Semi-interactive construction of 3D event logs for scene investigation
Dang, T.K.

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
3D models of indoor scenes have many applications. But automatically building a 3D model is a challenge in practice. Many methods exist for outdoor scenes, but indoor scenes are more difficult. Due to the limited space for movement, the input is very often close to being degenerate. For such input making a model is impossible without smart input processing. This chapter presents a framework for modeling of indoor scenes. The key idea is to segment the input video into general and degenerate parts, and instead of throwing away degenerate segments, use them to build a more complete model. To do so, we first develop a frame filtering method that preserves all the information of the input while reducing the computational load. Results show that the remaining frames, significantly smaller in number, are nearly as informative as the original input and are suitable for the later steps of the modeling process. By utilizing the information from degenerate segments, our framework indeed produces a model that is far more complete than the ones obtained with traditional frameworks.

*A preliminary version of this chapter has been published as “Dealing with degenerate input in 3D modeling of indoor scenes using handheld cameras” in IEEE International Conference on Multimedia and Expo [31]
4.1 Introduction

In many applications such as real estate management, crime scene investigation, and tourism [54, 141], having a 3D model of the scene enhances human-computer interactivity as you can navigate through the scene as if you are there. In addition, if 3D models are made they can be used to search for similar models [33], form the basis for efficient transmission [25], or be rendered on mobile devices [10] to provide mobile navigation.

One way of obtaining 3D information about a scene is by capturing it using specific devices, e.g. laser-scanners or stereo cameras. Those devices, however, are not commonly available and are not flexible. Widely available handheld cameras have the portability and flexibility to capture scenes in various conditions. For that reason, 3D modeling from video sequences captured by handheld cameras is our preferred choice.

Good results have been shown for modeling outdoor scenes and isolated objects [121, 122, 58, 24]. Applying those methods to indoor scenes, however, is not trivial. The 3D coordinates of a point in the scene can only be recovered if it has been seen from two different views. The larger the baseline, i.e. the distance between the views, the better the result. For 3D modeling a suitable camera movement is therefore an arc-track with the camera pointing towards the target. This movement maximizes the baseline while keeping the target within the range of view. Both outdoor scenes and isolated objects are easy to capture in this configuration since the space of movement is quite open. Within an indoor scene the space is limited and the cameraman is close to the target. The typical moves in this case are tilt, pan, and zoom interlaced by dolly. These moves are not suitable for 3D modeling as they produce input where the baseline is very small. In general we can say that the moves typically made by cameramen in indoor scenes are quite different from those used in outdoor scenes.

Theoretically, it is impossible to recover depth information from video captured with tilt, pan, and zoom as there is no translation component [61]. For indoor scenes proper handling of degenerate input is thus essential. In previous work, e.g. [117, 125], degenerate input is either discarded or avoided by applying a strict capturing guideline. The solution guarantees that the remaining frames are good for modeling, but raises a problem for indoor scenes. Since the distance from the cameraman to the target is small, only a small part the scene is captured in each frame. Thus the remaining frames only cover a small part of the scene limiting the completeness of the model. The arc-track move of the strict capturing guideline is typically done with a camera mounted on a track, which is inappropriate for a person capturing a scene with a handheld camera. So, in order to make 3D modeling from video work for indoor scenes, there should be a method that enables people to capture the scene in a convenient way with the degeneracy handled automatically.

In this chapter we propose a framework for reconstruction of indoor scenes particularly designed to deal with the degeneracy problem. By first filtering the input and applying segmentation, we identify (near) degenerate input. Instead of throwing away those data, we use proper methods to recover the corresponding camera motions. The information is used in later stages to improve the completeness of the reconstructed model. Consequently the modeling process is more robust against degeneracy.
4.2 Background

4.2.1 The 3D modeling process

The process of 3D modeling with handheld cameras [121, 61] usually consists of four steps (Figure 4.1): feature processing, structure recovery, stereo mapping, and finally model creation.

