Behind the Scenes: What Moving Targets Reveal About Static Scene Geometry

Geoffrey Taylor Fei Mai
Canon Information Systems Research Australia
PO BOX 313, North Ryde, NSW 1670, Australia
{Geoffrey.Taylor, fei.mai}@cisra.canon.com.au

Abstract

Reasoning about 3D scene structure is an important component of visual scene understanding. Often, reasoning proceeds from low-level cues without resorting to full 3D reconstruction. However, existing geometric cues may require multiple viewpoints, supervised training, constraints on scene structure or information from auxiliary sensors. To address these limitations, this paper demonstrates how geometric context for a single static camera can be recovered from the location and shape of moving foreground targets. In particular, we propose methods to compute the likelihood of a static occlusion boundary and floor region at each pixel. Importantly, these cues do not require supervised training, or prior knowledge of camera geometry or scene structure. Finally, we show how the proposed geometric cues can be used to infer an ordinal depth map and demonstrate its use in compositing with correct occlusion handling.

1. Introduction

The connection between 3D structure and image understanding is an important and long running theme in computer vision, largely motivated by models of human perception from computational neuroscience, including Marr's 2.5D sketch [16]. Geometric cues, whether arising from direct 3D reconstruction or low-level geometric reasoning, have been utilized in a range of visual tasks including tracking [11], object detection [8] and visual saliency prediction [13]. Conventional geometric methods for recovering scene structure require multiple images with varying viewpoint, focus, lighting or other intrinsic or extrinsic parameters. However, practical applications such as surveillance and monitoring are dominated by static cameras in uncontrolled environments, typically with wide-baseline or non-overlapping views. In these instances, structure recovery must rely on less robust single-view geometric cues.

Since single-view scene reconstruction is inherently under-constrained, existing methods for recovering scene structure must incorporate prior knowledge. For example, the early blocks world experiments interpreted line drawings based on known 3D polyhedral shapes [18]. Other recent approaches adopt the Manhattan world assumption [4] as a constraint on urban and indoor scenes with orthogonal planes. Similarly, assuming a planar ground and pin-hole camera enables recovery of metric ground plane properties [1]. Several authors have adopted supervised machine learning to learn the mapping between low-level image features and 3D geometry [9, 21].

Figure 1: Scene geometry from moving targets: (left) video frame; (right) occluding (green), occluded (blue) and supporting (red) segments, and occlusion boundary (black).

Without the above assumptions of planar structures, known camera model or supervised training data, what can we infer about scene geometry from a single static camera? The answer draws on an alternative line of research that treats moving targets as exploratory "probes". As shown in Figure 1, regions in a static scene fall into three classes: those that *occlude* foreground targets, are *occluded by* foreground targets, and *support* foreground targets (i.e. floor regions that contain target "footprints"). These classes are not mutually exclusive, i.e. complex scenes may contain regions that are both occluders and occludees. *Static occlusion boundaries* between occluding and occluded regions correspond to depth discontinuities and provide a strong cue for 3D scene structure by inducing a depth ordering on neighbouring regions. Similarly, floor regions induce a depth ordering since they do not occlude other regions. Thus, segmenting an image into these classes of regions results in an ordinal depth map.

Based on the above observations, this paper introduces two methods for single-camera geometry estimation based on observations of moving targets. The first method generates a static occlusion boundary likelihood map which indicates the likelihood that each pixel occurs on a boundary between occluding and occluded regions. This

method assumes that static occlusions lead to a consistent change in the shape of partially occluded foreground targets, which otherwise move non-uniformly. The second method generates a floor likelihood map, which indicates the likelihood that each pixel occurs in a target support region. This method assumes that the floor has a unique appearance compared to non-floor regions. Importantly, these methods are applicable to many scenes of practical interest and impose less restrictive assumptions than the single-view scene geometry methods described above.

The remainder of this paper is organized as follows. Section 2 discusses related work on single-view structure recovery from moving foreground targets. Section 3 provides an overview of the proposed methods. Technical details for recovering the occlusion likelihood map and floor likelihood map are described in Sections 4 and 5 respectively. Section 6 presents a quantitative evaluation of the proposed cues and a comparison with competing methods. Finally, Section 7 presents an MRF-based model for recovering an ordinal depth map from the detected occlusion boundaries and floor regions.

2. Related Work

The recovery of single-view scene structure from the interactions of moving targets with static background has been previously explored by several authors. Most of these methods are based on analysing the location and shape of foreground regions detected using statistical background models such as described in [12].

