# 3D Layout Propagation to Improve Object Recognition in Egocentric Videos

Alejandro Rituerto(✉), Ana C. Murillo, and José J. Guerrero

Instituto de Investigación en Ingeniería de Aragón, University of Zaragoza, Zaragoza, Spain
{arituerto,acm,josechu.guerrero}@unizar.es

**Abstract.** Intelligent systems need complex and detailed models of their environment to achieve more sophisticated tasks, such as assistance to the user. Vision sensors provide rich information and are broadly used to obtain these models, for example, indoor scene modeling from monocular images has been widely studied. A common initial step in those settings is the estimation of the 3D layout of the scene. While most of the previous approaches obtain the scene layout from a single image, this work presents a novel approach to estimate the initial layout and addresses the problem of how to propagate it on a video. We propose to use a particle filter framework for this propagation process and describe how to generate and sample new layout hypotheses for the scene on each of the following frames. We present different ways to evaluate and rank these hypotheses. The experimental validation is run on two recent and publicly available datasets and shows promising results on the estimation of a basic 3D layout. Our experiments demonstrate how this layout information can be used to improve detection tasks useful for a human user, in particular sign detection, by easily rejecting false positives.

**Keywords:** Scene understanding · Egocentric vision · Object detection

## 1 Introduction

Vision systems have become an essential perception component in all kinds of autonomous and intelligent systems, including assistance oriented systems such as household robots or wearable visual assistance devices [5,29]. There is a growing interest on applications using wearable cameras for vision assistive approaches, frequently towards assistance for impaired people [4,25].

These applications are based on visual recognition systems, and it has been shown many times that context information is essential to achieve better recognition performance in real world problems [26]. Even a basic 3D model of the

We would like to thank Prof. Roberto Manduchi for his comments and suggestions, which helped us to improve the present work. This work was supported by the Spanish FPI grant BES-2010-030299 and Spanish projects DPI2012-31781, DGA-T04-FSE and TAMA.


L. Agapito et al. (Eds.): ECCV 2014 Workshops, Part III, LNCS 8927, pp. 839–852, 2015.
DOI: 10.1007/978-3-319-16199-0_58

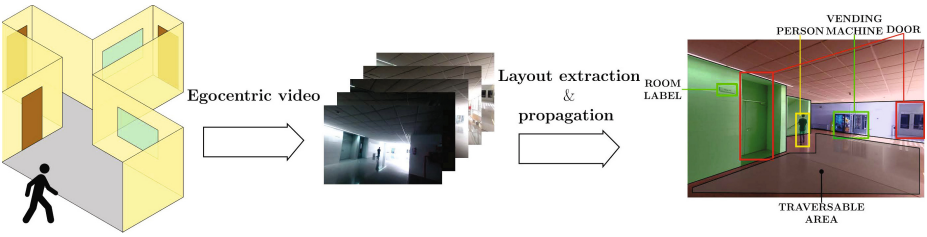



**Fig. 1.** Our goal is to process video acquired from a wearable camera to obtain the 3D layout of the scene in each frame (Red: floor and ceiling; Green and Blue: walls from different orientations). This layout information is a strong prior to facilitate the detection of other scene details: persons, signs, doors or the traversable area.

scene provides useful information about the environment structure that facilitates automatic scene understanding. For example, it can help us identify the type of area traversed (e.g., a corridor or a room) or provide strong priors for detection and recognition of objects [14].

This work extends the work presented in [20]. Our goal is to provide a basic 3D model of the environment traversed while recording a video (Fig. 1) to enhance the performance of more complex tasks. We aim to achieve this goal without computing accurate camera motion or 3D maps of the environment.

## 2 Related Work

The estimation of the 3D layout of a scene from an image is a widely studied problem, as well as the advantages on using this layout information to facilitate further tasks. Prior work demonstrates the advantages of using scene layout information to improve recognition tasks. A simple 3D model of the scene or 3D spatial relationships between elements of the environment allows to better understand the content of the scene and provide strong priors for detection and recognition of objects [14,24]. Recent approaches [2,12,16] propose to solve simultaneously the problems of estimating the layout of the scene and detecting the objects that appear in it.

