3D Spatial Layout Propagation in a Video Sequence

Alejandro Rituerto[1(✉)], Roberto Manduchi[2], Ana C. Murillo[1], and J.J. Guerrero[1]

[1] Instituto de Investigación en Ingeniería de Aragón, University of Zaragoza, Zaragoza, Spain
[2] Computer Vision Lab at University of California, Santa Cruz, USA
{arituerto,acm,josechu.guerrero}@unizar.es,
manduchi@soe.ucsc.edu

Abstract. Intelligent autonomous systems need detailed models of their environment to achieve sophisticated tasks. Vision sensors provide rich information and are broadly used to obtain these models, particularly, indoor scene understanding has been widely studied. A common initial step to solve this problem is the estimation of the 3D layout of the scene. This work addresses the problem of scene layout propagation along a video sequence. We use a Particle Filter framework to propagate the scene layout obtained using a state-of-the-art technique on the initial frame and propose how to generate, evaluate and sample new layout hypotheses on each frame. Our intuition is that we can obtain better layout estimation at each frame through propagation than running separately at each image. The experimental validation shows promising results for the presented approach.

1 Introduction

This paper investigates the construction of indoor scene models given an image sequence. The models contain essential information about the environment structure that may allow us to better understand the image. Prior approaches demonstrate the fact that obtaining information about the 3D structure of the scene is a powerful tool to improve the accuracy of other tasks [11].

Previous approaches on layout estimation use to assume certain constrains, like the Manhattan World assumption, and try to solve this problem for single images [3,8,9,13]. Other papers use sequential information to model the environment from a mobile camera. These approaches use to rely on SLAM or Structure-from-Motion [4,5,19].

The goal of this work is to provide semantic information about the scene layout traversed during the sequence (Fig. 1). Our approach propagates the estimated scene by taking advantage of restrictions in the sequential data and

This work was supported by the Spanish FPI grant BES-2010-030299 and Spanish projects DPI2012-31781, DGA-T04-FSE and TAMA.

A. Campilho and M. Kamel (Eds.): ICIAR 2014, Part II, LNCS 8815, pp. 374–382, 2014.
DOI: 10.1007/978-3-319-11755-3_42

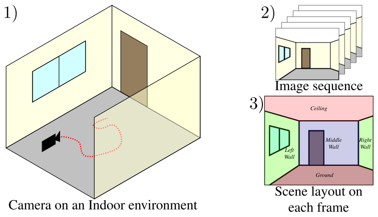


Fig. 1. 3D layout estimation along a sequence. 1) We have a mobile camera recording indoors, and we want to process the acquired sequence 2) to estimate the scene layout describing the 3D information of the environment at each frame 3).

prior knowledge of the projection of man made environments in images. We show how to achieve this task without the need to compute accurate camera motion or 3D maps.

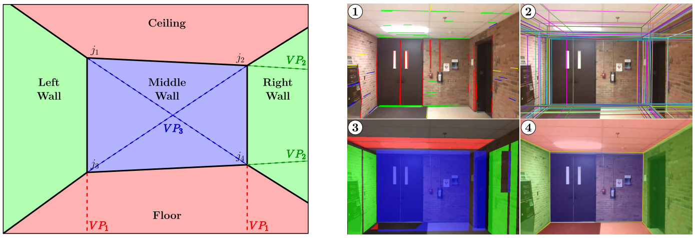
Related Work. Recognizing the 3D structure of an environment captured in an image is a widely studied problem. To solve the scene structure in general scenes, study in [10] proposes to learn appearance-based models of the scene parts and describe the scene geometry using these labels. Similarly, Markov Random Fields are used to infer plane parameters for homogeneous patches extracted from the image [18].

For indoor environments, additional assumptions can be made, such as the Manhattan World assumption [2]. Using this constrain, a Bayesian network model is proposed in [3] to find the "floor-wall" boundary in the images. The work in [13] generate interpretations of a scene from a set of line segments extracted from an indoor image. Similarly, the approach in [7] models the scene as a parametric 3D box. The spatial layout in omnidirectional images was solved in [14]. Extending similar ideas to outdoors, the work in [6] proposes to create physical representations where objects have volume and mass.