Brostow and Essa [3] proposed a method to assign ordinal depth to a static image segmentation based on foreground occlusions, by 'pushing' occluded regions to lower layers and 'popping' occluding regions to higher layers. However, this approach is sensitive to the initial segmentation and target path. Schödl and Essa [22] later proposed a more robust minimum description length optimization for depth layer assignment based on the assumption that persistent edges of foreground targets coincide with occlusion boundaries. However, this approach tends to over-segment the scene, especially in the presence of stationary targets.

Based on similar ideas, Guan *et al.* [7] build a simple three-layer model that segments scenes into static background, moving targets and static foreground occlusions. Evidence for occlusions follows from two assumptions: occlusions result in persistent foreground edges perpendicular to the direction of motion, and occlusion edges never appear inside foreground regions. The simple three-layer model approach is restricted to relatively simple scenes, e.g. scenes in which a person walks behind but never in front of an occluding object. Jackson *et al.* [10] additionally assume foreground targets change rapidly in area during occlusion. Both approaches do not require static occlusion boundaries to coincide with intensity edges, and have the advantage of avoiding over-segmentation. However, neither method detects horizontal occlusion boundaries parallel to target motion. Our proposed methods avoid the limitations of [22] and [7] by directly estimating occlusion boundaries rather than ordinal depth segments. Ordinal depth may still be recovered by combining occlusion boundaries with floor segments (see Section 7).

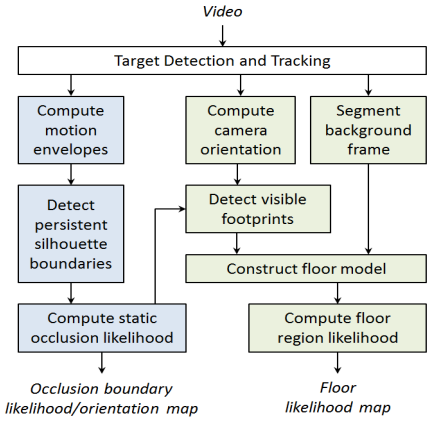


Figure 2: Proposed framework for scene geometry estimation.

Several authors have previously exploited moving targets to segment floor regions. Rother *et al.* [19] analyses moving people to recover camera calibration, floor segmentation and ground plane parameters in a static scene. The floor segmentation approach models floor appearance based on seed pixels underneath detected target footprints. The floor segment is constructed as a connected region of pixels within a threshold colour distance from the seeds. Bovyrin and Rodyushkin [2] proposed a similar model to detect floor regions by iteratively growing a floor region around seed pixels. Both approaches work well in simple scenes with homogeneous floor appearance and sufficiently distributed footprints. However, neither approach considers the impact of partially occluded targets with no visible footprint, or occlusions that divide the floor into disconnected segments. Our method employs occlusion reasoning and a floor appearance model based on superpixel statistics to overcome these limitations.

Finally, the recent work of Fouhey *et al.* [6] moves beyond simple foreground shape analysis by detecting human actions to probe "sittable" and "walkable" surfaces of a scene, which are subsequently used to infer single-view 3D structure. This approach nevertheless relies heavily on single-view human pose estimation, which remains an open and challenging research problem.

3. System Overview

Figure 2 outlines the proposed framework for single-view geometry from foreground targets. The algorithm takes video from a single static camera as input, and

(a) (b) (c)

Figure 3: Example of persistent silhouette boundary pixel detection: (a) first occluded target; (b) overlapping target behind the same occlusion; (c) superimposed silhouette boundaries.

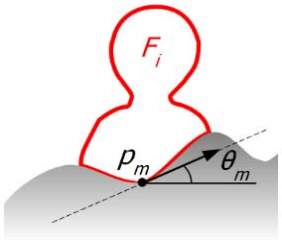


Figure 4: Persistent silhouette pixel orientation.

determines the likelihood of a static occlusion boundary and a floor region at each pixel. Associated with the occlusion likelihood is a per-pixel orientation map that encodes the relative depth of regions neighbouring a discontinuity. The method makes no assumptions about scene geometry or target motion except that sufficient targets are observed to reveal the major structures.

As shown in Figure 2, the first step is to detect and track foreground targets in each video frame. Our current implementation uses the SAMMI change detector [23], although any method that produces a per-pixel foreground mask is suitable[1]. Foreground regions are naïvely tracked assuming Brownian motion and greedy nearest neighbour association based on centroid location and size ratio. This processing step results in a set of foreground regions and their temporal associations over the video sequence.

The processing pipelines for recovering occlusion likelihood (blue blocks in Figure 2) and floor likelihood (green blocks in Figure 2) are largely independent except where the occlusion likelihood is used to reason about occluded targets during footprint detection. The following sections detail these pipelines in turn.