Earlier approaches to estimate the layout for general scenes include the work by Hoiem et al. [13], that proposed to learn appearance-based models of the scene parts (sky, floor, vertical objects) and described the scene geometry using these coarse labels. Later, Saxena et al. [23] used Markov Random Fields to infer plane parameters, such as 3D location and orientation, for homogeneous patches extracted from the image. For indoor environments, where certain additional assumptions can be made, we find the work proposed by Delage et al. [7], where a dynamic Bayesian network model is used to find the "floor-wall" boundary in the images assuming a Manhattan World [6]. Lee et al. [17] presented a method to generate interpretations of a scene from a set of line segments extracted from an indoor image. Similarly, Hedau et al. [11] proposed how to model the scene

as a parametric 3D box. Gupta et al. [10] extended this idea to outdoor scenes and proposed how to create physical representations of outdoor scenes where objects have volume and mass, and their relationships describe the 3D structure and mechanical configurations.

Papers described so far analyze the structure of a single image. However, if we consider images that belong to a video sequence, we could propagate the information already obtained about the scene and obtain a better, more efficient or more robust result. Spatio-temporal restrictions between consecutive frames can provide both efficiency and accuracy improvements by accumulating the information obtained in each of them. This is one of the key ideas exploited in this work.

Most of the recent approaches taking advantage of sequential information, are based on SLAM or structure-from-motion techniques. For example, Flint et al. [8] combined geometric and photometric cues to obtain their scene model from a moving camera. They applied ideas from semantic reasoning in monocular images and 3D information obtained using structure-from-motion techniques. Similarly, Furlan et al. [9] proposed a method to estimate the 3D indoor scene layout from a moving camera. They pre-process the sequence to obtain the camera motion and a 3D map of the environment. From these results the method creates scene hypotheses that are evaluated and improved along the sequence. Tsai et al.[27] described a method to create a model of the environment using images acquired from a mobile robot. Since they focus on a robot moving indoors they can adopt constraints about the camera motion and the environment. The method uses different hypotheses describing the environment, that are updated with new details discovered while the robot moves.

Also related to our approach, we find papers on how to propagate semantic information in video sequences using different probabilistic frameworks. Badrinarayanan et al. [1] used a probabilistic graphical model. They are able to use pixel-wise correspondences from motion estimation, image patch similarities or semantical consistent hierarchical regions to propagate the labels. Vazquez et al. [28] presented the Multiple Hypothesis Video Segmentation method for unsupervised segmentation of video sequences. The method works with a few frames at a time and creates, propagates and terminates the labels without supervision. Rituerto et al.[21] focused on label propagation indoors using images acquired from a mobile robot. They learn the appearance of the different regions of interest from some training examples and propagate them trough the sequence using a non-parametric model. Similarly, Hussain et al. [19] estimate the 3D structure of outdoor video scenes by computing different appearance, location and motion features.

Our work also proposes a probabilistic framework to propagate semantic information in a sequence, in particular, we aim to propagate the 3D layout of the environment traversed by a camera. We use a hierarchical method for single image layout estimation adapted from [18], and we then propagate and update this information making use of spatio-temporal restrictions and the lines detected in each frame.