If we consider images of a video sequence we can propagate the scene information and get better and robust results. This is the idea exploited in this work. Acquiring sequential information is the usual scenario in mobile platforms, and the spatio-temporal restrictions between frames can provide both efficiency and accuracy advantages.

Most of the papers using sequential data to obtain scene models, are based on SLAM or structure-from-motion techniques. In. [4] geometric and photometric cues are combined to obtain the scene model from a moving camera. Similarly, structure-from-motion is used in [5]. The work in [19] describes a method to model the environment using images from a mobile robot.

Attending how to propagate semantic information in video sequences using probabilistic frameworks. We find the work in [1], that uses pixel-wise correspondences, image patch similarities and semantical consistent hierarchical regions in a probabilistic graphical model. The approach in [16] focus on label propagation indoors for mobile robots. Similarly, the work in [15] estimates the 3D structure of outdoor scenes by computing appearance, location and motion features.



(a)

(b)

Fig. 2. Scene model used to create the scene structure hypotheses, (a), and steps of the base method [13], (b): **1**, lines and vanishing points are detected, many structure hypotheses are proposed, **2**, and the orientation map is computed, **3**. Hypotheses are compared against the orientation map and the best is chosen as solution, **4**.

Our work proposes a probabilistic framework to propagate information in sequences. We aim to propagate the 3D layout of the environment traversed by the camera. The initial frame layout is obtained using a single image technique [13], and we then propagate this information in each consecutive frame.

2 Single Image 3D Layout

Our method uses the single image algorithm proposed by Lee et al. [13] to create layout hypotheses on the first frame. Their approach proposes several physically valid scene structures and validate them against an orientation map to find the best fitting model (Fig. 2(b)). Authors adopt the Indoor World model that combines the Manhattan World assumption [2] and a single-floor-single-ceiling model. Layout hypotheses are generated randomly from the lines detected in the image and then compared with an orientation map. The orientation map expresses the local belief of region orientations computed from the line segments.

To extract the image lines Canny edge detector (Kovesi [12] Matlab toolbox) is run and the vanishing points are detected following the method presented by Rother [17]. The generation of hypotheses is made based on the model showed in Fig. 2(a).

3 Propagating the 3D Layout in a Video Sequence

The objective of this work is to compute the 3D layout at every frame of a video sequence. We exploit the fact that consecutive frames in a video have certain spatio-temporal restrictions that constrain the possible variations in the scene acquired. By propagating the possible layouts, we improve the results and obtain a more robust estimation of the layout on each frame. We adopt a particle filter based strategy to track the posterior probability of the layout given all the observations up to the current frame.

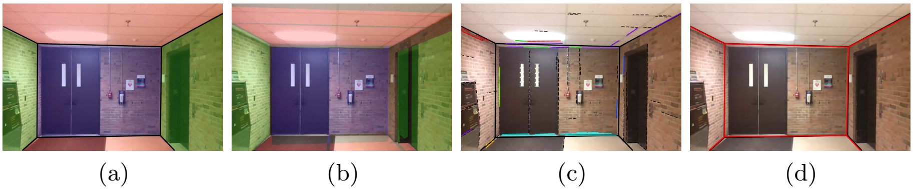


Fig. 3. Layout and correspondent plane orientation, (a), and the orientation map computed from the observed lines, (b). Both orientations maps are compared to compute S_{omap}. Lines supporting the model, S_{lines}, (c), and evaluated hypothesis (black) and closest observed layout (red), used to compute S_{model}, (d). Best seen in color.

The process followed by our approach is the following: For the first frame, hypotheses are created using the base algorithm (Section 2). These hypotheses are evaluated and ranked and the best one is selected as the solution for that frame. For next frames, new hypotheses (particles) are randomly generated depending on previous hypotheses and their evaluation score.

Layout Parametrization. We parametrized the scene layout model (Fig. 2(a)) by the image coordinates of the junctions defining the middle wall, j_k, and the directions of the scene vanishing points, VP_l: $x_i = \{j_1, j_2, j_3, j_4, VP_1, VP_2, VP_3\}$.