4. Static Occlusion Boundary Likelihood

Static occlusion boundary detection is based on the following three premises:

1. Static occlusion boundaries occur at edges of individual foreground targets, known as *silhouette boundary pixels*;
2. Static occlusion boundaries do not occur inside any detected foreground target that is behind the same occluder; and
3. Silhouette boundary pixels that lie on static occlusion boundaries have *persistent* location and orientation across different detections.

These concepts are illustrated in Figure 3. Figure 3(a) shows silhouette boundary pixels (magenta) of a partially occluded detection. Premise (1) states that a subset of the silhouette boundary pixels lies on the occlusion boundary, and premise (3) states that this subset has a persistent location and shape across multiple detected targets, since it is defined by the occlusion rather than the target. To identify this subset, Figure 3(b) shows an overlapping detection behind the same occluder. The superposition of both detections is shown in Figure 3(c). Pixel locations from the first target that are interior to the second target (shown in red) violate premise (2) and are discarded. The remaining boundary pixels shown in white are persistent across the two detections. Accumulating evidence for these persistent locations across all overlapping detected targets provides evidence for static occlusion boundaries.

4.1. Persistent Silhouette Boundary Detection

This section provides an algorithmic implementation of the concepts outlined above. Let $F = \{F_i\}$ represent the set of all detected foreground regions in all frames of a video sequence. Detection F_i is divided into boundary pixels $B_i = \{b_k\}$ (i.e. pixels that are 8-connected to a background pixel) and interior pixels $\bar{B}_i = \{\bar{b}_l\}$ such that $F_i = B_i \cup \bar{B}_i$. The set $\hat{F}_i$ of detections that overlap F_i is

$$\hat{F}_i = \{F_j : |F_i \cap F_j| > \rho \max(|F_i|, |F_j|), \forall j \neq i\}, \quad (1)$$

where ρ is a minimum overlap ratio threshold. Provided ρ is sufficiently high, $\hat{F}_i$ will generally include overlapping targets behind the same occluder as F_i. Then, the set of non-interior silhouette boundary pixels $P_i \subseteq B_i$ for detection F_i is

$$P_i = \{(p_m, \theta_m) : p_m \in B_i,\ p_m \notin \bar{B}_j\ \forall\ F_j \in \hat{F}_i\}, \quad (2)$$

where p_m is the pixel location and θ_m is the quantized boundary orientation, determined as a tangent orientation to the boundary of F_i (see Figure 4). The directed tangent is defined with detected target F_i on the counter-clockwise side. Boundary orientation thus encodes the relative depth of neighbouring regions, with the closer region (the *occluder*) on the clockwise side. Our current implementation quantizes θ_m into eight orientation bins.

The likelihood of persistence $L_i(p_m)$ at each pixel location $p_m \in P_i$ is measured as the proportion of overlapping targets that have a boundary at this location:

$$L_i(p_m) = \frac{1}{|\hat{F}_i|} \sum_{F_j \in \hat{F}_i} \sum_{b_k \in B_j} \delta(p_m, b_k), \quad (3)$$

where $\delta(\cdot,\cdot)$ is the Kronecker delta. Finally, the total likelihood of persistence $L(p, \theta)$ for each discrete image location p and quantized orientation θ is accumulated over

[1] SAMMI actually produces a foreground label per 8×8 DCT block, so occlusion likelihood maps are generated for 8×8 subsampled images.

all foreground detections as

$$L(p,\theta) = \sum_i \sum_{(p_m,\theta_m)\in P_i} L_i(p_m)\cdot\delta(p,p_m)\cdot \delta(\theta,\theta_m). \quad (4)$$

4.2. Motion Envelopes

Inferring occlusions directly from foreground regions faces two challenges. Firstly, stationary targets (sitting, standing, etc.) can generate high occlusion likelihood where no occlusion exists, since static silhouette boundaries cannot be distinguished from persistent occlusion boundaries. Secondly, equation (1) involves N^2 set intersections where N is the number of detected targets, which can quickly become intractable. Our solution is to track a smaller set of *motion envelopes*, and apply persistent silhouette boundary detection to motion envelopes rather than raw detections.

Let $F = \{F_t : t = 1,\dots,T\}$ represent detections of a target tracked over T frames. We wish to segment F into a set of N $(< T)$ non-overlapping sub-intervals $F_i' = \{F_t : a_i \le t \le b_i\}$, $i = 1,\dots,N$. The motion envelope M_i in each sub-interval is then the union of foreground regions

$$M_i = \bigcup_{t=a_i}^{b_i} F_t. \quad (5)$$

Importantly, the motion envelopes defined above preserve the property that static occlusion boundaries coincide with persistent silhouette boundary pixels; as shown in Figure 5, the shape of the occlusion boundary is preserved in the boundary of M_i. However, this property is potentially lost when a target moves from behind to in front of an occluder. We avoid this problem by choosing spatially compact sub-intervals. While many boundary criteria exist, we choose a_i and b_i to yield a small overlap between F_{a_i} and F_{b_i} (blue shaded area in Figure 5(a)):