The first step, feature processing, detects image features, which are then matched throughout the sequence. The most common features used are points. Feature points can be detected and matched quite accurately [93, 12]. In the structure recovery step, the initial correspondences are used to compute basic geometric constraints, which define the relation of geometric elements (e.g. points, lines) between views. For two-view geometry, in case of general input they are represented by the fundamental matrix [95] or in the degenerate input case, by a homography [61]. Since the initial correspondences may contain outliers, those constraints are computed using a robust method, e.g. RANSAC [48]. Based on those constraints a projective structure, i.e. the 3D coordinates of feature points up to a projective transformation, is initiated and extended. Then the metric upgrading transformation can be recovered, e.g. using a method in [115], [124], or [122]. Using that transformation the projective structure is upgraded to a metric one, i.e. Euclidean structure up to a scale factor. To recover the complete structure, stereo mapping must be done over the frame sequence. This step densely maps all points from one frame to others. It can be done using a variety of methods [132]. Finally the model creation step turns the dense point cloud into a 3D mesh with texture extracted from frames mapped onto it.

![Figure 4.1: The traditional framework for uncalibrated 3D modeling with handheld cameras should be extended to deal with those cases.](image)

Among those steps, the structure recovery is very important since it recovers the “skeleton” of the final model. More importantly this step is the critical point determining whether the modeling succeeds or fails. Failure in this step is a consequence of degenerate input not detected in previous steps. As mentioned in the Introduction, degeneracy is very often met while doing modeling of indoor scenes. We will therefore elaborate on it in the next section.

4.2.2 Degeneracy

There are two kinds of degeneracy: (i) degenerate structure, and (ii) degenerate motion. The first class depends on the structure of the scene itself and the viewpoints used [74]. The latter depends on camera motion only and thus can happen with any scene and is also subject of recent research [26, 48]. The most frequently met degenerate structure case is the pure planar case. It happens, for example, when filming along a wall. It was first noticed in the paper of Torr et al [156]. They propose to use the geometric robust information criterion (GRIC) [154] to detect the situation...
from the two-view constraints. Later Pollefeys et al [117] extended GRIC for the three-view constraints to detect a specific case of general input not detected by two-view constraints. For a more reliable result Torr proposed multiple hypothesis tracking [155] but this is very costly. Within this chapter, for simplicity we use only two-view constraints and leave the 3-view constraints and the multiple hypothesis tracking for future work.

Another degeneracy case occurs when a scene has small objects in front of a planar structure. Since most of the feature points are on the same plane, the fundamental matrix computation using RANSAC usually terminates before finding the correct result. Thus this case is often misidentified as a pure planar structure. This problem is solved by using Degenerate RANSAC [26] or Quasi Degenerate RANSAC [48]. Those methods integrate model testing within RANSAC to select the correct geometric model. Degenerate RANSAC is used in our fundamental matrix computation.

<table>
<thead>
<tr>
<th>Camera move</th>
<th>Notation</th>
<th>Functionality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tilt</td>
<td><img src="icon.png" alt="Tilt" /></td>
<td>Vertical overview from observation point</td>
</tr>
<tr>
<td>Pan</td>
<td><img src="icon.png" alt="Pan" /></td>
<td>Horizontal overview from observation point</td>
</tr>
<tr>
<td>Dolly</td>
<td><img src="icon.png" alt="Dolly" /></td>
<td>Moving between observation points</td>
</tr>
<tr>
<td>Track</td>
<td><img src="icon.png" alt="Track" /></td>
<td>Moving between observation points</td>
</tr>
<tr>
<td>Arc-track</td>
<td><img src="icon.png" alt="Arc-track" /></td>
<td>Detail inspection</td>
</tr>
</tbody>
</table>

**Figure 4.2:** Possible basic/elementary camera moves using handheld cameras and their functionality. The first two are the more common ones, but they produce degenerate input.

Motion degeneracy is created by camera moves without a translation component such as data coming from a panning camera. Yet, of the common camera moves (Figure 4.2), pan and tilt are the preferable ones since they produce more steady videos. For this case only the motion can be recovered [61]. GRIC is also the means to detect this case. But there is a lack of tools to distinguish it from the planar structure case. Although in [154] two distinguishing techniques were suggested, they are either not practical or too costly. So in previous work for outdoor scenes or isolated objects, e.g. [125, 117] this part of the input is either discarded or avoided by a capturing guideline. Figure 4.4 illustrates the relation between filming moves, characteristics of the scene and degeneracy.

