Monocular Object Detection Using 3D Geometric Primitives

Peter Carr[1], Yaser Sheikh[2], and Iain Matthews[1,2]

[1]Disney Research, Pittsburgh
[2]Carnegie Mellon University

Abstract. Multiview object detection methods achieve robustness in adverse imaging conditions by exploiting projective consistency across views. In this paper, we present an algorithm that achieves performance comparable to multiview methods from a single camera by employing geometric primitives as proxies for the true 3D shape of objects, such as pedestrians or vehicles. Our key insight is that for a calibrated camera, geometric primitives produce predetermined location-specific patterns in occupancy maps. We use these to define spatially-varying kernel functions of projected shape. This leads to an analytical formation model of occupancy maps as the convolution of locations and projected shape kernels. We estimate object locations by deconvolving the occupancy map using an efficient template similarity scheme. The number of objects and their positions are determined using the mean shift algorithm. The approach is highly parallel because the occupancy probability of a particular geometric primitive at each ground location is an independent computation. The algorithm extends to multiple cameras without requiring significant bandwidth. We demonstrate comparable performance to multiview methods and show robust, realtime object detection on full resolution HD video in a variety of challenging imaging conditions.

1 Introduction

Occupancy maps [1–6] fuse information from multiple views into a common world coordinate frame and are particularly useful for detecting 3D objects perpendicular to a plane, such as people, as they describe the probability of every ground plane location being occupied by an object. An occupancy map is calculated by quantizing the ground plane into a set of discrete locations. The probability of a particular (X, Y) ground location being occupied is determined by projecting the world location at a series of heights above the ground location into each of the cameras and aggregating the image evidence. The number of objects and their positions are inferred directly from the occupancy map. Occupancy maps exploit the fact that all views of an object are consistent projections of an unknown 3D shape. Previous work [2, 4] has shown occupancy maps to be robust to changing lighting conditions, shadows, camera shake, and limited resolution.

The performance of occupancy maps improves with additional cameras, as each new vantage point provides additional projective consistency constraints.

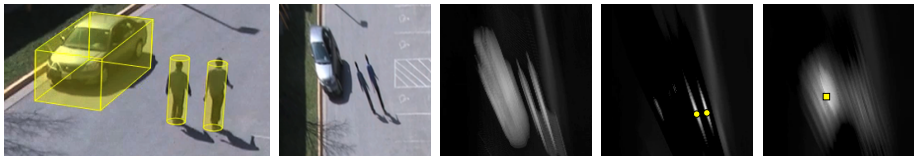


Fig. 1. Monocular Occupancy Maps: *The locations of objects are determined in the metric world frame. A monocular occupancy map does not exhibit well defined local maxima, making object detection difficult. Multiview methods generate strong sharp responses through consensus among different perspectives (see Figure 2). We achieve a similar outcome, but from a single view, by modelling objects as geometric primitives of particular sizes (pedestrians as cylinders and vehicles as cuboids). For each world location we derive a projected response template for each 3D geometric primitive. Location maps specific to cylinders and cuboids are generated from template similarity. Object positions (yellow circles and squares) are estimated using mean shift. The approach is highly parallel, allowing realtime performance on 1080p video.*

However, since the occupancy calculation requires simultaneous access to all pixels in all views, the algorithm does not easily scale to a large number of cameras. Centralized processing requires synchronized cameras and significant data bandwidth to aggregate and analyze multiple video streams in realtime. As a result, live systems using occupancy maps are often only deployed over small areas with a limited number of low resolution cameras.

We present an occupancy map based object detection approach which requires only monocular video, yet remains competitive with multiview methods, by loosely characterizing 3D object shape using geometric primitives (see Figure 1). Our key contribution is an analytical formation model of occupancy maps that arises from the convolution of an object location map with a spatially-varying kernel function. We show that the camera calibration and geometric primitive uniquely determine the kernel function. Deconvolving the occupancy map recovers object locations, and we efficiently approximate this process using template similarity. Precise object locations are recovered by finding the modes of the estimated probability density function of object locations using the mean shift algorithm [7]. Our results illustrate how geometric primitives improve the performance of monocular detection, and are competitive with multiview occupancy map methods for detecting pedestrians and vehicles.

Our method extends to multiple cameras and handles both overlapping and non-overlapping scenarios. Unlike traditional multiview occupancy maps, our algorithm permits each camera to process data in isolation. Only minimal network bandwidth is needed to transmit detections from each vantage point to a central location. As a result, we are able to achieve robust detection of people or vehicles at 30 fps in 1080p video with negligible latency, over large areas, across multiple cameras, and in challenging weather and lighting conditions, demonstrated on over 700 minutes of video.