$$|F_{a_i} \cap F_{b_i}| < \lambda \min(|F_{a_i}|, |F_{b_i}|) \quad (6)$$

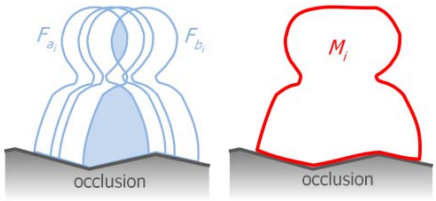
where λ is the maximum overlap ratio threshold. In general, the number of regions in F_i' varies with target motion, e.g. many detections may be combined into a single motion envelope while a target is stationary. Motion envelopes thus summarize target motion, reducing complexity and increasing robustness.

4.3. Likelihood Estimation

Finally, the static occlusion likelihood and orientation maps are computed from accumulated persistence $L(p,\theta)$:

$$L(p) = \max_\theta L(p,\theta)$$
$$\Theta(p) = \arg\max_\theta L(p,\theta). \quad (7)$$

The likelihood and orientation in equation (7) are only valid at pixel locations where more than one non-interior silhouette boundary was observed. This constraint is represented as a confidence mask:



(a) (b)

Figure 5: Motion envelope construction: (a) foreground detections; (b) resulting motion envelope (union of foregrounds).

$$C(p) = \sum_i \sum_{p_m\in P_i} \delta(p,p_m). \quad (8)$$

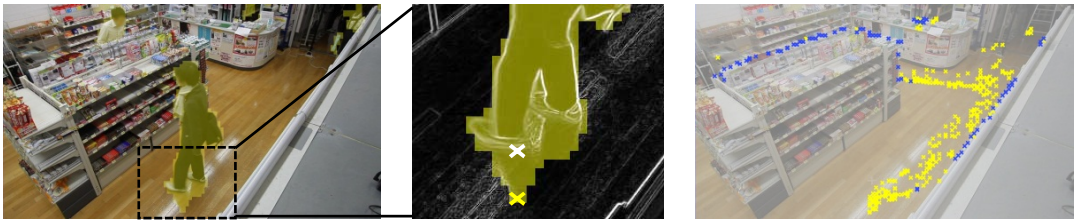
5. Floor Likelihood Estimation

Our proposed floor likelihood estimation method detects regions that support moving targets, i.e. regions likely to contain target footprints. This has two significant challenges: floor regions are often piecewise planar rather than globally planar, and targets typically walk on a small subset of the visible floor region. Existing footprint-based floor detection methods which make assumptions about geometry (e.g. planar ground) or connectivity (e.g. region growing) are therefore likely to find only part of the floor.

To overcome these limitations, we instead assume that the floor has a unique appearance compared to non-floor regions. This appearance is modelled using a non-parametric, multi-modal colour distribution over super-pixel regions containing footprints. These superpixels are the result of an over-segmentation of the background image into coherent regions with uniform colour or texture. An important step is to detect and filter out partially occluded targets to avoid corrupting the model with false footprints. The non-parametric, multi-modal colour distribution model is used to assign a floor likelihood value to each superpixel in the view, regardless of whether the superpixel contains footprints. This approach is especially suited to finding floors with coloured patterns (e.g. tiled floors), and disconnected regions due to occlusions, since it does not require targets to traverse all areas of the floor. Details of the algorithm are provided below.

5.1. Visible Footprint Detection

For simplicity, our current implementation assumes that the camera is in the upright orientation and the lowest point on the target is the footprint (generally, this assumption can be avoided by estimating the vertical direction as described in [15]). However, due to noise from shadows and reflections, edges of foreground regions may not coincide with silhouette edges of the target. Thus, a simple refinement step shown in Figure 6 is adopted to localize the true footprint. The refinement starts at the lowest point on a foreground mask closest to the central vertical axis of the target (yellow cross in Figure 6(b)) and



(a) (b) (c)

Figure 6: Footprint detection and filtering: (a) detected foreground (yellow); (b) initial footprint (yellow cross) and refined footprint (white cross) at intensity edge; (c) detection of occluded footprints (blue) and visible footprints (yellow) based on occlusion likelihood.

takes the strongest intensity edge above this point as the true footprint (white cross).