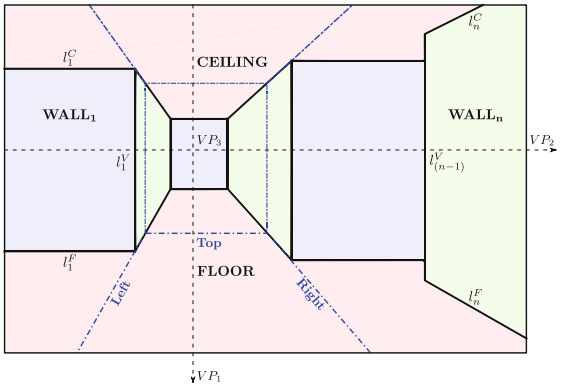


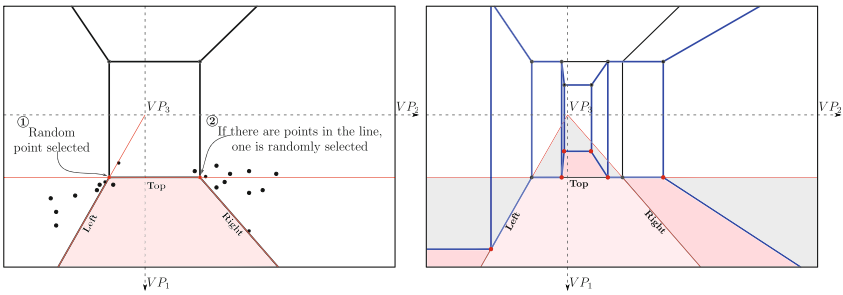

**Fig. 2.** Scene layout model. The *colored areas* in the image encode the different planes of the scene according to the surface orientation (red for horizontal, green and blue for vertical). The *black lines* define the scene structure and are grouped as floor ($l_i^F$), vertical ($l_i^V$) and ceiling lines ($l_i^C$). A scene with $n$ planes contains $n$ ceiling and floor lines and $n-1$ vertical lines. *Blue dashed lines* denote the basic scene layout that originated the complete layout. *Black dashed lines* are the horizon and vertical lines defined by the vanishing points of the scene. (Best seen in color).

## 3 Initial 3D Scene Layout

This section presents our approach for single-view layout estimation. The proposed method provides a set of scene layout hypotheses that will be automatically evaluated, ranked and propagated accordingly. We adapt the hierarchical method proposed in [18] to compute the scene layout from an omnidirectional image. They proposed to start the process by looking for a basic scene layout: a rectangular space around the camera. Once this basic layout has been detected, they expand it by looking for plausible walls and corners. The method uses floor points as base to build and expand the hypothesis. Fig. 2 shows the scene model that we adopt inspired by those ideas. Since we are using conventional cameras, with smaller field of view than omnidirectional cameras, the basic scene layout is formed by just three walls, Left, Top and Right. As the original work, we expand the basic layout in a hierarchical process.

*Lines, vanishing points and intersections.* In the first step of the method line segments and vanishing points of the image are computed. To extract the image lines Canny edge detector (Kovesi [15] Matlab toolbox) is run and the vanishing points are detected following the method presented by Rother [22]. Then, intersections between the detected line segments are computed.

*Building a basic room layout.* Room hypotheses are randomly generated from the intersections computed. The process is shown in Fig. 3. To build a basic room, a floor intersection is randomly chosen. The vanishing lines crossing in that point are computed. To finish the hypothesis, another floor intersection




Basic layout hypothesis Expanded layout hypothesis

**Fig. 3.** Basic and expanded scene layout hypotheses. In both figures, *red lines* represent the vanishing lines used to build a hypothesis and *red points* are the points used to define one hypothesis. To build a basic layout hypothesis (*black lines*), a floor intersection point is chosen randomly and vanishing lines in the directions intersecting on that point are computed. To complete the hypothesis another point is randomly selected among those where vanishing lines intersect. To expand a basic layout hypothesis (*blue lines*), we follow each of the basic room boundaries (left, top and right) and look for intersections that enlarge the room area. *Gray areas* show the expansion area where these intersections occur. (Best seen in color).

is chosen between those aligned with the computed vanishing lines. The lines crossing in those points compound the basic room hypothesis.

*Expanding a basic room layout.* Given a basic room hypothesis, it can be expanded to fit more complex environments. Fig. 3 shows this process. For each boundary of the basic room model, we look for intersections that could enlarge the floor area. The gray areas show where these intersections appear. The expansion process depends on the kind of boundary that we try to expand:

- *Top boundary:* we start with a random floor intersection in the Top boundary. From this point, a vanishing line is computed and another point aligned with this new line is chosen. This process is repeated until we close the area.
- *Left or Right boundaries:* in the case of Left or Right boundaries we define two ways of expanding the floor. We can choose a point aligned with the Top boundary as show for the Right boundary in Fig. 3, or choose a point in the correspondent boundary as done for the Left boundary in the same figure.