Fig. 2. Multiview Occupancy Maps: *A soccer match is recorded from six cameras [10]. For clarity, cameras on far side of the pitch (top) have been laterally reversed, and the occupancy map superimposed over a schematic. The occupancy map exhibits 'X'-shaped patterns where players are located because the cameras are relatively low, and players are typically visible in two cameras simultaneously.*

2 Previous Work

Occupancy maps were previsouly used by Franco and Boyer [8] and Khan and Shah [1] for detecting people. Their insight was that an image to ground homography mapped a person's feet to a consistent ground location from any view point, making it possible to estimate the number of people and their locations by finding local maxima in the occupancy map generated from multiple cameras. Consistent mappings between different vantage points does not apply to just the ground plane. A person's head maps to the same location on the horizontal plane at the height of the person's head; Eshel and Moses [3] solved for the optimal object height, whereas Khan and Shah [4] used a fixed average height to increase robustness. Recently, occupancy maps have been formulated using an infinite number of planes. The projected silhouette of each vertical column at (X, Y) is evaluated in the image plane. If the image is warped such that areas become rectangles, the computation can be optimized using integral images [5, 9]. Alternatively, one can approximate the areas as rectangles [2, 6].

Modeling the salient shape of objects using geometric primitives for visual perception was proposed in the 1970's by Binford and colleagues in a series of papers [11–13], where they explored the generalized cylinder as a basic primitive. For humans in particular, Marr and Nishihara [14] introduced the idea of hierarchical geometric primitives with a single cylinder at the coarsest level and an articulated collection of progressively finer cylinders at more detailed levels. A number of other geometric primitives such as spheres, superquadrics, and ellipsoids have also been used [15–18] to represent bodies and limbs for recognition and tracking. Detailed reviews of shape representations used in a variety of tracking and recognition can be found in [19–21].

A number of papers have focused on joint understanding of object and scene geometry, such as [22–25]. Typically, values for some camera parameters are assumed (such as intrinsics [25] or height [22]) and the remainder are estimated through geometric consistencies between known object sizes and their projected images. [22] assumes upright cameras located at eye level and works for this specific case. Objects are represented as 2D bounding boxes. In 3D, the representation of [22] is equivalent to a series of infinitely thin frontoparallel planes extruding from the ground (i.e., parallel to the image plane). As camera height

increases (typical in surveillance, sporting, and other applications), the necessary assumptions of [22] break down: tops of objects become visible, and vertical lines in the world no longer align with the image axes. At higher vantage points, objects must be modelled in 3D. Additionally, their silhouettes rarely resemble perfect rectangles in the image plane (consider, for example, a camera with roll).

Pedestrian detection methods using sliding windows and/or features [26–30], or HOG-based detectors [31], in particular, perform well for a variety of objects. However, these methods often require additional information to remain robust to adverse imaging conditions. They are typically trained for a specific vantage point, and in the case of human detection, struggle with complex body poses. The algorithms are computationally intensive, making realtime operation rare [32]. Efficient implementations through approximation [33] and/or parallel execution [34] have been investigated recently.

3 Detecting Geometric Primitives in Monocular Occupancy Maps

Our goal is to estimate a set of object locations $\mathcal{L} = \{(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)\}$ on the ground plane $Z = 0$. For convenience, we define a 2D location map $\mathbf{L}(X, Y)$ to represent the collection of objects using 2D delta functions

$$\mathbf{L}(X, Y) = \sum_{i=1}^{n} \delta(X - X_i, Y - Y_i). \tag{1}$$

An occupancy map $\mathbf{O}(X, Y)$ describes the probability of each (X, Y) ground location being occupied by a portion of an object (see Figure 2). Every 3D volume element at (X, Y, Z) is assigned an occupancy probability based on the image evidence (which could be a binary image or continuous probability measure). The volume element probabilities are then integrated over height at each (X, Y) location.

Occupancy maps formulated from multiple overlapping views exhibit strong isolated peaks for tall thin objects [2, 4], making it reasonable to assume $\mathbf{O} \approx \mathbf{L}$. As a result, an estimate $\hat{\mathbf{L}}$ of object locations is typically formulated by searching for significant local peaks in $\mathbf{O}$ followed by non-maxima suppression to enforce object solidity [35]. Generally, $\mathbf{O}$ and $\mathbf{L}$ will be significantly different. Unlike the location map, the occupancy map contains projection artifacts that depend on the object and camera geometries. Since a pixel back-projects as a ray, the occupancy map will not contain delta function responses at object locations. Instead, object locations will coincide with the maxima of broader functions.

We characterize objects of interest as 3D geometric primitives of constant height. Vehicles and pedestrians, for instance, resemble cuboids and cylinders. For a camera located at $\mathbf{C} = (C_X, C_Y, C_Z)$, a geometric primitive $\phi \in \{\text{cylinder}, \text{cuboid}, \ldots\}$ at ground location $\mathbf{P}$ defines the 2D *projected primitive kernel* $\mathbf{K}_\phi(X, Y; \mathbf{P}, \mathbf{C})$. This kernel specifies the local spread in the occupancy map and is determined by the shape of its base (a circle for a cylinder and a rectangle

