Coherent Object Detection with 3D Geometric Context from a Single Image

Jiyan Pan
Carnegie Mellon University
Pittsburgh, PA, U.S.A.
jiyanpan@cs.cmu.edu

Takeo Kanade
Carnegie Mellon University
Pittsburgh, PA, U.S.A.
Takeo.Kanade@cs.cmu.edu

Abstract

Objects in a real world image cannot have arbitrary appearance, sizes and locations due to geometric constraints in 3D space. Such a 3D geometric context plays an important role in resolving visual ambiguities and achieving coherent object detection. In this paper, we develop a RANSAC-CRF framework to detect objects that are geometrically coherent in the 3D world. Different from existing methods, we propose a novel generalized RANSAC algorithm to generate global 3D geometry hypotheses from local entities such that outlier suppression and noise reduction is achieved simultaneously. In addition, we evaluate those hypotheses using a CRF which considers both the compatibility of individual objects under global 3D geometric context and the compatibility between adjacent objects under local 3D geometric context. Experiment results show that our approach compares favorably with the state of the art.

1. Introduction

When we look at the two image patches shown in the lower-left corner of Figure 1, it is hard to tell what is contained in the green box, yet it is likely that the red box encloses a pedestrian dressed in black. However, when we look at the entire image and obtain a sense of the 3D scene layout behind the image, we can tell for sure that the green box actually contains a car, since it rests on the road with the right pose, size, and location. Also, the red box cannot contain a pedestrian, since our sense of 3D geometry tells us that the content in the red box would have been too short for a pedestrian. This example illustrates the importance of 3D geometric context in resolving visual ambiguities encountered in object detection.

While the majority of existing research focuses on 2D context such as object co-occurrences and relative locations in the 2D image plane [11, 7, 19, 2, 4], some researchers have done pioneering works on utilizing 3D geometric context in object detection and have shown promising results [14, 3, 18]. The general idea is when 3D geometry is recovered together with object detection, the joint optimal solution would enforce 3D geometric coherence among detected objects and therefore improve object detection performance. Our work follows this line of research, yet with several important differences in multiple aspects.

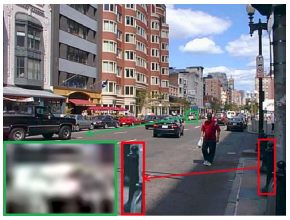

Figure 1. 3D geometric context explains away visual ambiguities.

To handle 3D geometric context, the first issue to consider is modeling the 3D geometry of the scene (a.k.a. global 3D geometry). Hoiem *et al*. represents the scene geometry with a ground plane parameterized by its pitch angle (*i.e*. horizon position) and height with respect to the camera [14]. The ground plane is more flexible in [3] and [18] where it is allowed to have a non-zero roll angle (*i.e*. horizon could be tilted in the image). In all those works, ground plane parameters are quantized into a small number of bins for tractability.

Different from the existing works, we also model gravity direction in addition to the ground plane (please see Figure 3), so that scenes like sloped streets can be represented as well. In addition, our approach does not require the quantization of continuous parameters, allowing for a much larger value range and a higher precision.

The second issue is relating the 2D appearance of individual objects to the global 3D geometry. Hoiem *et al*. derives an approximate relationship between ground plane parameters and the position and height of object bounding box in the image [14]. As only bounding box information is used, the simplified geometric relationship is effective only when ground plane roll is zero and ground plane pitch is small. The methods proposed in [3] and [18] use object 2D appearance to estimate its pitch angle with respect to the camera, and compute the ground plane from at least 3 objects.

In our proposed approach, we use object 2D appearance to estimate a richer set of properties (both pitch and roll angles as well as landmark locations), such that each individual object is able to establish the ground plane or gravity direction, removing the restriction that at least 3 objects must be present.

The third issue is jointly estimating global 3D geometry and detecting objects. Hoiem *et al*. build a Bayes net in which every object candidate is attached to the common global 3D geometry (*i.e*. ground plane pitch and height) [14]. Inference over the Bayes net gives the optimal ground plane parameters and the validity of each candidate. To make the inference tractable, the ground plane is assumed to have zero roll angle, and the quantization of the ground plane parameters is relatively coarse. Bao *et al*. propose to enumerate all possible quantized ground planes within a predefined range [3]. For each enumerated ground plane hypothesis, the validity of object candidates are checked against it, and the ground plane with the highest compatibility is chosen. As this approach performs exhaustive search, the search space has to be confined in a narrow range (2 degrees in pitch and 20 degrees in roll). An improved method is presented in [18], where all object candidates cast votes for the ground plane parameters in a Hough voting space, and the peak in the voting space is regarded as the optimal ground plane. A weakness of this approach is that when object detection is noisy, which is often the case, false detections would corrupt the votes.

We propose a novel way to generate global 3D geometry hypotheses using a generalized RANSAC algorithm that a) does not require quantization or exhaustive search over limited range, b) suppresses corruption of hypotheses caused by false detections, and c) improves accuracy of hypotheses by reducing the noise of inaccurate estimates obtained from true detections (please see Figure 6). Another novelty of our approach is that surface regions are also involved in generating and evaluating global 3D geometry hypotheses.