*Ceiling detection.* We adopt the Indoor World model that combines the Manhattan World assumption [6] and a single-floor-single-ceiling model. This model applies to most indoor environments and introduces symmetry between ceiling and floor shapes, something useful when the floor-walls boundaries are occluded. Once the floor boundaries have been defined, we look for the ceiling boundaries. We assume floor-ceiling symmetry, so we just have to detect the height of the room. We compute the first vertical line of our model, and look for ceiling intersections aligned with that line: one is randomly chosen. When we have computed the height of one vertical line, the rest of the ceiling boundaries can be computed

drawing parallel lines to the floor boundaries and computing the intersections with the rest of vertical lines.

## 4 Propagating the 3D Layout in a Video Sequence

Once we are able to compute the scene layout from a single image, our goal is to propagate this 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 variations in the acquired images. As seen before, the proposed method for single-view scene layout estimation is based on line detection. Image lines are very informative, but their detection is noisy. By propagating the possible layouts computed in one frame to the next frames, we are hoping to improve the results and obtain a more robust estimation for each frame in the sequence. We adopt a particle filter based strategy to track the posterior probability of the layout given all the observations up to the current frame.

Algorithm 1 presents the main steps of our approach. $\mathsf{I}_t$ is the frame at time $t$ and $X_t$ is the layout state, compound by $n$ layout hypotheses, $\mathbf{X}_t = \{x_1, x_2, \ldots x_n\}$. For the first frame, hypotheses are created using the single-view algorithm (Section 3). These hypotheses are evaluated and ranked as detailed in the following subsections, 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. Again, these hypotheses are evaluated in a similar manner and the best one is selected as the solution in each of the following frame

**Algorithm 1** Particle filter based algorithm for hypothesis sampling

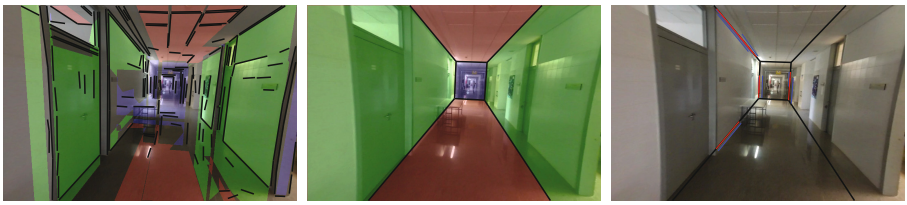
**Require:** Video sequence: $\mathsf{I}_t | t = 0 \ldots$ # frames
**Ensure:** 3D Scene Structure Layout: $bestHyp_t$
  $\mathbf{X}_0$ = generateHypothesisFromImage($\mathsf{I}_0$)
  $\mathbf{p}_0$ = evalHypotheses($\mathbf{X}_0$, $\mathsf{I}_0$)
  **for** $t = 1 \ldots$ # Frames **do**
    $\mathbf{X}_t$ = sampleNewHypotheses($\mathbf{X}_{t-1}$, $\mathbf{p}_{t-1}$)
    $\mathbf{p}_t$ = evalHypotheses($\mathbf{X}_t$, $\mathsf{I}_t$)
  **end for**

### 4.1 Layout Parametrization

The model used to define the 3D scene layouts is shown in Figure 2. A room hypothesis $x_i$ compound of $n$ walls is parametrized as sets of floor, vertical and ceiling lines, and the vanishing points of the scene:

$$x_i = \{(l_1^F \ldots l_n^F),\ (l_1^V \ldots l_{(n-1)}^V),\ (l_1^C \ldots l_n^C),\ (VP_1,\ VP_2,\ VP_3)\} \qquad (1)$$

where $l_i^F$ is the $i$-th floor line, $l_i^V$ the $i$-th vertical line and $l_i^C$ the $i$-th ceiling line, that is aligned with the same vanishing point than $l_i^F$ (they are parallel in the scene). The model also includes the vanishing points: $VP_1$, $VP_2$ and $VP_3$.



Observed orientation map Hypothesis orientation map Overlapping lines

**Fig. 4.** Evaluation of the hypotheses. The observed orientation map, computed from the detected lines, and the hypothesis orientation map are compared to compute $S_{omap}$. $S_{overlap}$ is computed as the length of overlapping (red) divided by the total length of the hypothesis lines (black). Blue lines are the detected lines that are parallel and close to the model lines. (Best seen in color).