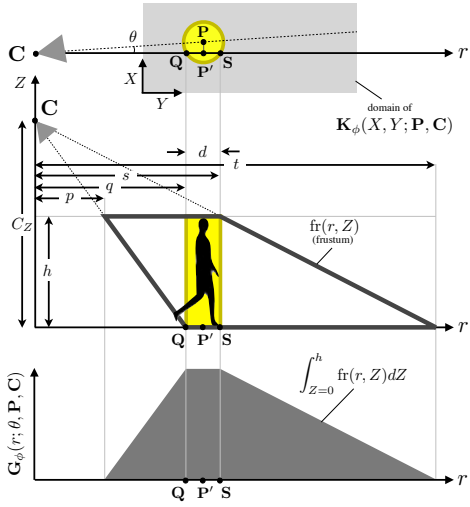


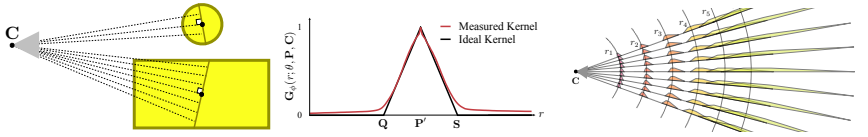
Fig. 3. Projected Primitive Kernel Profiles: *(Top) The overhead view of a camera and cylinder located at* $\mathbf{C}$ *and* $\mathbf{P}$ *respectively. We consider an arbitrary vertical cross section along the camera's line of sight passing through an interior point* $\mathbf{P}'$ *of the object. (Middle) The cylindrical cross section is a rectangle of height* h *and depth* $d = \|\mathbf{S}-\mathbf{Q}\|$*. The bounding rays from the camera intersect the ground at distances* q *and* t*, and an elevated horizontal plane at* p *and* s*, producing the frustum outlined in grey. (Bottom) The profile* $\mathbf{G}_\phi(r; \theta, \mathbf{P}, \mathbf{C})$ *of the projected primitive kernel* $\mathbf{K}_\phi(X, Y; \mathbf{P}, \mathbf{C})$ *for this particular cross section is generated by integrating the frustum vertically to produce a distinctive trapezoidal response.*

for a cuboid) and its extrusion height. The nature of the projected primitive kernel is best understood as a series of radial profiles $\mathbf{G}_\phi(r; \theta, \mathbf{P}, \mathbf{C})$. For convenience, we switch from rectilinear world coordinates (X, Y, Z) to cylindrical coordinates (r, θ, Z) originating at the camera's ground location $(C_X, C_Y, 0)$. We consider an arbitrary point $\mathbf{P'} = (X, Y, 0) \equiv (r, \theta, 0)$ lying within a geometric primitive located at $\mathbf{P}$. A vertical cross section passing through $\mathbf{C}$ and $\mathbf{P'}$ will result in a 2D rectangle of fixed height h and varying depth d (see Figure 3). The primitive's cross section will be bounded by rays which intersect the ground plane at distances q and t, and an elevated horizontal plane $Z = h$ at p and s. The projected primitive kernel's profile along this cross section θ is the integration of the frustum along the vertical axis between $Z = 0$ and $Z = h$, and is a trapezoid

$$\mathbf{G}_\phi(r; \theta, \mathbf{P}, \mathbf{C}) = \int_{Z=0}^{h} \text{fr}(r, Z)dZ. \tag{2}$$

The locations of $\mathbf{Q}$ and $\mathbf{S}$ are determined by the primitive's size, shape, and position; as well as the location of the camera $\mathbf{C}$ and the particular interior point $\mathbf{P'}$. From similar triangles, the extent of the integrated response before $\mathbf{Q}$ and after $\mathbf{S}$ is respectively $q - p = \frac{qh}{C_Z}$ and $t - s = \frac{sh}{C_Z - h}$.

Pedestrians and Vehicles. We represent pedestrians as cylinders 1.8m high and 0.5m in diameter. In practice, the cylinder is too large to approximate by



(a) (b) (c)

Fig. 4. Projected Primitive Kernels and Profiles: *(a) For a location* $\mathbf{P}$*, internal points are computed along the perpendicular to the camera's line of sight. Pedestrians (top) require fewer cross sections than vehicles, since they are narrower. (b) The expected projected pedestrian kernel profile along each cross section (right) is plotted against an average of more than* 3000 *detections in actual occupancy maps. There is good agreement between our model and experimental data. (c) The projected primitive kernel for a particular camera location and object shape varies with object position. The trapezoid extent and asymmetry increase with larger radial distances and lower camera heights.*

a single cross section (see Figure 4a). However, the top of the trapezoid cross section response is extremely narrow, so pedestrians appear as triangular profiles in occupancy maps (see Figure 4b).

We coarsely model vehicles as cuboids 2m wide, 4m long and 1.5m high. Unlike a cylinder, the depth of any cross section through the cuboid will depend on its orientation with respect to the camera. Vehicles often align with the direction of the road, and in some circumstances, it may be possible to infer the orientation of the cuboid for a given location. Generally, a series of orientation specific signatures are needed. In practice, four models are often sufficient, as that provides angular resolution of $\pm 22.5°$ (since the geometric primitive has no distinction between front/back). Vehicles are significantly wider than pedestrians, so several cross sections are necessary.

Formation Model. For a set of object locations $\mathbf{L}_\phi(X, Y)$ specific to a particular geometric primitive ϕ, the corresponding occupancy map will be the convolution of the location map with the spatially-varying projected primitive kernel. If multiple object types are present in a scene, the observed occupancy map will be a sum of the shape specific occupancy maps plus noise ϵ

$$\mathbf{O}(X, Y) = \sum_{\phi} \mathbf{L}_\phi(X, Y) * \mathbf{K}_\phi(X, Y; \mathbf{P}, \mathbf{C}) + \epsilon. \tag{3}$$

The process is analogous to the image formation model involving a point spread function. However, $\mathbf{K}_\phi$ differs from common lens point spread functions in that it is spatially varying and strongly asymmetric (see Figure 4c). Object detection now requires finding significant local peaks in the deconvolution of $\mathbf{O}$, where the spatially varying kernel is $\mathbf{K}_\phi$.