In addition, different from the aforementioned works that only consider the constraints of global 3D geometry imposed on each individual object candidate, we also introduce local 3D geometric constraints between object candidates, and integrate the global and local 3D geometric constraints in a Conditional Random Field(CRF) [16] (please see Figure 8). Experiments on 422 challenging outdoor images from the LabelMe dataset [17, 14] confirm the effectiveness of our RANSAC-CRF framework.

The rest of this paper is organized as follows. After an overview of our algorithm in Section 2, we describe in Section 3 how we generate the hypotheses of global 3D geometry in a way that suppresses outliers and reduces noise simultaneously. Evaluation of those hypotheses that incorporates both global and local 3D geometric constraints is detailed in Section 4. We present our experiment results in Section 5, and conclude the paper in Section 6.

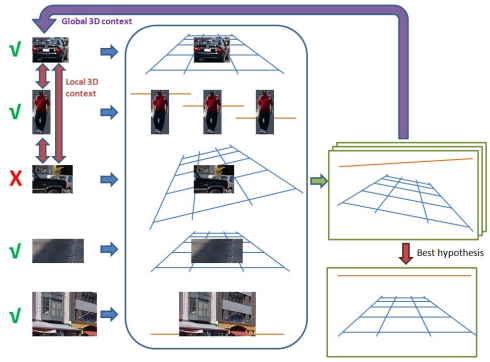

Figure 2. The overall scheme of our algorithm. Here, gravity direction w.r.t. camera is represented by the orange horizon line, and the ground plane is represented by the blue mesh.

2. Overview

The overall scheme of our algorithm is illustrated in Figure 2. We start by generating object/surface candidates (including cars, pedestrians, vertical and horizontal surface regions) using state-of-the-art object detectors (*e.g*. Deformable Part Model [8, 9]) and surface segmentation algorithms (*e.g*. Surface Layout Model [13, 12]). Each object/surface candidate gives an estimate of the global 3D geometry (*i.e*. gravity direction and ground plane parameters) based on their 2D appearance. Those noisy estimates are then pooled together using an generalized RANSAC algorithm to generate a set of global 3D geometry hypotheses. Given each hypothesis, we compute the compatibility of each object/surface candidate and infer their validity according to global and local 3D geometric context. The quality of each hypothesis is obtained as a result of the inference procedure. Finally the hypothesis with the highest quality is selected as the optimal estimate of the global 3D geometry, and the inference result of object candidate validity associated with the best hypothesis gives the final object detection result.

3. Generating global 3D geometry hypotheses

3.1. Modeling the scene and objects in it

In our work, we use exclusively the camera coordinate system illustrated in Figure 3. Aside from the focal length f, we categorize all the variables in Figure 3 into two groups. The first group contains global variables depicting the global 3D geometry: (inverse) gravity direction $\mathbf{n_g}$, ground plane orientation $\mathbf{n_p}$, and ground plane height h_p. The second group contains local variables specific to individual objects: object vertical orientation $\mathbf{n_v}$, object pitch angle θ, object roll angle γ, object depths d_t and d_b for top and bottom landmarks , locations $\mathbf{x_t}$ and $\mathbf{x_b}$ of the top and bottom landmarks in the image, real world height H of the

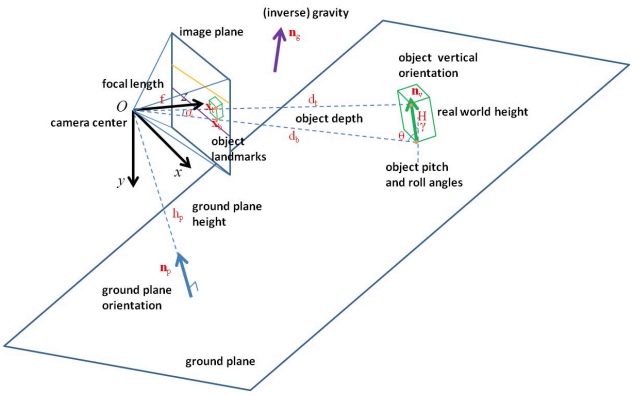

Figure 3. Modeling scene and objects in the camera coordinate system. The orange and purple lines in the image plane are ground horizon and gravity horizon, respectively.

top landmark, and object vertical viewing angle α.

Now we list all the geometric relationships that exist among those variables:

$$d_t = H \sin\theta / \sin\alpha; \tag{1}$$

$$d_b = H(\sin\theta / \tan\alpha + \cos\theta); \tag{2}$$

$$\mathbf{n_v} = \mathbf{m}\{d_t \mathbf{r}\{\mathbf{x_t}, f\} - d_b \mathbf{r}\{\mathbf{x_b}, f\}\}; \tag{3}$$

$$\mathbf{n_v} = \mathbf{g}\{-\mathbf{r}\{\mathbf{x_b}, f\}, \theta, \gamma\}; \tag{4}$$

$$\alpha = \arccos\{< \mathbf{r}\{\mathbf{x_t}, f\}, \mathbf{r}\{\mathbf{x_b}, f\} >\}; \tag{5}$$

$$h_p = -d_b < \mathbf{n_p}, \mathbf{r}\{\mathbf{x_b}, f\} >; \tag{6}$$

$$\mathbf{n_v} = \begin{cases} \mathbf{n_p} & \text{for cars} \\ \mathbf{n_g} & \text{for pedestrians.} \end{cases} \tag{7}$$

Here, function $\mathbf{r}\{\mathbf{x}, f\}$ computes the unit vector pointing from the camera center towards pixel location $\mathbf{x}$ in the image plane. Function $\mathbf{m}\{\mathbf{v}\}$ normalizes a vector $\mathbf{v}$ to unit length. Function $\mathbf{g}\{\mathbf{v}, \theta, \gamma\}$ rotates a unit vector $\mathbf{v}$ by pitch angle θ and roll angle γ. Function $< \mathbf{v_1}, \mathbf{v_2} >$ computes the inner product of two vectors.