### 4.2 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 two measurements for this evaluation computed for each layout hypothesis $x_i$:

- *Orientation map:* the orientation map is presented in [17]. It expresses the local belief of region orientations computed from detected line segments (Fig. 4). This orientation map, $omap(\mathrm{l}_i)$, is compared with the orientation map defined by the hypothesis being evaluated, $omap(x_i)$ (Fig. 4). 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, $nPix = width \times height$

$$S_{omap\ i} = \frac{\sum_{k=0}^{nPix} omap(\mathrm{l}_i)_k = omap(x_i)_k}{nPix} \tag{2}$$

  where $k$ is the pixel index. This score is the only evaluation used in [17] where the highest $S_{omap}$ gives the chosen solution.
- *Observed lines overlap:* this evaluation measures the length of the overlapping between the observed lines and the lines of the hypothesis being evaluated. The layout parametrization used defines model lines delimiting the layout areas (Fig. 4). We look for lines parallel and close to these model lines and compute their overlapping length with the model lines. The score of this evaluation is computed as the total overlapping length divided by the total length of the model lin:

$$S_{overlap\ i} = \left( \frac{\sum overlap\ length}{\sum model\ lines\ length} \right)_i \tag{3}$$

  Both scores are used together to evaluate the hypotheses:

$$S_{total\ i} = mean(S_{omap\ i},\ S_{overlap\ i}) \tag{4}$$

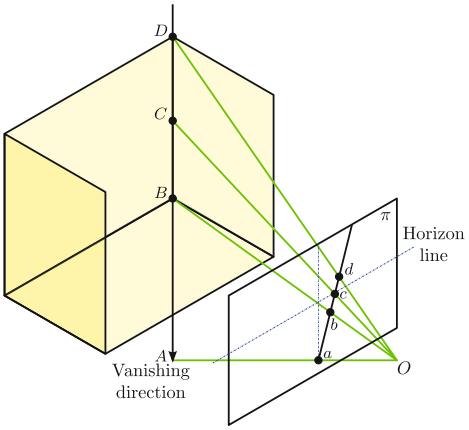



**Fig. 5.** The cross-ratio of the four points showed remains constant between consecutive views. This relation is used to locate the ceiling points when sampling new hypotheses. (Best seen in color).

### 4.3 Sampling New Hypotheses

A new set of hypotheses is created by sampling from the hypotheses of the previous frame and their evaluation score. For each hypothesis, a score has been computed, $S_{total\ i}$. The number of new hypothesis sampled from each previous hypothesis depends on this score. The probability of generating a new hypothesis, $x_i'$, from previous hypothesis $x_i$ is $p_i = S_{total\ i}$. New hypotheses are created randomly, high scores will generate more new hypotheses since they are more probable, and low scores hypotheses will receive few samples or even disappear.

The model used parametrizes the layout as sets of lines describing planes. Given the camera motion, a homography relates the projection of the coplanar points between frames and the vanishing points are related by the rotation matrix. We work with a moving camera where rotation and translation are unknown. 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. Random rotation and translation are created, $\mathsf{R}$ and $\mathbf{t}$, from 3 random angles ($\mathsf{R} = f(roll, pitch, yaw)$ and 3 random translations ($\mathbf{t} = [t_x, t_y, t_z]^T$). The homography $\mathsf{H}$ relating coplanar points can be computed as

$$\mathsf{H} = \mathsf{R} - \frac{\mathbf{t}\ \mathbf{n}^T}{d} \tag{5}$$

where $\mathbf{n}$ is the normal of the plane where the junction points lie and $d$ the distance between the camera and the plane. The plane used is the floor plane, where we have computed the room hypothesis. We assume $d$ distance as unitary so the scale of the random translation $t$ is defined by the real distance to the plane.