Approximate Deconvolution. Ideally, objects of a specific size and shape are detected by searching for significant local maxima in the deconvolution of the occupancy map $\mathbf{O}(X, Y)$. However, deconvolution is slow and sensitive to noise and precise camera calibration, and occupancy maps often contain errors from background subtraction and approximating the object's actual geometry by a geometric primitive. If a scene is not overly crowded, the convolution kernels will not overlap, and the projected primitive kernel will closely match the local

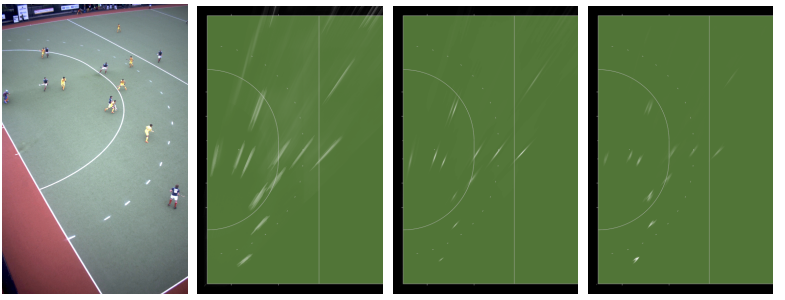


Image Occupancy Map Template Similarity Deconvolution

Fig. 5. Deconvolution vs Template Similarity: *Deconvolving the occupancy map for a spatially varying kernel using the Richardson Lucy algorithm produces strong responses at player locations. The similarity between the occupancy map and each location-specific projected primitive kernel produces similar strong responses at player locations, although the responses are broader than the deconvolution.*

occupancy scores. As a result, template matching can be employed instead of more computationally expensive deconvolution (see Figure 5).

The projected primitive kernel changes size and shape depending on the location of the object, making efficient template matching difficult. Evaluating local similarity to a spatially varying template is well suited to parallel execution. We use a GPU to compute the template matching score for each occupancy map location by comparing the local scores to the expected projected primitive kernel. For efficiency, we exploit the intrinsic properties of occupancy maps and evaluate similarity at a reduced number of samples along each cross section. The number of samples regulates a trade-off between detection performance and processing time. More samples produce sharper responses in the estimated deconvolution, but require more computation time. The estimated deconvolution $\hat{\mathbf{L}}_\phi(X,Y)$ at location (X,Y) is computed using the sum of squared differences

$$\hat{\mathbf{L}}_\phi(X,Y) = \exp\left(-\frac{\|\mathbf{K}_\phi(X,Y) - \mathbf{O}(X,Y)\|^2}{\|\mathbf{K}_\phi(X,Y)\|^2}\right). \tag{4}$$

The value is normalized with respect to unoccupied space to give context as to whether the difference between $\mathbf{K}_\phi$ and $\mathbf{O}$ is insignificant. For additional sensitivity, the values of $\mathbf{O}(X,Y)$ can be normalized for gain and bias to better match $\mathbf{K}_\phi(X,Y;\mathbf{P},\mathbf{C})$.

Mean Shift. The estimated $\hat{\mathbf{L}}_\phi$ deconvolution of the occupancy map will not resemble a combination of delta functions (see Figure 6). No solidity constraint has been enforced, i.e., a valid set of object locations $\mathbf{L}$ should not have objects occupying the same physical space [35]. We infer the number of objects and their locations using the mean shift algorithm. For efficiency, only ground plane locations having scores above a specified threshold are used as initial modes. The mean shift algorithm adjusts the number of modes and their locations to recover the final location map $\mathbf{L}_\phi$.

The bandwidth parameter of mean shift gauges the closeness of two locations. Since $\mathbf{L}$ is defined on the metric ground plane, object solidity is enforced by

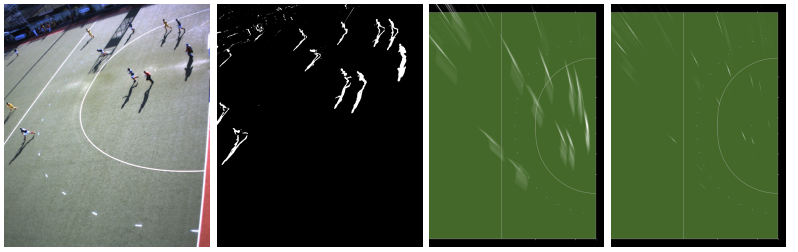


Fig. 6. Approximate Deconvolution: *In strong sunlight, shadows are detected as foreground objects. The occupancy map does not adequately suppress the background subtraction errors. However, a threshold applied to estimated deconvolution of a projected cylinder kernel discards the majority of the errors.*

combining modes that are less than one object width apart. We use the sample point estimator [36], which considers the projective uncertainty of every ground plane location when evaluating the distance between two locations.