In addition, we could also estimate the distributions of object pitch and roll angles given object appearance I and category c:

$$\theta \sim p_1(I, c); \tag{8}$$

$$\gamma \sim p_2(I, c). \tag{9}$$

Another cue we could use is the prior knowledge that predicts the distribution of object height given its category c:

$$H \sim p_3(c). \tag{10}$$

Among all the geometric variables, only $\mathbf{x_t}$ and $\mathbf{x_b}$ are directly observable. When multiple object candidates are present, we are faced with a large set of equations containing non-linear and even non-deterministic constraints. Moreover, many false detections would produce invalid equations. Therefore, instead of attempting to directly solve them, we use those equations to propose hypotheses of the global 3D geometry $(\mathbf{n_g}, \mathbf{n_p}, h_p)$, and to evaluate those hypotheses.

Input: I -- The image patch enclosing the object candidate
Output: $D[\mathbf{n_v}]$ -- Non-parametric distribution of the vertical orientation of the object candidate
Predefined: $\mathbf{G_n} = \{\mathbf{n_i}\}$ -- A dense grid on the upper half of the unit sphere
Procedure:
1. Apply pose regressor to image patch I to obtain the mean and variance of the pitch angle θ and the roll angle γ. (Constraints 8 and 9)
2. Apply landmark regressor to image patch I to obtain the mean and variance of the top landmark $\mathbf{x_t}$ and bottom landmark $\mathbf{x_b}$.
3. Apply prior knowledge to obtain the mean and variance of the real world height H of the top landmark. (Constraint 10)
4. Sample a large number of $\{\theta, \gamma, \mathbf{x_t}, \mathbf{x_b}, H\}$ according to the mean and variance of those variables
5. For each sample of $\{\theta, \gamma, \mathbf{x_t}, \mathbf{x_b}, H\}$:
   5.1. Apply constraint 4 to obtain a sample of vertical orientation $\mathbf{n_{v1}}$.
   5.2. Apply constraints 5, 1, 2, and 3 to obtain another sample of vertical orientation $\mathbf{n_{v2}}$.
6. Pool together all samples of $\mathbf{n_{v1}}$ and $\mathbf{n_{v2}}$, and apply kernel density estimation (KDE) to obtain the density at each node $\mathbf{n_i}$ on the grid $\mathbf{G_n}$.
7. Normalize the densities by the total number of samples to obtain the final distribution $D[\mathbf{n_v}]$.

Figure 4. Algorithm for generating a non-parametric distribution of the vertical orientation of an object candidate. The space under consideration covers the entire upper dome of unit sphere.

3.2. Each object/surface candidate gives an estimate of global 3D geometry

For each object candidate: Using the constraints from the equations listed above, we estimate a non-parametric distribution of the vertical orientation $\mathbf{n_v}$ of each object candidate from the appearance of the image patch enclosing it. Instead of resorting to a multi-view object detector that returns only a handful of discrete object poses, we directly estimate the mean/variance of the pitch and roll angles of a detected object by applying a regressor trained using the Hierarchical Discriminant Regression model (HDR) [15] with HoG features [6] on the LabelMe training dataset [14, 17]. We also estimate the mean/variance of the 2D positions of the landmarks using the HDR regressor. Here, the top and bottom landmarks for a car are the top and bottom locations of the wheel closest to the camera (which are stabler and better-defined), and for a pedestrian, they are the head and foot locations. The pitch and roll angles, together with landmark locations, produce a non-parametric distribution of object vertical orientation $\mathbf{n_v}$, according to the algorithm summarized in Figure 4. Following constraint 7, the vertical orientation distributions obtained from cars are regarded as the estimations of the ground plane orientation $\mathbf{n_p}$, and those from pedestrians are for the gravity direction $\mathbf{n_g}$.

In addition to estimating $\mathbf{n_g}$ and $\mathbf{n_p}$, given the vertical orientation, each object candidate also provides cues for the ground plane height h_p according to its size and location. The algorithm for using an object candidate to generate a non-parametric distribution of h_p is summarized in Figure 5.