From hypothesis $x_i$, sampled hypothesis $x'_i$ will be related by the random $\mathsf{R}$ and $\mathbf{t}$. Points, $pt$, of the floor lines are related by a homography:

$$pt' = H \cdot pt = (\mathsf{R} - \frac{\mathbf{t}\ \mathbf{n}^T}{d})pt \quad (pt \in l_i^F | i = 1 \dots n) \tag{6}$$

and the vanishing points are related by the rotation matrix:

$$VP'_k = \mathsf{R} \cdot VP_k \quad (k = 1 \dots 3) \tag{7}$$

*Ceiling points relation.* Through the computed homography, we are able to relate the points on the floor, however we cannot relate the points in the ceiling of the scene, since they are part of a different plane. To relate the ceiling points, we assume that the distance between camera and floor remains the same between consecutive frames. We use the cross ratio to relate the height of the scene between images, Fig. 5. Given 4 collinear points, $A$, $B$, $C$ and $D$, the cross ratio, $C_R$ remains invariant for any perspective. The collinear points in our case are the two intersections defining the height of the room in the image, $b$ and $d$, the vertical vanishing point, $a$, and the intersection between the vertical line and the line of the horizon, $c$. Since we consider the camera height to be the same between consecutive frames, the horizon is the same and the cross ratio remains constant. So, the cross ratio, $C_R$, is computed in the current image as:

$$C_R = \frac{|ac|}{|ad|}\frac{|bd|}{|bc|} \tag{8}$$

where $|ac|$ is the signed distance between $a$ and $c$. Therefore, we obtain the ceiling point from the the floor point in next image as:

$$|b'd'| = C_R \frac{|a'd'||b'c'|}{|a'c'|} \tag{9}$$

## 5 Experimental Validation

We have run experiments in public datasets to show the performance of the proposed method and how its use can improve object recognition tasks.

### 5.1 Analysis of the Method Performance

We analyze the performance of the layout estimation obtained by our method by comparing it with a well know state-of-the-art method [17] as baseline[1].

[1] We have used the code provided by the authors in http://www.cs.cmu.edu/~dclee/code/index.html. This version does not include the complete environment model presented in the paper.


**Fig. 6.** Example images of all the sequences included in the dataset [9] with the best fitting layout obtained with our method

**Experimental Settings.** We have tested our method on the 10 sequences included in the dataset presented in [9]. These sequences have been acquired indoors with two different mobile cameras (Fig. 6) and include between 203 and 965 images. For all the sequences, the ground-truth has been manually annotated in one of each ten images. Figure 6 shows example frames of all the sequences included in the dataset and the correspondent resulting layout. Note that the Manhattan World assumption cannot be applied in some of the sequences, like Room 1 where walls are not orthogonal.

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 are the same divided by the total number of pixels of the image

$$Accuracy = 100 \frac{\sum_{k=0}^{nPix} omap(GT)_k = omap(x_i)_k}{nPix} \tag{10}$$

where $k$ is the pixel index, $GT$ denotes the ground-truth layout, $x_i$ is the layout hypothesis being analyzed and the number of pixels in the image is $nPix = width \times height$.

**Method Evaluation.** Table 1 shows the accuracy for all the sequences included in the dataset for our method and the base method [17]. 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 and both methods, the mean

**Table 1.** Mean accuracy of the layout solutions for each sequence obtained with Lee et al. method [17] and the proposed method

| | Lee et al. [17] | Proposed method |
|---|---|---|
| Corridor | 56.84 | 71.77 |
| Entrance 1 | 80.13 | 72.49 |
| Entrance 2 | 74.27 | 66.45 |
| Lounge 1 | 47.40 | 55.43 |
| Lounge 2 | 36.38 | 57.38 |
| Room 1 | 50.73 | 55.99 |
| Room 2 | 66.79 | 78.93 |
| Room 3 | 36.82 | 74.49 |
| Room 4 | 25.93 | 63.29 |
| Room 5 | 64.70 | 78.85 |
| Average | 53.99 | 67.50 |

of the accuracy obtained for the solution hypothesis in all frames is shown. Our method performs better for the majority of sequences. Main performance differences correspond to Lounges 1 and 2 and Rooms 3 and 4 sequences, where the algorithm in [17] performances are low while our method produces good results. In these sequences there is more clutter than in the rest. On average, our method performs better that the baseline algorithm. Fig. 6 shows the resulting layout on one image of each sequence.