We assume every point $\mathbf{p}$ on the image plane has constant isotropic uncertainty (which we arbitrarily define as 1% of the image diagonal) described by a covariance matrix $\Sigma_{\mathbf{p}}$. The corresponding location $\mathbf{P} = \mathtt{H}\mathbf{p}$ on the ground is determined by the image position $\mathbf{p}$ and a homography $\mathtt{H}$ extracted from the projection matrix, which also determines the covariance matrix $\Sigma_{\mathbf{P}} = \mathtt{H}\Sigma_{\mathbf{p}}\mathtt{H}^{\mathsf{T}}$ of the ground plane location $\mathbf{P}$ [37].

Multiple Views. Although our algorithm is designed for monocular views, it readily extends to multiple perspectives (which is useful for large and/or densely crowded areas) and naturally handles both overlapping and non-overlapping scenarios. Our monocular detector is run on multiple cameras in parallel, with each camera outputting a series of detected (X, Y) locations. Since cameras may overlap, it is entirely possible that the same object is detected in more than one camera simultaneously. Aggregating detections by concatenating the monocular results will not resolve multiple detections of the same object.

Detections which correspond to the same individual are identified by computing the Mahalanobis distances between all pairs of detections. Any detections which are less than one unit apart are clustered into a single detection. For a given set $i = \{1, 2, \ldots, n\}$ of ground plane detections, the best estimate of the object's position $\bar{\mathbf{P}}$ and uncertainty $\bar{\Sigma}_{\bar{\mathbf{P}}}$ is determined as [38]

$$\bar{\mathbf{P}} = \bar{\Sigma}_{\bar{\mathbf{P}}} \sum_{i=1}^{n} \left(\Sigma_{\mathbf{P}_i}^{-1} \mathbf{P}_i \right) \qquad \text{and} \qquad \bar{\Sigma}_{\bar{\mathbf{P}}} = \left(\sum_{i=1}^{n} \Sigma_{\mathbf{P}_i}^{-1} \right)^{-1}. \tag{5}$$

In other words, detections are combined by weighting each view by its uncertainty (see Figure 7). If an object is close to one camera but also detected in a distant camera, the distant detection will have significantly less weight because the uncertainty in its position will be much higher than that of the nearby camera.

Camera Height. The height C_z of the camera strongly influences detection robustness and localization precision. A camera which is low to the ground can

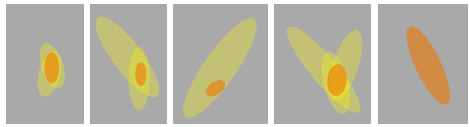


Fig. 7. Fused Detections: *Five examples of monocular detections fused using Eq. 5. Yellow ellipses represent the confidence interval of monocular detections, and red ellipses are the resulting fused detections. Large elliptical regions correspond to distant detections, while nearby detections appear as small ellipses. Objects detected in a single view (far right) appear as red ellipses.*

discriminate object height quite accurately, but the position estimate is imprecise. At the other extreme, the perspective of a top-down view makes it difficult to identify objects of a particular height, but the uncertainty in the location is quite small. The relation between image uncertainty and ground uncertainty is governed by a homography, but we can coarsely model the trend through trigonometry. We assume a camera at height C_z is oriented to look directly at the top of an object of height h and distance r (see Figure 3). The tilt angle θ of the camera is governed by $\tan\theta = \frac{C_z - h}{r}$. The derivative $\frac{dr}{d\theta} = (h - C_z)\csc^2\theta$ determines how the image uncertainty propagates to the ground plane uncertainty. Near the principal point the change in angle $d\theta = \frac{du}{f}$. We verify our localization uncertainty obeys this model using a constant image plane detection uncertainty for eight different cameras heights (see Figure 8).

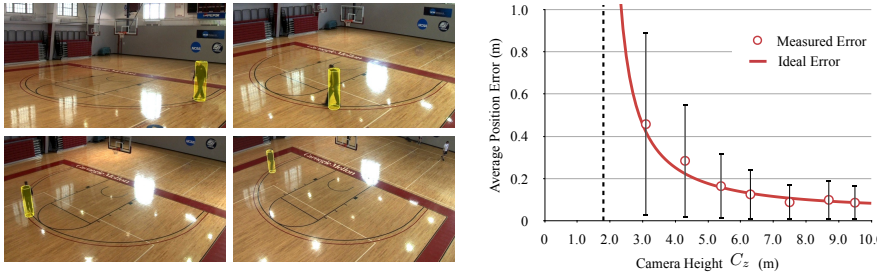


Fig. 8. Camera Height: *We observe a person from different camera heights (left) walking along a known curve on the ground plane. The average localization error (computed from approximately 1000 data points) is plotted as a function of camera height (right). As expected, the data point for $C_z = 1.7m$ failed to produce any detections, since the camera was not above the modelled pedestrian height. We fit a simplified trigonometric model between image plane error and detection uncertainty to the average position error at each camera height. The asymptote indicates the required assumption $C_z > h$.*

4 Experiments

We compare our approach to the POM algorithm [2] using its publicly available implementation[1]. For all experiments, we use an ATI Radeon HD 5770 GPU to compute the occupancy map for horizontal and vertical resolutions of 10 pixels/m, similar to [4]. Binary foreground masks are computed for each video