For each surface candidate: Given a vertical surface region like a building facade, we extract long edges within it and compute vertical and horizontal vanishing points (VP) using Gaussian sphere [1]. To account for uncertainty, each vanishing point is represented by a set of circle-intersection points *on the Gaussian sphere*. The vertical direction of the surface $\mathbf{n_v}$ can be estimated directly from the vertical

Input: I -- The image patch enclosing the object candidate
$\mathbf{n_v}$ -- Vertical orientation of the object candidate
$\mathbf{n_p}$ -- Ground plane orientation
Output: $D[h_p]$ -- Non-parametric distribution of the ground plane height
Predefined: $\mathbf{G_h} = \{h_i\}$ -- A dense grid on the scalar range from 0 to 50
Procedure:
1. Apply landmark regressor to image patch I to obtain the mean and variance of the top landmark $\mathbf{x_t}$ and bottom landmark $\mathbf{x_b}$.
2. Apply prior knowledge to obtain the mean and variance of the real world height H of the top landmark. (Constraint 10)
3. Sample a large number of $\{\mathbf{x_t}, \mathbf{x_b}, H\}$ according to the mean and variance of those variables
4. For each sample of $\{\mathbf{x_t}, \mathbf{x_b}, H\}$:
   4.1. Apply constraints 4, 5, 2, 6 to obtain a sample of ground plane height h_p.
5. Pool together all samples of h_p, and apply kernel density estimation (KDE) to obtain the density at each node h_i on the grid $\mathbf{G_h}$.
6. Normalize the densities by the total number of samples to obtain the final distribution $D[h_p]$.

Figure 5. Algorithm for using an object candidate to generate a non-parametric distribution of the ground plane height. The range under consideration covers 50m.

VP, or from the cross-product of a pair of horizontal VPs. Therefore, for the vertical VP, it directly yields a set of $\mathbf{n_v}$ samples from its constituent circle-intersection points. For each pair of horizontal VPs, we compute the cross product of their respective constituent circle-intersection points and generate a set of $\mathbf{n_v}$ samples. The $\mathbf{n_v}$ samples from the vertical VP and all pairs of horizontal VPs are pooled together to generate a non-parametric distribution of $\mathbf{n_v}$ over the dense grid $\mathbf{G_n}$ using Kernel Density Estimation (KDE).

Estimating the vertical direction of a horizontal surface region (*e.g.* a road) is similar, except that its vertical VP is highly unreliable and therefore not used.

In outdoor street scenes, vertical surfaces usually correspond to building facades which typically agree with the gravity direction, while horizontal surfaces usually fall on roads which relate to the ground plane orientation. Therefore, we have $\mathbf{n_v} = \mathbf{n_g}$ for vertical surfaces and $\mathbf{n_v} = \mathbf{n_p}$ for horizontal surfaces.

An example of each object/surface candidate giving an estimate of the global 3D geometry is shown in Figure 6a and b. The estimates look messy, partly due to the existence of several false detections, and partly due to the estimation noise in true detections. In the next subsection, we discuss how to generate at least one good hypothesis from those messy estimates.

3.3. Generating hypotheses of global 3D geometry with generalized RANSAC

One of the keys for RANSAC to succeed is that at least one hypothesis should be close to the ground truth. In our case, a single object/surface candidate (*i.e.* observation) alone can generate a hypothesis of the global 3D geometry. Ideally, we could simply use a single observation (*i.e.* the minimal set) to generate a hypothesis. However, as single observations (even if they are true detections) tend to be noisy, it is likely that none of the hypotheses generated by the minimal set is close to the ground truth. On the other hand, if we use all the observations with equal weights, false

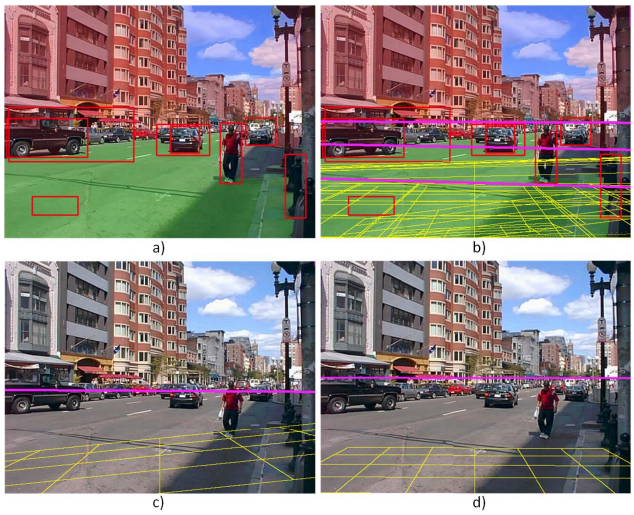

Figure 6. Generate hypotheses of global 3D geometry from object/surface candidates. a) Object/surface candidates. Here, red and green shades indicate vertical and horizontal surface candidates, respectively. b) Estimates of global 3D geometry given by individual object/surface candidates. Here, magenta lines represent gravity horizons, and yellow grids indicate ground planes, where the grid size is 1m. For display purposes, only the mode of each non-parametric distribution is shown. c) A bad hypothesis. d) A good hypothesis.

detections would corrupt the hypothesis. Therefore, we propose a generalized RANSAC algorithm to both inhibit outliers and reduce noise.

After each object/surface candidate has estimated a distribution of the global 3D geometry, we generate a set of mixed distributions by mixing individual distributions together with randomly generated weights. For each mixed distribution, we find its modes using the mean-shift algorithm [5] and take those modes as hypotheses. When the set of mixed distributions is large enough, at least one of them would mostly come from valid object/surface candidates. Furthermore, by finding modes of their mixed distribution (which is equivalent to averaging) we also reduce the noise level.