The assumption that the detected footprint touches the floor may be violated when the target is partially occluded, which occurs frequently in cluttered scenes. Detection and removal of partially occluded targets is therefore critical to ensure the estimated floor model remains free of non-floor pixels. Observing that partial occlusions typically generate false footprints on occlusion boundaries, occlusions can be detected by testing whether the occlusion boundary likelihood at putative footprint locations exceeds an empirical threshold. While it is conceivable that a partial occlusion creates a false footprint away from an occlusion boundary, this occurs rarely in practice. Figure 6(c) shows the result of footprint filtering over all detected targets in a video sequence, based on the observed occlusion likelihood map (see Section 4). False footprints along occlusion boundaries, most notably along the top of the shelves, are detected and rejected with high precision.

5.2. Floor Appearance Mixture Model

The underlying assumption in the floor appearance model is that the statistics of the entire floor are captured in a few representative superpixels, which are selected based on the presence of footprints. The superpixel segmentation is computed on a background frame free of foreground targets so that the pixel statistics capture only the static scene content. Any good superpixel algorithm is suitable for this purpose; Section 6 demonstrates experimental results for both watershed superpixels [14] and geodesic superpixels [17].

The presence of a footprint within a superpixel is a strong indicator that the superpixel represents a floor region. However, numerous sources of noise, including spurious foreground targets due to camera jitter, reflections, shadows, and incorrect filtering of occluded targets, lead to false footprint detections. Robustness to noise is strengthened by selecting only superpixels that contain N (empirically chosen) or more footprints. Figure 7 illustrates a geodesic superpixel segmentation, along with the representative floor superpixels (outlined in yellow) selected based on the footprints in Figure 6(c).

Each selected superpixel is assumed to capture a different aspect of the floor's appearance (e.g. different coloured tiles), and contributes a unique mode to a non-parametric multi-modal mixture model. Some modes may be identical, e.g. in large homogeneous floor regions. While further processing could be applied to cluster identical modes, this is not necessary in practice.

Figure 7: Selected superpixels (yellow) for floor model.

Supposing K superpixels are selected, the k-th mode, $k = 1, \ldots, K$, is modelled as a triplet of normalized colour histograms corresponding to the marginal distributions of pixel values in YCrCb space. The histogram for the c-th channel, $c = 1, \ldots, 3$, is $H_{c,k} = \{h_{c,k,b}\}$, where $\Sigma_b h_{c,k,b} = 1$, and $b = 1, \ldots, B$ ranges over the B histogram bins. The overall floor appearance is the set of histograms for all channels of every mode, $H = \{H_{c,k}\}$.

5.3. Likelihood Estimation

Intuitively, any superpixel sufficiently similar to at least one of the superpixels containing N or more footprints is also likely to be floor. For the j-th superpixel, the corresponding similarity metric is the minimum distance, $\hat{d}_j$, between the normalized colour histogram triplet $G_{c,j}$ of the j-th superpixel and all modes in the floor model:

$$\hat{d}_j = \min_k \sum_{c=1}^{3} d(G_{c,j}, H_{c,k}), \quad (9)$$

where $d(\cdot)$ is the earth mover's distance [20]. While $\hat{d}_j$ can be interpreted directly as a likelihood, for convenience we map it to the range [0,1] by the likelihood function L_j:

$$L_j = \exp\left(-\left(d_j/\sigma\right)^2\right), \quad (10)$$

where σ is an empirically chosen scaling factor encoding the expected noise and variation in floor appearance. The value L_j is proportional to the conditional probability of observing $\hat{d}_j$ given that the j-th superpixel is a floor region. Finally, the floor likelihood map is constructed by