[1] http://cvlab.epfl.ch/software/pom

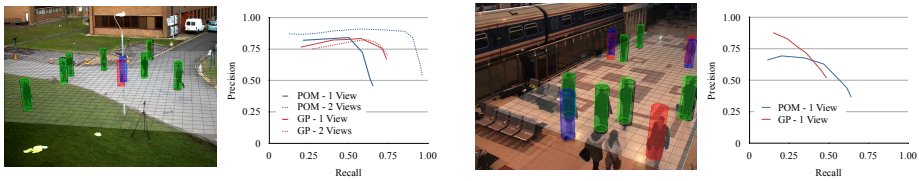


Fig. 9. Pedestrians: *Geometric primitives produce results competitive with POM on monocular sequences from the PETS 2009 (left) and 2006 (right) data sets. In the PETS 2009 data set, POM exhibits a significant boost in performance with multiple views, while GPs' results are similar to monocular performance (as expected). Correct (green), missed (blue) and false (red) detections for monocular geometric primitives are shown in the two exemplar camera images.*

frame using a per-pixel Gaussian appearance model (moving average over ten seconds). The occupancy score $\mathbf{O}(X, Y)$ is determined at each ground location using the average foreground score of all sampled vertical locations in the column. Non-geometric parameters, such as noise tolerance, were held constant for all experiments involving geometric primitives (GPs). As we will show, POM is more sensitive to good background subtraction results, and we found it necessary to re-tune the algorithm's non-geometric parameters for many experiments.

Pedestrians. We use views #1 and #2 from the PETS 2009 S2-L1 data set (a 20m $\times$ 20m area), and views #3 and #4 from the PETS 2006 S7-T6 data set (a 20m $\times$ 8m area). All cameras were calibrated using manually specified correspondences between camera images and a reference ground plane image. We computed the total number of true positives *tp*, false positives *fp* and false negatives *fn* over the entire sequence, and plot precision $= \frac{tp}{tp+fp}$ versus recall $= \frac{tp}{tp+fn}$ curves (see Figure 9).

Both data sets exhibit common trends in both monocular and multiview performance. In the monocular case, GPs and POM have similar *multiple object detection accuracy* [39] MODA $= 1 - \frac{fn+fp}{tp+fn}$ scores for a typical tolerance of 1m (see Table 1). GPs exhibit slightly higher recall and lower precision, but the discrepancy is due to the specific noise tolerance settings used in these experiments. The recall of both algorithms increases when a second view is added. However, the MODA performance of POM increases dramatically, whereas GPs' remains

	Monocular		Multiview	
	POM	GP	POM	GP
PETS 2009	0.527	**0.679**	**0.807**	0.645
PETS 2006	0.285	**0.425**	0.446	**0.472**

Table 1. At a tolerance of 1m, GPs' have slightly higher MODA scores than POM. POM's MODA scores improve significantly with multiple views, while GPs' remain similar to monocular performance, since our current fusion algorithm does not include extensive multiview reasoning.

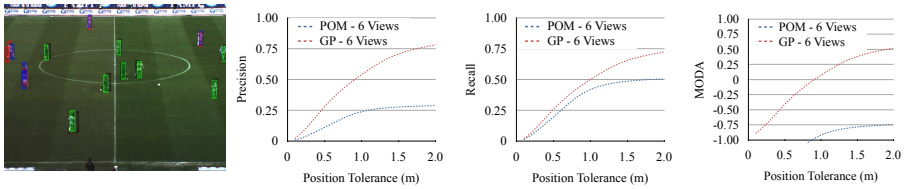


Fig. 10. Sports Players: *We consider the performance of the two algorithms at a tolerance of* $1m$*, since the misalignment (both spatially and temporally) between cameras makes precise measurements unlikely. Both algorithms produce roughly the same recall scores, but geometric primitives has half the number of false detections as POM.*

unchanged from the monocular case. POM simultaneously analyses information from both views, and is therefore able to reason about occlusion and projective consistency before detection. GPs, on the other hand, combine detections into a single result. Our algorithm does not attempt to suppress false detections through multiview occlusion reasoning. So, we expect GPs' multiview MODA characteristic to be close to the monocular case. Our MODA scores for POM on the PETS 2009 data set are slightly lower than those reported in [40]. Our gauge for a correct detection based on ground plane distance is a more difficult measure compared to rectangle overlap in the image, which explains the difference in the two performance numbers.

Sports Players. Outdoor sports have varying lighting and weather conditions, and the binary foreground masks are often noisy. We use a publicly available soccer data set [10] of six cameras (see Figure 2) to compare the performance of GPs and POM in these conditions. The POM algorithm failed to detect players using the parameter settings of the PETS data sets, so we increased its sensitivity. The following results are not fully optimized, but the long offline processing time limits tuning. The data set has synchronization errors, so ground truth locations that do not always overlap with the pixels of the actual players. As a result, the absolute performance numbers reported here are lower than the true values because of the background subtraction noise and calibration errors.

We also demonstrate the robustness and efficiency of our algorithm on a data set of ten complete field hockey games (each 70 minutes in length) collected from a two-week tournament. Eight 1080p cameras covered the 91.4m $\times$ 55.0m field and live results with one frame latency were generated in real-time at 30fps. Streams of detected (X, Y) locations were aggregated at a central machine. Games were played during the days and evenings, and in a variety of weather and lighting conditions (see Figure 11).