To verify this claim, we perform experiments on the 100 images that are provided with the ground truth horizon in Hoiem's dataset [14], and the results are plotted in Figure 7. Here, we compare the error of hypothesis generation in there cases: 1) obtaining hypotheses by directly using the modes from each individual distribution estimated by each object/surface candidate (red circle); 2) obtaining hypotheses by computing the modes of the average distribution over *all* the distributions estimated by object/surface candidates (magenta square); 3) obtaining hypotheses using our generalized RANSAC approach (blue curve). The error of hypothesis generation is defined as the difference between the ground truth and the *best* hypothesis among all the hypotheses generated.

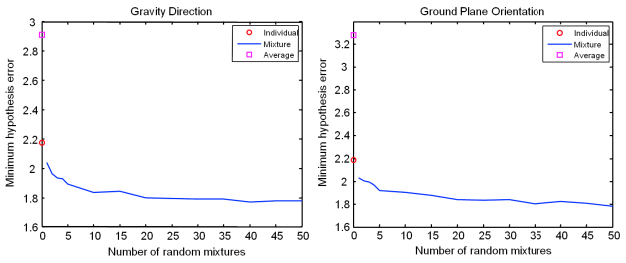

Figure 7. Comparing different methods of generating hypotheses. The unit of the y-axis is degree. Please see text for details.

We can see that the generalized RANSAC approach has the smallest error. Also, as the size of the set of random mixtures grows, the error decreases. This is expected, because the ground truth is more likely to be covered when we try more combinations. In our experiment, generating 50 random mixtures is sufficient.

Two qualitative examples of bad and good hypotheses from the hypothesis set are shown in Figure 6c and d.

4. Evaluating global 3D geometry hypotheses

Given a global 3D geometry hypothesis $(\tilde{\mathbf{n}}_\mathbf{g}, \tilde{\mathbf{n}}_\mathbf{p}, \tilde{h}_p)$, we evaluate its quality by measuring how well it is supported by object/surface candidates after excluding the influence of outliers. For this purpose, we evaluate the global and local geometric compatibilities of each object/surface candidate, and employ a CRF to infer the validity of each candidate. The optimal score of the objective function used in the CRF inference is regarded as the quality of the current global 3D geometry hypothesis.

4.1. Global geometric compatibility

Global geometric compatibility refers to the compatibility of an individual object/surface candidate w.r.t. the global 3D geometry such as ground plane and gravity. An illustrative example is given in Figure 8 where the red objects violate global geometric constraints.

Individual object candidate: We use two sources of geometric constraints to compute the global compatibility of an individual object candidate. In both the two sources, landmark locations $\mathbf{x_t}$ and $\mathbf{x_b}$ take the mean value produced by the landmark regressor. The first source compares the pitch and roll angles predicted by the pose regressor (using constraints 8 and 9) with those directly computed from the current hypothesis of the global 3D geometry (using constraint 4 where $\mathbf{n_v} = \tilde{\mathbf{n}}_\mathbf{g}$ for pedestrians and $\mathbf{n_v} = \tilde{\mathbf{n}}_\mathbf{p}$ for cars). The resulting compatibility score is

$$s_{g1} = \exp\{-\frac{(\tilde{\theta} - \theta_0)^2}{2\sigma_\theta^2}\} \cdot \exp\{-\frac{(\tilde{\gamma} - \gamma_0)^2}{2\sigma_\gamma^2}\} - 0.5, \quad (11)$$

where θ_0, γ_0 and σ_θ^2, σ_γ^2 are the mean and variance of the pitch and roll regressor outputs, respectively. $\tilde{\theta}$ and $\tilde{\gamma}$ are computed according to constraint 4.

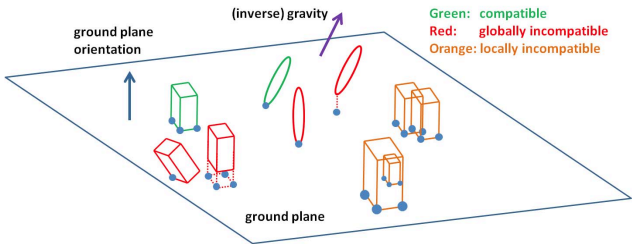

Figure 8. Different types of geometric context. Here, cubes represent cars and ellipsoids represent pedestrians. Blue dots indicate ground touching points.

The second source of geometric constraints involves the real-world height of the object. Given the ground plane hypothesis $\tilde{\mathbf{n}}_\mathbf{p}$ and $\tilde{h}_p$, we compute the bottom landmark depth d_b using constraint 6. This gives us the 3D coordinate $\mathbf{X_b}$ of the bottom landmark. According to constraint 3, we search for the optimal 3D coordinate $\mathbf{X_t}$ of the top landmark along its line of sight using gradient descent, such that the direction of $\mathbf{X_t} - \mathbf{X_b}$ has the best match with $\tilde{\mathbf{n}}_\mathbf{v}$ (which equals $\tilde{\mathbf{n}}_\mathbf{g}$ for pedestrians or $\tilde{\mathbf{n}}_\mathbf{p}$ for cars). The length of $\mathbf{X_t} - \mathbf{X_b}$ is checked against the prior knowledge of the real-world height H of the top landmark. Denote the angle between $\tilde{\mathbf{n}}_\mathbf{v}$ and the direction of $\mathbf{X_t} - \mathbf{X_b}$ as δ, then the resulting compatibility score is

$$s_{g2} = \exp\{-\frac{\delta^2}{2\sigma^2}\} \cdot \exp\{-\frac{(\|\mathbf{X_t} - \mathbf{X_b}\| - H_0)^2}{2\sigma_H^2}\} - 0.5, \quad (12)$$

where σ_H^2 is the variance of H according to prior knowledge, and σ is a parameter set as 20 degrees.