Hypotheses Evaluation. The evaluation of the hypotheses is performed on every frame. For all the images, lines and vanishing points are computed and used to evaluate the compatibility of the layout hypotheses. We define three evaluation measurements computed for each layout hypothesis x_i:

Orientation Map. The orientation map [13] expresses the local belief of region orientations computed from line segments (Fig. 3(b)). This observed orientation map, $omap(\mathsf{I}_i)$, is compared with the orientation map defined by the hypothesis being evaluated, $omap(x_i)$ (Fig. 3(a)). The evaluation score is computed as the number of pixels where the orientation of both maps is the same divided by the total number of image pixels, $nPix = width \times height$

$$S_{omap\ i} = \frac{\sum_{k=0}^{nPix} omap(\mathsf{I}_i)_k = omap(x_i)_k}{nPix} \qquad (k = 1 \ldots nPix) \tag{1}$$

Lines Supporting the Model. This evaluation measures how many of the observed lines support the hypothesis being evaluated (Fig. 3(c)). To evaluate how a hypothesis fits the new observation, we look for lines parallel and close to the model lines. The score is computed as the number of lines supporting the model divided by the total number of lines detected.

$$S_{lines\ i} = \left(\frac{\#\ supporting\ lines}{\#\ total\ lines} \right)_i \tag{2}$$

Table 1. Accuracy of the method for different evaluation methods (50 hypotheses)

	mean	*Max*
S_{omap} [13]	70.58	95.12
S_{lines}	75.47	93.78
S_{model}	59.38	93.78
$mean(S_{omap}, S_{lines})$	78.70	93.78
$mean(S_{omap}, S_{model})$	84.05	95.57
$mean(S_{lines}, S_{model})$	75.14	93.78
S_{total}	86.86	95.93

Distance to the Closest Layout Obtained with New Observed Lines. This last evaluation scores a propagated layout hypothesis, x_i, by computing the closest lines to this layout in the current image and determining a layout based on these lines, x_{obs} (Fig. 3(d)). The distance between layouts, $d(x_i, x_{obs})$, is computed as the mean distance between the junctions, j_k, of both layouts:

$$S_{model\ i} = \frac{1}{1 + d(x_i, x_{obs})} \qquad \text{where} \qquad d(x_i, x_{obs}) = \underset{k=1\ldots4}{mean}(||j_k, j_{k\ obs}||) \tag{3}$$

The mean of the three scores, $S_{total\ i}$, is used as evaluation.

Sampling New Hypotheses. A new set of hypotheses is created by sampling from the hypotheses of the previous frame. The probability of generating a new hypothesis, x'_i, from previous hypothesis x_i is $p_i = S_{total\ i}$.

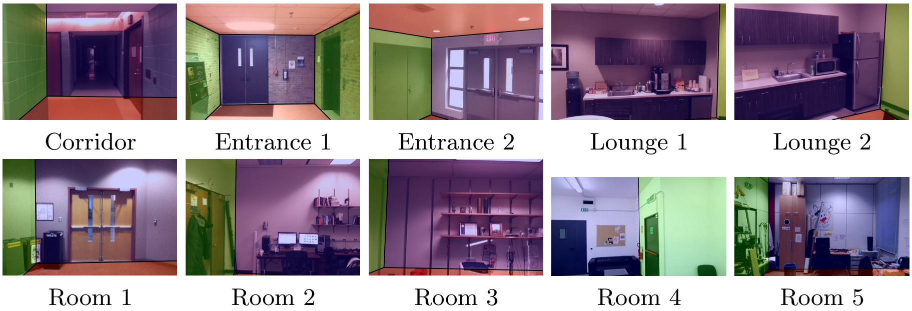
Given the camera motion, a homography relates the projection of the coplanar junctions between frames and the vanishing points are related by the rotation matrix. To create a new hypothesis from a previous one, we assume a random motion of the camera, with zero velocity and random noise in camera translation and rotation. From hypothesis x_i, sampled hypothesis x'_i will be related by the random rotation, R, and translation, $\mathbf{t}$. The junctions are related by a homography, H, and the vanishing points are related by the rotation matrix:

$$j'_k = \mathsf{H} \cdot j_k = (\mathsf{R} - \frac{\mathbf{t}\ \mathbf{n}^T}{d})j_k \text{ and } VP'_l = \mathsf{R} \cdot VP_l \qquad (k = 1\ldots4,\ l = 1\ldots3) \tag{4}$$

where $\mathbf{n}$ is the normal of the plane where the junction points lie and d the distance between the camera and the plane. We assume d distance as unitary so the scale of the random translation t is defined by the real distance to the plane.

4 Experiments

Experimental Settings. We have tested our method on the 10 sequences included in the dataset presented in [5]. These sequences have been acquired indoors with two different mobile cameras and include between 203 and 965 images. For all the sequences, the ground-truth has been manually annotated in one of each ten images. Fig. 4 shows example frames of the dataset sequences and the layout computed.



Corridor Entrance 1 Entrance 2 Lounge 1 Lounge 2

Room 1 Room 2 Room 3 Room 4 Room 5

Fig. 4. Examples of the resulting layout in some frames of all the dataset sequences. Best seen in color.

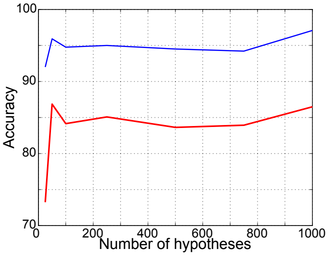


Fig. 5. Accuracy of our method for different number of hypotheses. Mean (red) and maximum accuracy (blue) of the layout solution along the sequence.

The accuracy of the solution is computed as the number of pixels where the orientation defined by the ground-truth and the orientation computed from the layout hypothesis is the same divided by the total number of pixels of the image.

Analysis of the Proposed Method Parameters. The accuracy of our method varies with two important choices: a) the evaluation measurements used and b) the number of particles used.

Influence of the Evaluation Measurement. Table 1 shows the mean and maximum value of the accuracy of the solution hypothesis on all the frames of the sequence Entrance 1. Results show that combining the different evaluation measurements we get to choose always a better solution. Therefore, all the evaluation measurements are used together in the remaining experiments.

Influence of the Number of Particles. Fig. 5 shows the accuracy of the algorithm presented depending on the number of particles. Results show poor accuracy when few hypotheses are considered (25 particles), and how the accuracy grows rapidly with the number of particles. For more than 50 particles, augmenting the number of hypotheses do not represent a big change in the method accuracy.

Table 2. Mean and standard deviation (σ) of the accuracy of our algorithm and the base algorithm for all the dataset sequences (50 hypotheses)

	Our Algorithm		Lee et al. Algorithm [13]	
	mean	σ	*mean*	σ
Corridor	42.93	**11.99**	**56.84**	30.78
Entrance 1	**86.80**	**9.90**	80.13	11.47
Entrance 2	71.34	**15.49**	**74.27**	15.76
Lounge 1	**56.52**	**17.47**	47.40	30.68
Lounge 2	31.78	31.17	**36.38**	**28.20**
Room 1	**60.69**	**14.52**	50.73	25.97
Room 2	**75.91**	**10.20**	66.79	25.11
Room 3	**63.42**	**16.83**	36.82	35.82
Room 4	20.64	**13.41**	**25.93**	24.24
Room 5	**69.27**	**16.07**	64.70	22.63
Average	**57.93**	**15.71**	54.00	25.07

Method Evaluation. Table 2 shows results of our method run on all the dataset and compared with the base method [13]. The base method is intended to work on single images so we run this algorithm over all the frames of the sequence independently. For each sequence, the mean and the deviation of the accuracy obtained for the solution hypothesis in all frames are shown. Our method performs better for the majority of sequences and the average accuracy value is higher. At the same time, our solutions are more stable across all frames, since the standard deviation is smaller.