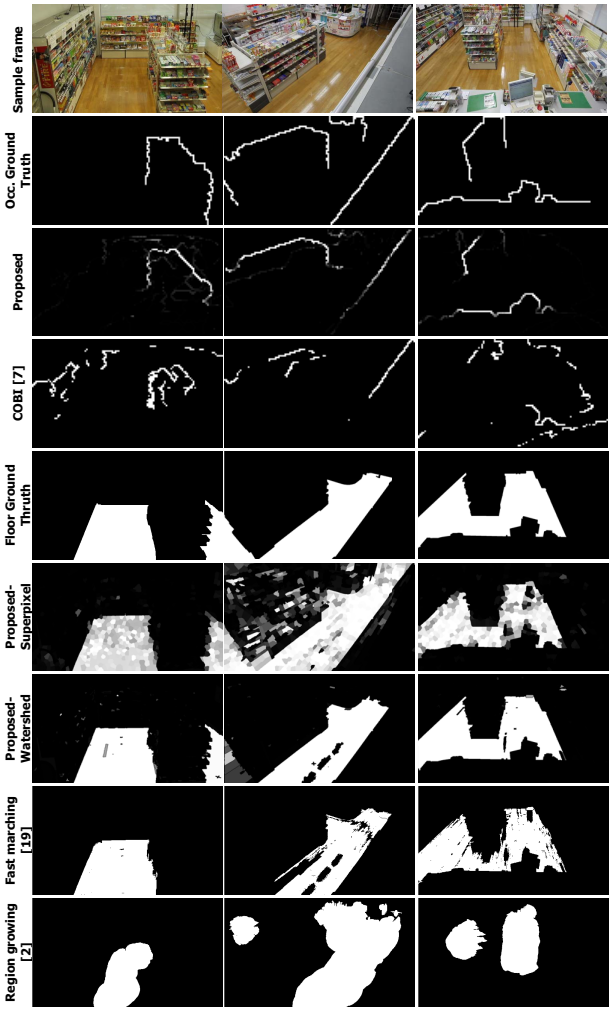


Figure 8: Experimental result for static occlusion boundary likelihood estimation and floor likelihood estimation (see text for discussion).

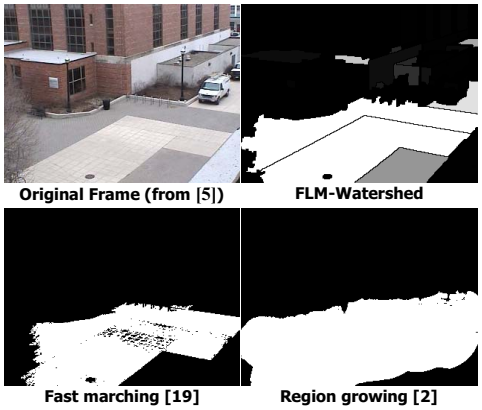


Figure 9: Comparison of floor detection with prior art for ground with multiple appearance modes.

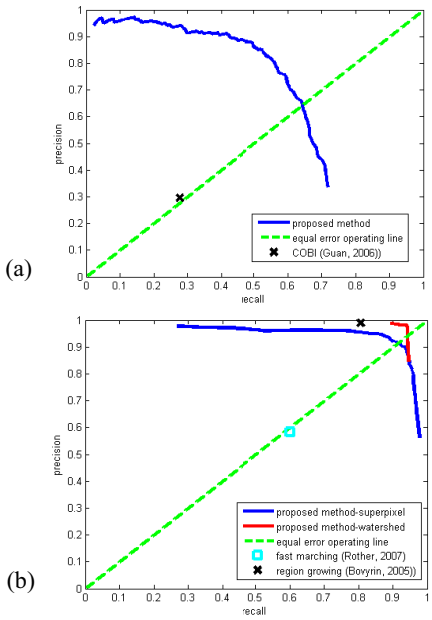


Figure 10: Precision-recall curve for proposed and competing methods for: (a) static occlusion boundary detection; (b) floor region detection.

assigning L_j to all pixels in the j-th superpixel.

6. Experimental Results

The proposed geometric cues were tested on an indoor dataset comprising three cameras recording actors performing typical customer behaviours in a convenience store. Approximately 10,000 frames and 10 targets were processed per camera, and the results are shown in Figure 8. All experiments used $\rho = 0.37$ in equation (1), $\lambda = 0.6$ in equation (6) and $\sigma = 0.9$ in equation (10) . Rows 2-4 provide results for occlusion likelihood detection. Row 2 shows the manual ground truth occlusion boundaries. Row 3 shows the static occlusion boundary likelihood maps generated by the method proposed in Section 4, with likelihood linearly mapped to intensity. Finally, for comparison row 4 shows the Cumulated Occluder Boundary Image (COBI) computed according to the method presented in [7]. The poor result of this method is due to the over-simplifying three-layer scene assumption.

Rows 5-9 of Figure 8 show the results for floor detection. Row 5 shows the manually generated floor region ground truth. Row 6 shows the floor likelihood estimated by the method proposed in Section 5 using geodesic superpixel segmentation, while row 7 shows the same algorithm on watershed superpixels. For comparison, Rows 8 and 9 show the binary floor regions generated by the fast marching segmentation method proposed in [19] and the region growing method proposed in [2]. Figure 9 shows another comparison of the proposed method with

(a) (b) (c) (d)

Figure 11: Video compositing: (a) superpixel segmentation; (b) optimal depth order assignment found by MRF; (c) synthetic targets; (d) synthetic targets composited into scene with correct occlusion handling.

[19] and [2] for an outdoor scene from [5]. This result highlights the advantage of the proposed method in finding all floor areas consistent with the appearance model, despite targets walking in only a relative narrow part of the scene.