### 5.2 Improving Object Recognition Tasks

This subsection shows results on object recognition tasks, poster detection in this case, using an egocentric vision dataset.

**Experimental Settings.** The images used in this second experiment are part of a wearable vision system dataset publicly available[2]. It consists of several indoor sequences acquired with wearable vision sensors. We have selected certain frames along those sequences that contain poster or signs, our objects of interest, to be able to demonstrate how the context information provided by the layout helps to automatically discard wrong detections.

We analyze how the performance of a sign detector can be improved by using the layout information as prior information. We consider a sign detection method that detects rectangular hypothesis in the scene that could correspond to signs. We compute the Precision and Recall of the correctly detected signs given the rectangles provided by this detector, i.e., among those hypothesis given by the detector, which ones are correct or wrong after the filtering achieved thanks to the layout information.

[2] https://i3a.unizar.es/es/content/wearable-computer-vision-systems-dataset

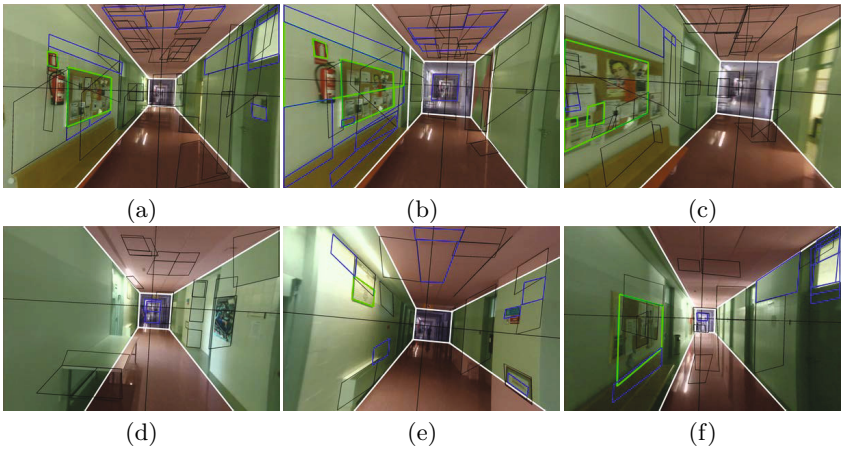



**Fig. 7.** We run the (rectangular) sign detector presented in [3]. We use the layout information to filter the rectangles and detect which ones are actual signs. Black rectangles show rectangles that have been correctly discarded with our filtering: they are not aligned with the scene vanishing directions or are part of more than one layout region. Blue rectangles are aligned with the layout and the vanishing directions, but they are not classified as posters by our filtering because of the relative location in the scene. Green and red show detections accepted by our filtering (correct or incorrectly respectively) and magenta rectangles are signs that have been incorrectly rejected. (Best seen in color).

**Poster Detection Evaluation.** Fig. 7 shows how the layout information improves the detection of posters in the images. We run the sign detector presented in [3] on a selection of images of the dataset. The detector creates detection hypotheses all over the image, but just some of them are correct. The rectangle hypothesis detected can be easily filtered using the scene layout and prior knowledge about man made environments to decide which hypothesis actually correspond to posters/signs or not:

- Scene objects are aligned with the scene vanishing points and with the vanishing points of the scene plane where they lay.
- Interesting objects, posters in our case, appear in walls, nor in floor or ceiling.
- Posters height is smaller than the wall height, they appear close to the eyes height (camera height in our case) and do not appear on top or bottom parts of the wall.

Sign detector detects 25 sign candidates per frame on average, and about 18 of these candidates are rejected (not aligned with the vanishing directions or are part of more than one layout region). The sign candidates remaining after the filtering are classified into poster or no poster. The precision of this classification is 95.24% and the recall is a bit lower 88.19%. The main reason for these high values is the filter step, where rectangles no fitting the structure are rejected. Fig. 7 shows examples of the poster detection.