Vehicles. Geometric primitives are not limited to people. We illustrate the ability to detect vehicles on a publicly available surveillance data set [41] (see Figure 12). Four orientation specific detectors are constructed for a single geometric primitive to represent vehicles.

Run Time. Our implementation effectively operates in constant time (and orders of magnitude faster than POM). There are generally negligible linear dependencies on the number of cameras and image resolution. The mean shift stage

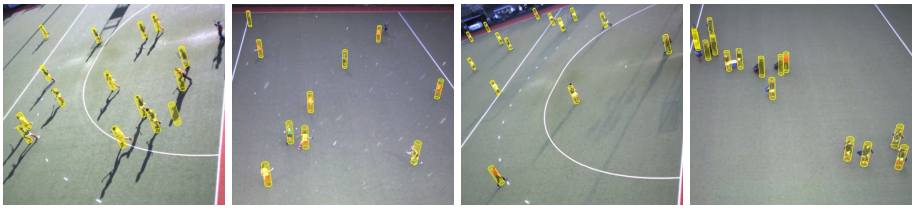


Fig. 11. Robust Realtime Performance: *Monocular 3D geometric primitives are able to handle strong shadows, rain, and long shadows. In addition to the extreme body poses, the detector is rarely confused with additional objects such as hockey sticks or equipment bags.*

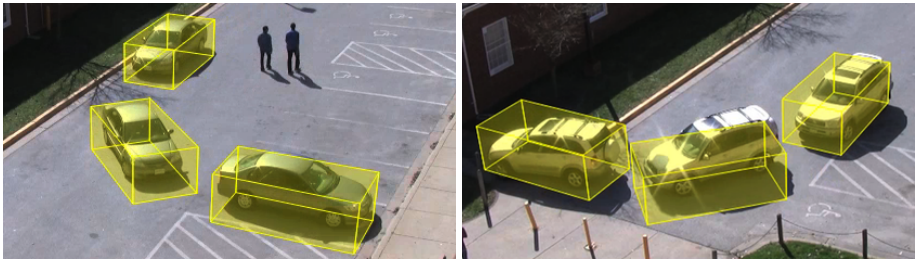


Fig. 12. Vehicle Detection: *Geometric primitives which are not radially symmetric, such as cuboids, must be detected in specific orientations. We detect vehicles using a fixed size cuboid for four specific orientations in the world. Since there is no distinction between front and back, we achieve angular resolution of* $\pm 22.5^\circ$.

has $O(N^2)$ complexity, but the size of N is usually insignificant (and a maximum number of iterations can be enforced if necessary). GPU readback speed is the major bottleneck.

5 Summary

Occupancy maps computed from a single camera exhibit significant blurring along the line of sight, making it difficult to localize objects precisely. The blur pattern, which we call a projected primitive kernel, is indicative of the object's size and shape. We define a formation model for occupancy maps which convolves object location maps with shape-specific spatially-varying projected primitive kernels. By modelling vehicles and pedestrians as cuboids and cylinders of fixed sizes, we are able to estimate the deconvolution of the occupancy map, and recover object locations.

Because object locations can be determined in each camera in isolation, our approach facilitates realtime detection across a large number of cameras. We have demonstrated detection on over 700 minutes of HD video footage from eight cameras (see accompanying video). Our current data fusion algorithm combines multiple detections of the same object from different cameras, but cannot perform multiview occlusion reasoning like POM. Our monocular performance is competitive with state of the art offline algorithms. Future work will explore better multiview data fusion algorithms.