Figure 10 shows the precision-recall curves for occlusion boundary and floor region detection, based on varying a detection threshold in the range [0, 1] for the likelihood maps shown in Figure 8. At each threshold, the recall was computed as the ratio of the number of true positive boundary/floor pixels (with likelihood exceeding the threshold) to the total number of ground truth boundary/floor pixels, and the precision was computed as the ratio of true positive boundary/floor pixels to the total number of detected boundary/floor pixels. At the equal error rate operating point (where the false detection rate equals the false rejection rate), our occlusion boundary detection method achieves a precision and recall of 0.64. By comparison, the COBI result shown in row 4 of Figure 8 achieves a recall of 0.28 and precision of 0.30. Similarly, our floor detection method achieves a precision and recall of 0.91 for superpixel based segmentation, and 0.94 for watershed based segmentation at the equal error operating point. By comparison, region growing achieves a recall of 0.81 and precision of 0.99, while fast marching achieves a recall of 0.60 and precision of 0.58.

7. Application: Video Compositing

Video compositing is the process of inserting synthetic objects into a real scene, and is a common task in applications such as augmented or mixed reality. In typical cluttered environments, the principal challenge is to ensure inserted objects respect scene occlusions. This can be achieved by decomposing the scene into partially transparent, depth ordered layers, such that opaque pixels in higher layers occlude those in lower layers. Synthetic content with correct occlusion handling can then be added by inserting new objects between the existing scene layers. We now show that a depth ordered layered representation suitable for compositing can be recovered from the proposed occlusion boundary and the floor likelihood cues.

Layer extraction may be posed as an optimal depth label assignment on a superpixel segmentation of the scene, which we solve using an MRF-based approach. An underlying assumption is that superpixel boundaries align with floor region boundaries and static occlusion boundaries. Let $D = \{d_0, \ldots, d_N\}$ represent the set of decreasing depth labels that can be assigned to superpixels. Now, a label assignment is desired that satisfies depth order constraints on superpixels separated by an occlusion boundary, and assigns greatest depth d_0 to floor regions with high likelihood.

The MRF is constructed with a node per superpixel and an edge joining neighbouring superpixels. Let S_i represent the set of pixels in the *i*-th superpixel, and x_i represent the assigned depth label. Then, the unitary node potentials $E_{floor}(x_i)$ penalize labels inconsistent with the average floor likelihood in S_i according to

$$E_{floor}(x_i) = \begin{cases} 1 - \sum_{p \in S_i} L_{floor}(p)/|S_i|, & if\ x_i = d_0 \\ \sum_{p \in S_i} L_{floor}(p)/|S_i|, & if\ x_i \neq d_0 \end{cases} \quad (11)$$

where $L_{floor}(p)$ is the floor likelihood at image location p.

To determine the pairwise node potentials, $E_{occ}(x_i, x_j)$, we first determine whether an occlusion boundary exists between regions s_i and s_j. A simple decision rule involves thresholding the occlusion boundary likelihood map and counting the number of high-likelihood occlusion pixels that occur on the boundary between s_i and s_j. If this number exceeds an empirically chosen threshold, the superpixels are assumed to lie at different depths. The relative depth of s_i and s_j may be determined from the average orientation of high-likelihood occlusion pixels on the boundary between s_i and s_j (see Section 4.1). Finally, the pairwise potential $E_{occ}(x_i, x_j)$ is constructed to penalize depth labels that violate the observed depth order. For example, if s_i is determined to be closer than s_j, the pairwise potential $E_{occ}(x_i, x_j)$ is constructed as

$$E_{occ}(x_i, x_j) = \begin{cases} \omega_{occ}, & if\ x_i \geq x_j \\ 0, & otherwise \end{cases}. \quad (12)$$

For superpixels that are not separated by an occlusion boundary, the pairwise potential enforces a smoothness constraint, given by

$$E_{occ}(x_i, x_j) = \begin{cases} \omega_{smooth}, & \text{if } x_i \neq x_j \\ 0, & otherwise \end{cases}. \quad (13)$$

In equations (12) and (13), ω_{occ} and ω_{smooth} are empirically chosen to balance the contribution of each potential function. Figure 11(b) shows the optimal label assignment (five depth layers encoded by increasing

intensity) found by solving the above MRF for the convenience store scenario using the floor and occlusion likelihood results shown in the second column of Figure 8.

Given the recovered depth layers, synthetic targets can now be inserted into the scene based on standard compositing techniques. Figure 11(c) shows three synthetic targets to be inserted just behind layers 4, 5 and 2 (from top to bottom). The composited result shown in Figure 11(d) is visually pleasing since all partial occlusions are faithfully reproduced.

8. Conclusions

This paper has demonstrated that static 3D scene structure can be recovered by analysing only the location and shape of moving targets in images of a video sequence from a static camera. Two cues related to ordinal depth were proposed: per-pixel static occlusion boundary likelihood and floor region likelihood. The proposed methods were shown to outperform similar methods in the literature. Furthermore, the ability to infer a pixel-wise ordinal depth map from these cues was described and demonstrated in a compositing application.

