use of classical dense stereo methods and facilitates the reconstruction of the ground. Last, the global 3D model is obtained by applying a local model selection: each local model is partitioned in patches, and a patch is rejected if an other local model is available to reconstruct the patch with less uncertainty. Such 3D models have never been obtained before with catadioptric cameras, because matching is always difficult in practice and the last step has been largely ignored in previous studies.

Many improvements are possible and include: better use of visibility in the local model selection, a better choice of patches near occluding contours, improving matching in low textured areas, and more investigations on the choice of local models to calculate (decrease their number, integrate different baselines for a same reference image).

References

[1] R. Bunschoten and B. Krose. Robust scene reconstruction from an omnidirectional vision system. *IEEE Transactions on Robotics and Automation*, pages 351–357, 2003.

[2] B. Curless and M. Levoy. A volumetric method for building complex models from range images. *SIGGRAPH*, 30, 1996.

[3] K. Daniilidis. The page of omnidirectional vision. www.cis.upenn.edu/˜ kostas/omni.html.

[4] P. Doubek and T. Svoboda. Reliable 3d reconstruction from a few catadioptric images. In *OMNIVIS'02*.

[5] S. Fleck, F. Busch, P. Biber, W. Strasser, and H. Andreasson. Omnidirectional 3d modeling on a mobile robot using graph cuts. In *IEEE ICRA'05*.

[6] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2000.

[7] S. Kang and R. Szeliski. 3-d scene data recovery using omnidirectional multibaseline stereo. *IJCV*, 25(2), 1997.

[8] V. Kolmogorov and R. Zabih. Computing visual correspondances with occlusions using graph-cuts. In *ICCV'01*.

[9] M. Lhuillier. Automatic structure and motion using a catadioptric camera. In *OMNIVIS'05*.

[10] M. Lhuillier and L. Quan. Match propagation for image-based modeling and rendering. *IEEE PAMI*, 24(8), 2002.

[11] M. Lhuillier and L. Quan. A quasi-dense approach to surface reconstruction from uncalibrated images. *IEEE PAMI*, 27(3), 2005.

[12] L. McMillan and G. Bishop. Plenoptic modelling: an image-based rendering system. In *SIGGRAPH'05*.

[13] B. Micusik and T. Pajdla. Structure from motion with wide circular field of view cameras. *IEEE PAMI*, 28(7), 2006.

[14] D. Nister. *Automatic Dense Reconstruction from Uncalibrated Video Sequence.* PhD thesis, Royal Institute of Technology KTH, Stockholm, Sweden, 2001.

[15] M. Okutomi and T. Kanade. A multiple-baseline stereo. *IEEE PAMI*, 15(4), 1993.

[16] M. Pollefeys, L. V. Gool, M. Vergauwen, F. Verbiest, K. Cornelis, J. Tops, and R. Koch. Visual modeling with a hand-held camera. *IJCV*, 59(3), 2004.

[17] M. Soucy and D. Laurendeau. A general surface approach to the integration of a set of range views. *IEEE PAMI*, 17(4), 1995.

[18] D. Strelow, J. Mischler, S. Singh, and H. Herman. Extending shape-from-motion estimation to noncentral omnidirectional camera. In *IEEE/RSJ IROS'01*.