In summary, degeneracy is common in input coming from recording of indoor scenes. Thus, in order to reconstruct indoor scenes, we should first identify degeneracy and then use a proper reconstruction strategy to utilize degenerate input.
Figure 4.3: Indoor filming path of a professional investigator. Sample frames at each point are given on the right. A few dolly, arc-track moves produce general input. All other moves, i.e. pan and tilt, produce motion degeneracy.
4.3 Modeling Framework for Indoor Scenes

This section introduces our framework for 3D modeling of indoor scenes. To filter out redundant data and achieve more complete 3D models, we extend the framework in Figure 4.1 to the one depicted in Figure 4.5. The feature processing step is extended with the two steps, Frame filtering and Frame segmentation. The frame filtering step pre-processes input by filtering out the redundant data and is discussed in more detail in 4.3.1. The frame segmentation step, discussed in 4.3.2, separates the filtered input into general and degenerate segments. To reconstruct more complete 3D models, the structure recovery step is decomposed into a Core structure recovery step and a Structure extension step. The core structure recovery uses the general part of the input to recover the initial core structure and motion. After that the structure extension iteratively tries to add other frames to extend the structure. Those steps are detailed in 4.3.3. The rest of the framework is the same as the general one.

4.3.1 Frame Filter

Since it is hard to identify degeneracy from very close frames [156], frame filtering is an essential preprocessing step before frame segmentation. Besides, it reduces the data redundancy. The data redundancy in a video maybe useful to make a more complete and detailed model, but requires much more computing power. Filtering out redundant frames, e.g. occurring when filming with slow movement, helps to reduce the computational load. Finally, the frame quality problem, e.g. disregarding frames having motion blur, should also be handled in this step.

In previous work, frame filtering is usually merged with frame segmentation. From the last selected frame, the next frame is selected using two traditional criteria: the number of
matches and the baseline. The larger the number of matches the more reliable the computation of geometric constraints. And, as mentioned, the large baseline helps to recover depth accurately. The number of matches \(n\) can be computed directly. Since the baseline is unknown, another metric, the error of the homography fitting between two views \(\varepsilon_H\), is often used as an indicator of viewpoint change. The product of \(n\) and \(\varepsilon_H\) can then be used as the frame filtering criterion \([117, 141]\). It, however, has the disadvantage that degenerate frames will be discarded because the \(\varepsilon_H\) is small no matter how different the two viewpoints are. This problem is either ignored as in \([141]\) or solved using GRIC, but this only works if there is no rotation \([117]\). We propose to use the difference of views as the second criterion. It is represented by the average length of the feature point motion vector \(\bar{m}\). This new measure increases as the view changes for both translation and rotation. The new criterion for frame filtering is:

\[
c_f = n \cdot \bar{m} \quad (4.1)
\]

For simplicity, we assume that the input video is one continuous shot and frames have similar quality. In case of dramatically varying quality, the degraded frame should be filtered out first. The first step of the filtering procedure is to find the second frame by calculating the frame filtering criterion \(c_f\) of the first frame \(s_0\) with respect to any frame in a certain range \(r_0\). The parameter \(r_0\) should be large since the speed of camera movement is unknown in the beginning. The one that produces the highest \(c_f\) is selected as \(s_1\). Other in-between frames are discarded. The next tentative \(s_{i+1}\) is computed as \(s_{i+1} = s_i + (s_i - s_{i-1})\). The procedure is the same for finding frames in subsequent steps except that, to improve the speed, we limit the search range to \([s_{i+1} - r_1, s_{i+1} + r_1]\). The parameter \(r_1\) can be smaller than \(r_0\) since we expect that image content changes gradually.