The proposed geometric cues are obtained in the absence of supervised training or prior information about camera calibration or scene geometry. Furthermore, these methods do not require occlusion boundaries to coincide with intensity boundaries, or require the floor appearance to be homogeneous or even smooth. The main assumptions are that sufficient foreground targets with varying shape and motion are present in the scene to reveal static geometry, and the appearance of the floor differs from non-floor regions.

While we expect the proposed methods to perform well in general based on our evaluation, there exist special cases which violate our assumptions. . Floor detection relies heavily on accurate footprints, which can be corrupted by shadows and reflections, especially on polished floors. This problem may be solved by detecting foreground targets using pedestrian classification instead of change detection. The requirement that the floor has a different appearance from other regions is generally reasonable, but will fail in specific scenarios or alternative sensing modalities such as thermal IR. For occlusion detection, persistent boundary pixels can arise away from occlusion boundaries in certain scenarios. This occurs in particular for combinations of repetitive or rigid motion of similar objects, such as trains, industrial robots or people standing in a queue. Aside from these specific scenarios, the described methods are effective in environments populated with free-ranging targets with varying shape, which applies to a wide range of practical applications.

References

[1] B. Bose and E. Grimson. Ground Plane Rectification by Tracking Moving Objects. In *VS-PETS*, 2003.

[2] A.V. Bovyrin and K.V Rodyushkin. Human Height Prediction and Roads Estimation for Advanced Video Surveillance Systems". In *AVSS*, 2005.

[3] G. J. Brostow and I. A. Essa. Motion Based Decompositing of Video. In *ICCV*, 1999.

[4] J. Coughlan and A. Yuille. Manhattan World: Orientation and Outlier Detection by Bayesian Inference. *Neural Computation*, 15(5):1063-1088, 2003.

[5] J. Davis and V. Sharma. Background-Subtraction using Contour-based Fusion of Thermal and Visible Imagery. *CVIU*, 106(2-3):162-182, 2007.

[6] D. F. Fouhey, V. Delaitre, A. Gupta, A. A. Efros, I. Laptev and J. Sivic. People Watching: Human Actions as a Cue for Single View Geometry. In *ECCV*, 2012.

[7] L. Guan, S. Sinha, J.-S. Franco and M. Pollefeys. Visual Hull Construction in the Presence of Partial Occlusion. In *3DPVT*, 2006.

[8] A. Gupta, A. A. Efros, M. Hebert. Blocks World Revisited: Image Understanding Using Qualitative Geometry and Mechanics. In *ECCV*, 2010.

[9] D. Hoiem, A. A. Efros, and M. Hebert. Putting Objects in Perspective. In *CVPR*, 2006.

[10] B. Jackson, R. Bodor and N. Papanikolopoulos. Learning Static Occlusions from Interactions with Moving Figures. In *IROS*, 2004.

[11] M. Keck and J. W. Davis. Recovery and Reasoning About Occlusions in 3D Using Few Cameras with Applications to 3D Tracking. *IJCV*, 95(3):240-264, 2012.

[12] J. Krumm, B. Brumitt and B. Meyers. Wallflower: Principles and Practice of Background Maintenance. In *ICCV*, 1999.

[13] O. Le Meur. Predicting Saliency Using Two Contextual Priors: the Dominant Depth and the Horizon Line. In *ICME*, 2011.

[14] F. Meyer. Topographic distance and watershed lines. *Signal Processing*, 38:113-125, 1994.

[15] F. Lv, T.Zhao and R. Nevatia. Camera Calibration from Video of a Walking Human. *PAMI*, 28(9):1513-1518, 2009.

[16] D. Marr. *Vision. A Computational Investigation into the Human Representation and Processing of Visual Information*. W.H. Freeman and Company, 1982.

[17] T. Q. Pham. Geodesic Superpixel Segmentation. Australian Patent Application No. 2011265383, Dec. 20, 2011.

[18] L. Roberts. Machine Perception of Three-Dimensional Solids. MIT: Lincoln Laboratory, Tech. Report #315, 1963.

[19] D. Rother, K. Patwardhan, and G. Sapiro. What can casual walkers tell us about a 3D scene? In *ICCV*, 2007.

[20] Y. Rubner, C. Tomasi and L. J. Guibas. A Metric for Distributions with Applications to Image Databases. In *ICCV*, 1998.

[21] A. Saxena, M. Sun and A. Y. Ng. Make3d: Learning 3D Scene Structure From a Single Image. *PAMI*, 31(5):824-840, 2009.

[22] A. Schödl and I. Essa. Depth Layers from Occlusions. In *CVPR*, 2001.

[23] J. Springett and J. Vendrig. Spatio-activity based object detection. Unpublished. Available online: arXiv:0803.1586 [cs.CV], 2008.