The final compatibility score s_g for an individual object candidate is the average of s_{g1} and s_{g2}.

Individual surface candidate: The compatibility of a surface candidate also comes from two sources. Firstly, we check how well the vertical orientation distribution produced by the surface candidate agree with the current hypothesis $\tilde{\mathbf{n}}_\mathbf{g}$ (for vertical surface) or $\tilde{\mathbf{n}}_\mathbf{p}$ (for horizontal surface). This produces a compatibility score s_{g1} with range between -0.5 and 0.5. Secondly, we check the plausibility of the location of the surface candidate with respect to the ground horizon in the image. For the horizontal surface candidate, denote the proportion of the surface region *above* the ground horizon as r_h, then the compatibility score s_{g2} is $-r_h$ with range between -1 and 0. For the vertical surface candidate, this type of compatibility does not apply, as it usually straddles across the horizon. The final compatibility score s_g is the average of s_{g1} and s_{g2} for the horizontal surface candidate, and is s_{g1} for the vertical surface candidate.

4.2. Local geometric compatibility

Local geometric compatibility refers to the compatibility between nearby object candidates. Inspired by [10] (yet in

a totally different setting), we examine two types of local geometric compatibility. Firstly, if the bounding boxes of two candidates i and j overlap and the bounding box of the farther candidate j is located mostly or completely within the bounding box of the closer candidate i, then they are unlikely to co-exist due to the occlusion conflict resulting from the depth ordering, as is illustrated by the lower pair of orange cubes in Figure 8. Therefore, we define the pairwise compatibility score $s_{ij}^{(dep)}$ related to depth ordering as

$$s_{ij}^{(dep)} = -(|R_{ij}|/|R_j|)^{\lambda}, \tag{13}$$

where $|R_{ij}|$ is the overlapping area of candidates i and j, $|R_j|$ is the area of candidate j, and λ is a parameter set as 5.

Secondly, if the footprints of the two object candidates on the ground plane significantly overlap, they are unlikely to co-exist due to space occupancy conflict, as is illustrated by the upper pair of orange cubes in Figure 8. The pairwise compatibility score $s_{ij}^{(ocp)}$ related to space occupancy is therefore defined as

$$s_{ij}^{(ocp)} = -|R_i \cap R_j|/|R_i \cup R_j|, \tag{14}$$

where $|R_i \cap R_j|$ and $|R_i \cup R_j|$ are the intersection and union areas of the footprints of candidates i and j, respectively. The footprint of an object candidate is obtained by mapping several ground-touching landmarks in the image to the ground plane. Those landmarks are estimated by a HDR regressor.

4.3. Inferring candidate validity with CRF

We construct a CRF over the object/surface candidates to infer their validity. Each candidate forms a node, and two object candidates have an edge between them if their bounding boxes and/or footprints overlap. The objective function that the CRF attempts to maximize is

$$V(\mathbf{o}) = \sum_{k=1}^{2} \omega_k^{(s)}(o_k^{(s)}) + \sum_i \omega_i(o_i) + \sum_{(i,j)\in\mathbf{E}} \varphi_{ij}(o_i, o_j). \tag{15}$$

Here, o is the binary validity indicator. $\omega_1^{(s)}(o_1^{(s)})$, $\omega_2^{(s)}(o_2^{(s)})$, and $\omega_i(o_i)$ are the unary potentials for the vertical surface candidate, horizontal surface candidate, and object candidate i, respectively. Their values are defined as $\omega(o=1) = s_g + (s_d - 0.5)$ and $\omega(o=0) = 0$, where s_g is the compatibility score defined in the previous section, and s_d is the segmentation confidence or detection confidence returned from the surface segmentation algorithm or object detection algorithm. $\varphi_{ij}(o_i, o_j)$ is the pairwise potential between object candidates i and j. Its value is s_{ij} when both o_i and o_j are 1; otherwise its value is 0. Here, s_{ij} could either be $s_{ij}^{(dep)}$ or $s_{ij}^{(ocp)}$ depending on the type of the edge. If both of them exist, then s_{ij} is the smaller of the two.

After the inference is complete, the quality of the current global 3D geometry hypothesis is the maximum value V^* of the objective function. After all the hypotheses are evaluated, the one with the highest quality is selected. This optimal hypothesis is further refined by the valid object/surface candidates associated with it.

5. Experiment

We evaluate our approach on the test dataset compiled by Hoiem *et al.* [14] which contains 422 random outdoor images from the LableMe dataset [17]. Those images cover a multitude of outdoor urban scenes and include a wide variety of object pose and size, making the dataset very challenging. The dataset contains 923 cars and 720 pedestrians in total.