### 4.3.2 Frame Segmentation

The frame segmentation can be done by examining the multiple-view constraints in the general and degenerate cases. As mentioned, the two-view constraints in general and degener-
erate cases are the fundamental matrix $F$ and the homography $H$ respectively. $F$ defines a point-to-line mapping while $H$ defines a point-to-point mapping. Let $x_i$ and $x_j$ denote the homogeneous coordinates of points in two views, $F$ and $H$ are $3 \times 3$ matrices defined by:

$$x_j^T F_i x_i = 0$$

$$x_j = H_i x_i$$

(4.2)

(4.3)

Input to estimate $F$ and $H$ are pairs of 2D points, meaning that the dimension of input $r$ is 4. $F$ represents a 3D relation where $H$ represents a 2D relation. In other words, they define a 3D and a 2D surface in the 4D input space. Hence, the dimension of the constraints, $F$ and $H$ respectively, are 3 and 2. The degrees of freedom $k$ of $F$ and $H$ are 7 and 8 respectively [154]. In short, any constraint $M$ being evaluated by GRIC is represented by a set of parameters $(d, r, k)$.

To identify whether a frame is general or degenerate we use the GRIC [154] to evaluate the quality of fit of those two models on the data. GRIC takes into account both the model fitting error and the model complexity. Given the set of residuals $E = [e_1, e_2, ..., e_n]$ when fitting a model between two frames, and the standard deviation of the residuals $\sigma$, GRIC $g$ is formulated as:

$$g(M, E) = \sum_{e_i \in E} \min\left(\frac{e_i^2}{\sigma^2}, \lambda_3 (r-d)\right) + (\lambda_1 nd + \lambda_2 k)$$

(4.4)

The first term of (4.4), derived from fitting residuals, is the model fitting error. The minimum function used in $\rho$ is meant to threshold outliers. The second term, consisting of model parameters, is the model complexity. $\lambda_1$, $\lambda_2$, and $\lambda_3$ are parameters steering the influence of the fitting error and the model complexity on the criteria. In [156] suggested values for $\lambda_1$, $\lambda_2$, and $\lambda_3$ are $\log(r)$, $\log(rn)$ and 2 respectively.

The model with smaller GRIC is assumed the correct one. Given $g(H, E)$ and $g(F, E)$ as GRIC values for the homography model and fundamental matrix model, the model of choice is decided as follows:

$$model(E) = \begin{cases} F, \Delta_G > 0 \\ H, \Delta_G \leq 0 \end{cases}$$

(4.5)

where

$$\Delta_G = g(H, E) - g(F, E)$$

(4.6)

This criterion was successfully used in several papers [156, 117, 125].

Here we apply GRIC on the remaining frame sequence after filtering to identify degenerate segments. As noticed in [117], two frames with less than 50 matches are not reliable for structure recovery. So we assign those as degenerate frames to guarantee the best quality of the core structure. The result of this step together with the correspondences, $F$, and $H$ are passed to the core structure recovery and the structure extension.

### 4.3.3 Structure recovery

To recover the structure of the whole scene, we first recover parts of it and then try to extend them. From general segments, following the method described in [117], the core structure
4.4 Results

In this section we evaluate the frame filtering, and the frame segmentation from 4.3.1 and 4.3.2. In 4.4.3 we show that the structure extension helps to reconstruct a more complete model.

4.4.1 Evaluation of frame filtering

The objective of frame filtering is to reduce the number of frames as much as possible while keeping the input informative and useful. That means the remaining frames should cover the entire scene and it must be able to build a 3D model from them. In this work, for convenience of evaluation, we separate the two cases: general and degenerate input and evaluate them in isolation. For the general parts, which are normally very small compared to the complete video, the usability is more important since they are used to initiate the core structure. For degenerate parts, which are actually used to capture the scene, being informative is more important.

To evaluate the general case, we want to know the camera motion corresponding to each frame in order to evaluate the usability. The ALOI dataset [53], which contains objects (Figure 4.6) with motion ground truth, is suitable for this purpose. ALOI objects are filmed at 5 degree steps on a turntable. That is equivalent to the case of a camera moving around the object. For the degenerate cases, we use degenerate segments extracted from a self-recorded video of an indoor scene. This video is captured in a normal indoor environment without using a tripod.