Fig. 4 shows examples of the layout solution obtained for some frames of the dataset. The results for Entrances 1 and 2 are good (mean accuracies of 83.63 and 72.70, respectively). In sequences Lounge 1 and 2, Room 2, 3 and 5 the layout fits the environment, but the method fails in adjusting hypotheses lines to the structure. Finally, our method shows lower performance for sequences Corridor, and Room 4 where more than 3 walls appear, and Room 1 that violates the Manhattan World assumption.

5 Conclusions

In this paper we presented an approach to obtain the 3D spatial layout of all the frames of a video sequence. The method is designed to work indoors, makes use of the Manhattan World assumption and it is based in the previous work from Lee et al. [13]. This technique, is integrated with a Particle Filter to take advantage of the sequential information of video sequences. We have presented different evaluation methods to measure how well a spatial layout fits to an image. Experiments showed how our approach presents better accuracy than the state-of-the-art base algorithm.

References

1. Badrinarayanan, V., Galasso, F., Cipolla, R.: Label propagation in video sequences. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3265–3272 (2010)
2. Coughlan, J.M., Yuille, A.L.: Manhattan world: Compass direction from a single image by bayesian inference. In: IEEE International Conference on Computer Vision (ICCV), pp. 941–947 (1999)
3. Delage, E., Lee, H., Ng, A.Y.: A dynamic bayesian network model for autonomous 3d reconstruction from a single indoor image. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2418–2428 (2006)
4. Flint, A., Murray, D., Reid, I.: Manhattan scene understanding using monocular, stereo, and 3d features. In: IEEE International Conference on Computer Vision (ICCV), pp. 2228–2235 (2011)
5. Furlan, A., Miller, S., Sorrenti, D.G., Fei-Fei, L., Savarese, S.: Free your camera: 3d indoor scene understanding from arbitrary camera motion. In: British Machine Vision Conference (BMVC) (2013)
6. Gupta, A., Efros, A.A., Hebert, M.: Blocks world revisited: image understanding using qualitative geometry and mechanics. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part IV. LNCS, vol. 6314, pp. 482–496. Springer, Heidelberg (2010)
7. Hedau, V., Hoiem, D., Forsyth, D.: Recovering the spatial layout of cluttered rooms. In: IEEE International Conference on Computer Vision (ICCV), pp. 1849–1856 (2009)
8. Hedau, V., Hoiem, D., Forsyth, D.: Thinking inside the box: using appearance models and context based on room geometry. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part VI. LNCS, vol. 6316, pp. 224–237. Springer, Heidelberg (2010)
9. Hoiem, D., Efros, A.A., Hebert, M.: Geometric context from a single image. In: IEEE International Conference onComputer Vision (ICCV), pp. 654–661 (2005)
10. Hoiem, D., Efros, A.A., Hebert, M.: Recovering surface layout from an image. International Journal of Computer Vision **75**(1), 151–172 (2007)
11. Hoiem, D., Efros, A.A., Hebert, M.: Putting objects in perspective. International Journal of Computer Vision **80**(1), 3–15 (2008)
12. Kovesi, P.D.: MATLAB and Octave functions for computer vision and image processing
13. Lee, D.C., Hebert, M., Kanade, T.: Geometric reasoning for single image structure recovery. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2136–2143 (2009)
14. López-Nicolás, G., Omedes, J., Guerrero, J.: Spatial layout recovery from a single omnidirectional image and its matching-free sequential propagation. Robotics and Autonomous Systems (2014)
15. Raza, S.H., Grundmann, M., Essa, I.: Geometric context from video. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2013)
16. Rituerto, J., Murillo, A., Kosecka, J.: Label propagation in videos indoors with an incremental non-parametric model update. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 2383–2389 (2011)
17. Rother, C.: A new approach to vanishing point detection in architectural environments. Image and Vision Computing **20**(9), 647–655 (2002)

18. Saxena, A., Sun, M., Ng, A.Y.: Make3d: Learning 3d scene structure from a single still image. IEEE Transactions on Pattern Analysis and Machine Intelligence **31**(5), 824–840 (2009)
19. Tsai, G., Kuipers, B.: Dynamic visual understanding of the local environment for an indoor navigating robot. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 4695–4701 (2012)