References

1. Khan, S.M., Shah, M.: A multiview approach to tracking people in crowded scenes using a planar homography constraint. In: ECCV. (2006)
2. Fleuret, F., Berclaz, J., Lengagne, R., Fua, P.: Multicamera people tracking with a probabilistic occupancy map. PAMI **30** (2008) 267–282
3. Eshel, R., Moses, Y.: Homography based multiple camera detection and tracking of people in a dense crowd. In: CVPR. (2008)
4. Khan, S.M., Shah, M.: Tracking multiple occluding people by localizing on multiple scene planes. PAMI **31** (2009) 505–519
5. Delannay, D., Danhier, N., Vleeschouwer, C.D.: Detection and recognition of sports(wo)men from multiple views. In: ACM/IEEE International Conference on Distributed Smart Cameras. (2009)
6. Shitrit, H.B., Berclaz, J., Fleuret, F., Fua, P.: Tracking multiple people under global appearance constraints. In: ICCV. (2011)
7. Fukunaga, K., Hostetler, L.: The estimation of the gradient of a density function, with applications in pattern recognition. IEEE Transactions on Information Theory **21** (1975) 32 – 40
8. Franco, J.S., Boyer, E.: Fusion of multiview silhouette cues using a space occupancy grid. In: ICCV. (2005)
9. Yildiz, A., Akgul, Y.S.: A fast method for tracking people with multiple cameras. In: ECCV Workshop on HUMAN MOTION Understanding, Modeling, Capture and Animation. (2010)
10. D'Orazio, T., Leo, M., Mosca, N., Spagnolo, P., Mazzeo, P.L.: A semi-automatic system for ground truth generation of soccer video sequences. In: AVSS. (2009)
11. Binford, T.O.: Visual perception by computer. In: IEEE Conf. on Systems and Control. (1971)
12. Agin, G.J.: Representation and Description of Curved Objects. PhD thesis, Stanford University (1972)
13. Nevatia, R., Binford, T.O.: Description and recognition of curved objects. AI **8** (1977) 77–98
14. Marr, D., Nishihara, H.K.: Representation and recognition of the spatial organization of three-dimensional shapes. Proceedings of the Royal Society of London. Series B, Biological Sciences **200** (1978) 269–294
15. O'Rourke, J., Badler, N.: Model-based image analysis of human motion using constraint propagation. PAMI **2** (1980) 522–536
16. Barr, A.: Global and local deformations of solid primitives. Computer Graphics **18** (1984) 21–30
17. Azarbayejani, A., Pentland, A.: Real-time self-calibrating stereo person tracking using 3-D shape estimation from blob features. In: ICPR. (1996)
18. Farrell, R., Oza, O., Zhang, N., Morariu, V.I., Darrell, T., Davis, L.S.: Birdlets: Subordinate cateogrization using volumetric primitives and pose-normalized appearance. In: ICCV. (2011)
19. Moeslund, T.B., Granum, E.: A survey of computer vision-based human motion capture. CVIU **81** (2001) 231 – 268
20. Moeslund, T.B., Hilton, A., Krüger, V.: A survey of advances in vision-based human motion capture and analysis. CVIU **104** (2006) 90 – 126
21. Yilmaz, A., Javed, O., Shah, M.: Object tracking: A survey. ACM Comput. Surv. **38** (2006)

22. Hoiem, D., Efros, A.A., Hebert, M.: Putting objects in perspective. In: CVPR. (2006)
23. Leibe, B., Leonardis, A., Schiele, B.: Robust object detection with interleaved categorization and segmentation. IJCV **77** (2008) 259–289
24. Cornelis, N., Leibe, B., Cornelis, K., Gool, L.: 3d urban scene modeling integrating recognition and reconstruction. Int. J. Comput. Vision **78** (2008) 121–141
25. Wojek, C., Roth, S., Schindler, K., Schiele, B.: Monocular 3D scene modeling and inference: understanding multi-object traffic scenes. In: ECCV. (2010)
26. Haritaoglu, I., Harwood, D., Davis, L.: W4: real-time surveillance of people and their activities. PAMI **22** (2000) 809 –830
27. Leibe, B., Seemann, E., Schiele, B.: Pedestrian detection in crowded scenes. In: CVPR. (2005)
28. Tuzel, O., Porikli, F., Meer, P.: Human detection via classification on riemannian manifolds. In: CVPR. (2007)
29. Viola, P., Jones, M., Snow, D.: Detecting pedestrians using patterns of motion and appearance. In: ICCV. (2003)
30. Wu, B., Nevatia, R.: Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet part detectors. IJCV **75** (2007) 247–266
31. Dalal, N., Triggs, B.: Histograms of orientated gradients for human detection. In: CVPR. (2005)
32. Enzweiler, M., Gavrila, D.M.: Monocular pedestrian detection: Survey and experiments. PAMI **31** (2009) 2179–2195
33. Dollár, P., Belongie, S., Perona, P.: The fastest pedestrian detector in the west. In: BMVC. (2010)
34. Prisacariu, V.A., Reid, I.: fastHOG – a real-time GPU implementation of HOG. Technical Report 2310/09, University of Oxford (2009)
35. Hayes, P.J.: The second naive physics manifesto. In Hobbs, J., Moore, R., eds.: Formal Theories of the Commonsense World. Ablex (1985)
36. Sain, S.R., Scott, D.W.: On locally adaptive density estimation. Journal of the American Statistical Association **91** (1996) 1525–1533
37. Criminisi, A.: Accurate Visual Metrology from Single and Multiple Uncalibrated Images. PhD thesis, University of Oxford (1999)
38. Orechovesky, Jr., J. R.: Single source error ellipse combination. Master's thesis, Naval Postgraduate School (1996)
39. Kasturi, R., Goldgof, D., Soundararajan, P., Manohar, V., Garofolo, J., Bowers, R., Boonstra, M., Korzhova, V., Zhang, J.: Framework for performance evaluation of face, text, and vehicle detection and tracking in video: Data, metrics, and protocol. PAMI **31** (2009) 319 –336
40. Berclaz, J., Shahrokni, A., Fleuret, F., Ferryman, J., Fua, P.: Evaluation of probabilstic occupancy map people detection for surveillance systems. In: IEEE International Workshop on Performance Evaluation of Tracking and Surveillance. (2009)
41. Oh, S., Hoogs, A., Perera, A., Cuntoor, N., Chen, C.C., Lee, J.T., Mukherjee, S., Aggarwal, J., Lee, H., Davis, L., Swears, E., Wang, X., Ji, Q., Reddy, K., Shah, M., Vondrick, C., Pirsiavash, H., Ramanan, D., Yuen, J., Torralba, A., Song, B., Fong, A., Roy-Chowdhury, A., Desai, M.: A large-scale benchmark dataset for event recognition in surveillance video. In: CVPR. (2011)