Figure 4.6: Example objects of the ALOI dataset [53] captured in known configuration are used to evaluate Frame filtering in the case of general input.
In both experiments, we check the number of matches $n$ and the average length of the feature motion vector $\bar{m}$ (equation 4.1). The number of matches $n$ indicates whether remaining frames are useful for the structure recovery. The $\bar{m}$ indirectly shows how large the baseline is in the general case. Furthermore, in the general case, we take the advantage of having a ground truth for view angle to check the change of viewing angle $\alpha$. Because adding a new view requires 3D-2D correspondences, there must be features visible from three consecutive views. Since SIFT [93], which is used in our implementation, is able to detect features up to about 50 degrees of change of viewpoint angle [103] the suitable angle between two views should be smaller than 25 degrees. In practice the tolerable change of viewpoint angle is smaller since the objects are not a simple planar surface as in [103].

As stated before, being informative is more important in the second case. To evaluate, we manually define a representative surface in the scene. Given a set of frames and the representative surface of a scene, we project all the frames onto the surface and measure the coverage to tell how much the frame set covers the scene. If the kind of surface, its position, and orientation are well chosen, we can even tell whether the frames are evenly distributed over the scene. For the motion degenerate case, we use the view plane of the frame whose motion is closest to the average motion of all the frames. For the structure degeneracy, i.e. planar case, we use the view plane that is parallel to the scene plane, which contains all objects of the scene. To evaluate the informativeness of a frame set, we measure the ratio of the coverage of a frame set over the coverage of the original data. The coverage of a frame sequence $S = \{s_i\}_{i=1,N}$ is defined as:

$$\omega(S) = \bigcap_{\forall s_i \in S} H_i A_i$$

where $A_i$ is the area covered in the scene by frame $s_i$, and $H_i$ is the homography that transforms frame $s_i$ to the representative plane. The ratio of the coverage $\Omega$ for the frame sequence $S$ is defined as:

$$\Omega(S) = \frac{\omega(S)}{\omega(O)} \times 100\%$$

where $O$ is the original frame sequence.

The usability of the degenerate input is exhibited by the success of computation of homographies of consecutive pairs of frames, which are required by specific structure or motion recovery algorithms applied on degenerate input. We use $s_H$ as the notation of the successful homography computation rate, which is defined as:

$$s_H = \frac{S_H}{N-1} \times 100\%$$

where $S_H$ is the number of successful homography computations and $N$ is the total number of frames.

We run the filtering method described in section 4.3.1 with $r_0$ and $r_1$ set to 30 and 10 respectively. For the ALOI dataset, that means search in range of 50 degrees, more than SIFT descriptor [93] can match features. For the degenerate cases, these parameters are large enough so that the number of matches $n$ is nearly zero at each end of the searching window. Overall, judging on $n$ and $\bar{m}$, the results are very reasonable (Table 4.1 and 4.2). The number
4.4. Results

Table 4.1: Evaluation of frame filtering in the general input case using 20 ALOI objects

<table>
<thead>
<tr>
<th></th>
<th>Min</th>
<th>Max</th>
<th>Avg.</th>
<th>Sd.</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>11.0</td>
<td>740.0</td>
<td>187.1</td>
<td>n/a</td>
</tr>
<tr>
<td>(\bar{m}) (pixel)</td>
<td>18.4</td>
<td>81.9</td>
<td>44.8</td>
<td>18.2</td>
</tr>
<tr>
<td>(\alpha) (degree)</td>
<td>5.0</td>
<td>40.0</td>
<td>16.3</td>
<td>9.1</td>
</tr>
</tbody>
</table>

Table 4.2: Evaluation of frame filtering in the degenerate input case (self-captured video)

<table>
<thead>
<tr>
<th></th>
<th>Motion deg.</th>
<th>Structure deg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Our sel. (n)</td>
<td>321.9</td>
<td>92.1</td>
</tr>
<tr>
<td>(\bar{m}) (pixel)</td>
<td>68.8</td>
<td>9.4</td>
</tr>
<tr>
<td>Uniform sel. (n)</td>
<td>268.7</td>
<td>153.7</td>
</tr>
<tr>
<td>(\bar{m}) (pixel)</td>
<td>57.6</td>
<td>16.2</td>
</tr>
</tbody>
</table>

Table 4.3: Completeness \(\Omega\) of our selections and the uniform selections

<table>
<thead>
<tr>
<th></th>
<th>Motion deg.</th>
<th>Structure deg.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(\Omega) (%)</td>
<td>(s_H) (%)</td>
</tr>
<tr>
<td>Our sel.</td>
<td>97.2</td>
<td>100.0</td>
</tr>
<tr>
<td>Uniform sel.</td>
<td>95.6</td>
<td>79.5</td>
</tr>
</tbody>
</table>

of matches is large enough to reliably compute geometric constrains and the motion is fairly large. Particular criteria for each case also show good results.