## 6 Conclusions

This paper presents a new approach to obtain the 3D layout of a single image and propagate this layout along a video sequence. The approach is designed for indoor environments, so Manhattan World assumption is adopted. Our proposed method obtains an initial layout from a single image using certain assumptions typical for indoor environments and first-person perspective videos. Then, a particle filter framework is used to take advantage of the sequential information on video sequences and propagate the scene layout. The layout estimation method has shown better accuracy than a well known baseline and we show how to propagate the layout instead of computing all the model for each frame. Additionally, our experiments demonstrate how the 3D layout we obtain provides useful priors for recognition tasks. In particular we show how sign recognition can be improved by easily rejecting the numerous false positive detections.

## 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. Bao, S.Y., Sun, M., Savarese, S.: Toward coherent object detection and scene layout understanding. Image and Vision Computing **29**(9), 569–579 (2011)
3. Cambra, A.B., Murillo, A.: Towards robust and efficient text sign reading from a mobile phone. In: Int. Conf. on Computer Vision Workshops, pp. 64–71 (2011)
4. Chen, L., Guo, B.L., Sun, W.: Obstacle detection system for visually impaired people based on stereo vision. In: Int. Conf. on Genetic and Evolutionary Computing, pp. 723–726 (2010)
5. Ciocarlie, M., Hsiao, K., Jones, E.G., Chitta, S., Rusu, R.B., Şucan, I.A.: Towards reliable grasping and manipulation in household environments. In: Khatib, O., Kumar, V., Sukhatme, G. (eds.) Experimental Robotics. STAR, vol. 79, pp. 241–252. Springer, Heidelberg (2012)
6. 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)
7. Delage, E., Lee, H., Ng, A.Y.: A dynamic bayesian network model for autonomous 3d reconstruction from a single indoor image. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2418–2428 (2006)
8. 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)
9. 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)
10. 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)
11. 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)

12. 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)
13. Hoiem, D., Efros, A.A., Hebert, M.: Recovering surface layout from an image. International Journal of Computer Vision **75**(1), 151–172 (2007)
14. Hoiem, D., Efros, A.A., Hebert, M.: Putting objects in perspective. International Journal of Computer Vision **80**(1), 3–15 (2008)
15. Kovesi, P.D.: MATLAB and Octave functions for computer vision and image processing
16. Lee, D.C., Gupta, A., Hebert, M., Kanade, T.: Estimating spatial layout of rooms using volumetric reasoning about objects and surfaces. In: Advances in Neural Information Processing Systems (NIPS) (2010)
17. 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)
18. López-Nicolás, G., Omedes, J., Guerrero, J.: Spatial layout recovery from a single omnidirectional image and its matching-free sequential propagation. In: Robotics and Autonomous Systems (2014)
19. Raza, S.H., Grundmann, M., Essa, I.: Geometric context from video. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2013)
20. Rituerto, A., Manduchi, R., Murillo, A.C., Guerrero, J.J.: 3D Spatial layout propagation in a video sequence. In: Campilho, A., Kamel, M. (eds.) ICIAR 2014, Part II. LNCS, vol. 8815, pp. 374–382. Springer, Heidelberg (2014)
21. 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)
22. Rother, C.: A new approach to vanishing point detection in architectural environments. Image and Vision Computing **20**(9), 647–655 (2002)
23. 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)
24. Southey, T., Little, J.: 3D spatial relationships for improving object detection. In: 2013 IEEE International Conference on Robotics and Automation (ICRA), pp. 140–147 (May 2013)
25. Tapu, R., Mocanu, B., Bursuc, A., Zaharia, T.: A smartphone-based obstacle detection and classification system for assisting visually impaired people. In: Int. Conf. on Computer Vision Workshops (ICCVW), pp. 444–451 (2013)
26. Torralba, A., Murphy, K.P., Freeman, W.T.: Using the forest to see the trees: exploiting context for visual object detection and localization. Communications of the ACM **53**(3), 107–114 (2010)
27. 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)
28. Vazquez-Reina, A., Avidan, S., Pfister, H., Miller, E.: Multiple hypothesis video segmentation from superpixel flows. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part V. LNCS, vol. 6315, pp. 268–281. Springer, Heidelberg (2010)
29. Wexler, Y., Shashua, A., Tadmor, O., Ehrlich, I.: User wearable visual assistance device (ORCAM), uS Patent App. 13/914,792 (2013)