Hoiem *et al.* also collects a training dataset containing 51 images. We use this set to train our pose and landmark regressors. The prior distribution of pedestrian height in our experiment follows $N(H_p; 1.7, 0.09)$, and the prior distribution of wheel height is $N(H_w; 0.6, 0.25)$. When proposing hypotheses from distributions, 50 random mixtures are usually enough, and the total number of hypotheses to evaluate is in the hundreds. We run our algorithm under multiple focal lengths and the one that yields the highest value of V^* is adopted. It takes less than 10 minutes to process a 640-by-480 image with Matlab code.

Comparison with the state of the art: Using the top-notch Deformable Part Model (DPM) [8] as the baseline detector, we compare the object detection performance of our approach with Hoiem's algorithm in [14]. The result of Hoiem's algorithm is generated by running their published codes with the DPM detector outputs. The ROC curves are plotted in Figure 9a. The average precision (AP) of our approach is 50.5%, achieving a boost of more than 10% over the AP of the baseline detector at 40.1%. Surprisingly, Hoiem's algorithm performs worse than the baseline in the realm of lower false positive rates, yielding an AP of 30.8%. We observe that Hoiem's model is not effective for car candidates returned by the DPM detector. This is probably because Hoiem's algorithm takes the bounding box height as the object height in the image. This approximation is poor when a non-planar object, such as a car, is viewed from a non-zero pitch angle. As our algorithm explicitly estimates the landmark locations in the image, it does not have this problem.

In addition to comparing object detection, we also evaluate global geometry estimation. The dataset provides the ground truth of horizon in the form of the row index where the horizon is located. It does not distinguish between gravity and ground horizons, since the two are almost the same for most of the images in the dataset. After converting the row index of a horizon to the corresponding orientation vector, we compute the error of an estimated gravity direction (or ground plane orientation) by measuring the angle between it and the ground truth orientation vector. The re-

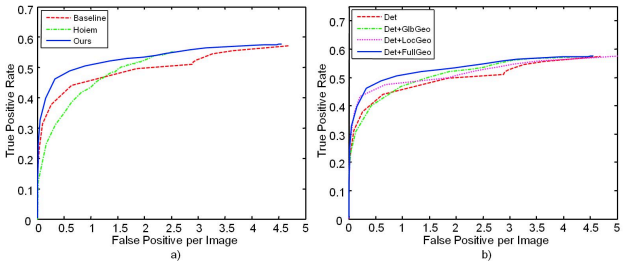

Figure 9. Comparison of object detection performance. a) Comparison with state of the art. The baseline detector is DPM [6]. Our algorithm significantly boosts the performance over the baseline detector. It also outperforms Hoiem's algorithm [14] while making fewer assumptions. b) Contribution of individual components. Here, "Det" shows the result of the DPM baseline detector; "Det+GlbGeo" shows the result of including global geometric context alone; "Det+LocGeo" shows the result of including local geometric context alone; "Det+FullGeo" shows the result of our full system using both types of context.

sults are shown in Figure 10. Despite not using any prior, our method has a smaller error in horizon estimation than Hoiem's algorithm. The estimated ground plane height by our method is centered around 1.5 meters, close to the typical eye level.

It is worth noting that, unlike Hoiem's algorithm, we do not assume the ground plane is perpendicular to the gravity direction and has zero roll and small pitch. Even with a greater flexibility, our approach still outperforms Hoiem's algorithm both in object detection and global geometry estimation, on the test dataset that largely satisfies those assumptions.

As we do not have access to the codes of the algorithms proposed in [3] or [18], we are not able to directly compare with their performance. Yet according to what the authors report in [3], their method does not perform as well as Hoiem's algorithm on a subset of the 422-image test dataset. In [18], the authors use their own baseline detector and a different subset of images. As a result, we could only compare the performance gain over the baseline. The algorithm in [18] achieves a gain of 5.1% in average precision, while our approach achieves a gain of 10.4%.

Global and local 3D geometric context: Different from existing works, we use both global and local 3D geometric context when inferring the validity of object candidates. The benefit of doing so can be seen in Figure 9b. Both the global and local 3D geometric context enhance detection performance, and the highest gain is achieved when they are applied simultaneously.

Benefit is mutual: Not only does 3D geometric context enhance object detection performance, but coherent object detection in turn improves the estimation of gravity and ground horizons. To verify this argument, we estimate the gravity direction and ground plane orientation from vertical

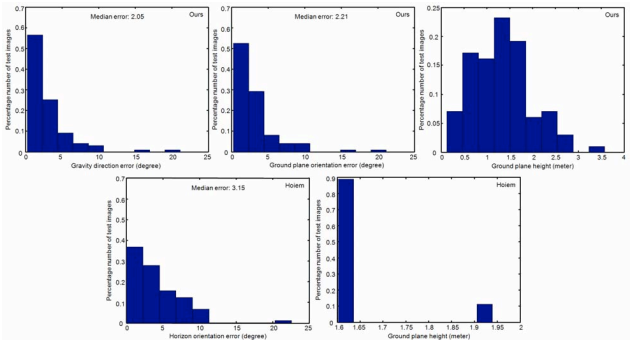

Figure 10. Comparison of global 3D geometry estimation performance. The first row shows the distributions of gravity direction error, ground orientation error, and ground height from our algorithm. The second row shows the results of Hoiem's algorithm [14]. Our algorithm has a smaller error in horizon estimation. We are not able to compute the error of the ground plane height estimation due to the lack of ground truth. However, both the algorithms peak at around 1.5 - 1.6 meters, roughly corresponding to the eye level. Best view on screen to zoom in.

and horizontal surfaces alone. The median estimation error is 2.62 degrees for gravity direction and 4.85 degrees for ground plane orientation. By contrast, the errors of our full system are 2.05 and 2.21 degrees, respectively.