In the general input case, 20 ALOI objects are used for the experiment. The average angle between two consecutive selected views is 16.3, which is reasonable in comparison with the “theoretical” limit of 25 degrees.

The rotation and the planar structure videos, corresponding to motion and structure degenerate cases, have 272 and 353 frames respectively. From those 38 and 39 frames remain after filtering. The \(\Omega(S)\) of the two selections are 97.2 and 96.5 percent respectively. They are slightly better than the uniform distribution selections, i.e. the same number of frames with a fixed step, whose \(\Omega(U)\) are 95.6 and 96.3 percent. But judging on the successful homography computation rate \(s_H\) our selection is better. Our approach has a \(s_H\) of 100 percent while uniform selections of motion and structure degeneracy cases have 79.5 and 87.5 percent respectively. That shows our selections are more related and more suitable for motion recovery than the uniform distribution. The standard deviation of \(n\) and \(\bar{m}\) in table 4.2 also confirms this observation.
4.4.2 Evaluation of frame segmentation

After the frame filtering, from the same 4225-frame video used in the previous subsection, we have 285 remaining frames. Based on the computation of GRIC of the $H$ and $F$ models ($g_H$ and $g_F$ respectively) we can identify the degenerate segments. For the evaluation we compare it with the manual segmentation of the video and calculate the false segmentation rate $f$:

$$f = \frac{N_F}{N} \times 100\%$$  \hspace{1cm} (4.10)

where $N_F$ is the number of incorrectly segmented frames and $N$ is the total number of frames.

Figure 4.8-a shows which model should be selected based on GRIC values. After the post-processing, the final result is shown in Figure 4.8-b. A correct point must fall onto the manual segmentation line. The false segmentation rate is $f = 2.8\%$, which is $8/285$ frames, of which 7 general frames are misidentified as degenerate. Those are at model switching points, which are hard to mark precisely even for humans. Only one degenerate frame is misidentified as general frame. This does not affect the next step since we only use general segments of at least three-frame length to initiate the core structure.

4.4.3 Final result

In Figure 4.9, we show that the proposed framework can reconstruct a scene in a more complete way than the standard methodology. We can clearly see the improvement in term of completeness. Comparing to the panorama of the scene, which is built from images and captures approximately the same as the video, we see that most of the elements of the scene are reconstructed.
4.5 Conclusion

In this chapter we have introduced a framework to do 3D modeling from video of indoor scenes recorded by a handheld camera. Using our new criterion for frame filtering, degenerate input are kept and used for modeling if possible. Hence, the final result is more complete than with existing methods. Our proposed framework also makes 3D modeling more robust by explicitly handling the degenerate input. Even in the outdoor case, it would provide users with a more comfortable, natural way of capturing. This work is a step forward in 3D modeling from unstructured videos.

Figure 4.8: (a) $\Delta G$ over the frame sequence. Frames with a value above zero are identified as general and others are degenerate. (b) Comparison between automatic GRIC-based segmentation and the manually generated ground truth. The few misidentified frames, are mostly due to the imperfect manual segmentation and the limitation of the GRIC.

4.5 Conclusion

In this chapter we have introduced a framework to do 3D modeling from video of indoor scenes recorded by a handheld camera. Using our new criterion for frame filtering, degenerate input are kept and used for modeling if possible. Hence, the final result is more complete than with existing methods. Our proposed framework also makes 3D modeling more robust by explicitly handling the degenerate input. Even in the outdoor case, it would provide users with a more comfortable, natural way of capturing. This work is a step forward in 3D modeling from unstructured videos.
Figure 4.9: (a) Two separated core structures built from general input segments, and (b) the final result after applying structure extension (models are generated using PMVS package [51]). A more complete structure is recovered thanks to the structure extension. The structure extension in fact even merges the two isolated structures of (a) into one model. Comparing to the panorama of the scene (c), most of the elements are present in the reconstructed model.