Qualitative evaluation: Several examples of the object detection results of our algorithm are shown in Figure 11. Please refer to the figure caption for the meanings of different types of boxes and some discussions about the results.

6. Conclusion

We have presented an object detection algorithm that ensures geometric coherence in the 3D world. Compared with existing approaches, the major contributions of our work include 1) a more flexible modeling of the scene that treats gravity direction and ground orientation separately, 2) a more systematic representation of the geometric relationships between scene and objects, 3) a generalized RANSAC algorithm that enables both outlier suppression and noise reduction in hypothesis generation, 4) incorporating both global and local 3D geometric context with a CRF, 5) including surface regions in estimating and evaluating global 3D geometry. Due to these factors, our algorithm achieves a superior performance on a challenging dataset. Future work would focus on 3D geometric constraints between surfaces and objects.

References

[1] D. G. Aguilera, J. G. Lahoz, and J. F. Codes. A new method for vanishing points detection in 3d reconstruction from a single view. *ISPRS*, 2005.

[2] C. Atanasoaei, C. McCool, and S. Marcel. A principled approach to remove false alarms by modeling the context of a face detector. *BMVC*, 2010.

Figure 11. Examples of our coherent object detection results. **Solid green box**: the object is detected by both the DPM detector and our algorithm. **Solid red box**: the object is missed by the DPM detector but recovered by our algorithm. **Dotted red box**: the object is detected by the DPM detector but rejected by our algorithm. **Magenta line**: gravity horizon. **Yellow grid**: ground plane where the grid spacing is 1m. The detection threshold of both the DPM detector and our algorithm is set as 0.5. We can see that some false detections from DPM are rejected due to inconsistency with global geometry (*e.g.* the huge "pedestrian" in (a)); some false detections from DPM are rejected due to inconsistency with local geometry (*e.g.* the rejected "cars" in (c)). Also, some true objects missed by DPM are recovered due to their geometric consistency (*e.g.* the pedestrians in (b)). The case when gravity direction and ground orientation do not agree is shown in (f), where the ground horizon is illustrated by the thick yellow line. It deviates from the gravity horizon. A failure case is shown in (k), where a "car" is mistakenly recovered due to its high geometric compatibility. In another failure case shown in (l), two truncated cars are wrongly rejected because our regressors are not trained on truncated cars.

[3] S. Y. Bao, M. Sun, and S. Savarese. Toward coherent object detection and scene layout understanding. *Image and Vision Computing*, 2011.

[4] M. J. Choi, A. Torralba, and A. S. Willsky. A tree-based context model for object recognition. *PAMI*, 2012.

[5] D. Comaniciu and P. Meer. Mean shift: a robust approach toward feature space analysis. *PAMI*, 2002.

[6] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. *CVPR*, 2005.

[7] S. K. Divvala, D. Hoiem, J. H. Hays, A. A. Efros, and M. Hebert. An empirical study of context in object detection. *CVPR*, 2009.

[8] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *PAMI*, 32(9):1627–1645, 2010.

[9] R. B. Girshick, P. F. Felzenszwalb, and D. McAllester. Discriminatively trained deformable part models, release 5. http://people.cs.uchicago.edu/ rbg/latent-release5/.

[10] A. Gupta, A. A. Efros, and M. Hebert. Blocks world revisited: image understanding using qualitative geometry and mechanics. *ECCV*, 2010.

[11] G. Heitz and D. Koller. Learning spatial context: using stuff to find things. *ECCV*, 2008.

[12] D. Hoiem, A. A. Efros, and M. Hebert. Code for recovering surface layout from an image. http://www.cs.illinois.edu/homes/dhoiem/.

[13] D. Hoiem, A. A. Efros, and M. Hebert. Recovering surface layout from an image. *IJCV*, 2007.

[14] D. Hoiem, A. A. Efros, and M. Hebert. Putting objects in perspective. *IJCV*, 2008.

[15] W. S. Hwang and J. Weng. Hierarchical discriminant regression. *PAMI*, 2000.

[16] J. Lafferty, A. McCallum, and F. Pereirac. Conditional random fields: probabilistic models for segmenting and labeling sequence data. *ICML*, 2001.

[17] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman. Labelme: a database and web-based tool for image annotation. *Technical report, MIT*, 2005.

[18] M. Sun, S. Y. Bao, and S. Savarese. Object detection with geometrical context feedback loop. *BMVC*, 2010.

[19] W. S. Zheng, S. Gong, and T. Xiang. Quantifying and transferring contextual information in object detection. *PAMI*, 2012.