PROCEEDINGS OF THE 2006 IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION
VOLUME II, PAGES 2410–2417, JUNE 2006

Depth from Familiar Objects: A Hierarchical Model for 3D Scenes

Erik B. Sudderth, Antonio Torralba, William T. Freeman, and Alan S. Willsky
Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology
sudderth@alum.mit.edu, torralba@csail.mit.edu, billf@mit.edu, willsky@mit.edu

Abstract

We develop an integrated, probabilistic model for the appearance and three-dimensional geometry of cluttered scenes. Object categories are modeled via distributions over the 3D location and appearance of visual features. Uncertainty in the number of object instances depicted in a particular image is then achieved via a transformed Dirichlet process. In contrast with image-based approaches to object recognition, we model scale variations as the perspective projection of objects in different 3D poses. To calibrate the underlying geometry, we incorporate binocular stereo images into the training process. A robust likelihood model accounts for outliers in matched stereo features, allowing effective learning of 3D object structure from partial 2D segmentations. Applied to a dataset of office scenes, our model detects objects at multiple scales via a coarse reconstruction of the corresponding 3D geometry.

1. Introduction

Detailed geometric models have played an important role in the design of methods for the detection of particular objects in cluttered scenes. However, most algorithms for generic object categorization use a simple 2D pixel representation. In discriminative approaches, scale invariance is often achieved by resizing an image in small steps, and detecting objects via a "sliding window". Alternatively, some part–based models include variables which account for global scaling of expected feature distances [3], or local affine warpings of feature templates [7]. Other methods discard geometry entirely following an initial stage of feature extraction [1, 13]. In all cases, scale invariance is based on transformations of the observed pixels or low–level features, and underlying 3D structure is ignored.

While a purely image based approach to object recognition is sometimes adequate, many applications require more explicit knowledge about the 3D world. For example, if robots are to navigate in complex environments and manipulate objects, they require more than a flat segmentation of the image pixels into object categories. Motivated by these challenges, we instead cast object recognition as a 3D problem, and develop methods which partition estimated 3D structure into object categories.

A few recent models ignore objects, learning direct mappings from images to 3D geometry [5, 12, 17]. However, knowledge of the objects present in a scene provides information about their expected 3D shape, regularizing the often ambiguous depth estimates produced by low–level features. In addition, geometry provides important cues for object recognition. To exploit these relationships, we use binocular stereo training images to train an approximately calibrated model of multiple objects' 3D geometry. Using this model, we achieve scale invariant object recognition via translations of 3D objects, rather than image transformations. Because we consider objects with predictable 3D structure, we also automatically recover a coarse reconstruction of the underlying scene depths.

Rather than learning classifiers for isolated objects, we propose a hierarchical, probabilistic model of multiple object scenes [14, 18]. Our approach extends an earlier 2D scene model based on the *transformed Dirichlet process* (TDP) [15]. Dirichlet processes are a flexible tool from nonparametric Bayesian statistics [2, 11, 16], which we use to allow uncertainty in the *number* of object instances depicted in each image. Generalizing [15], we automatically learn part–based descriptions of an *a priori* unknown set of visual categories. Previous TDP models also assembled 2D object models in a "jigsaw puzzle" fashion, and thus assumed images were normalized to a common scale. Extending the TDP to 3D scenes, we propose a robust stereo likelihood which captures ambiguities in low–level feature matching. We then develop Monte Carlo methods which learn 3D object models from partial stereo segmentations, and estimate 3D scene structure from monocular images.

We begin in Sec. 2 by introducing our feature representation, and formulate a robust stereo likelihood function. Sec. 3 then uses the TDP to develop a generative, part–based model for 3D feature locations and appearance. In Sec. 4, we describe a blocked Gibbs sampler which learns scene geometry from labeled stereo training images. We conclude in Sec. 5 with results on office scenes.

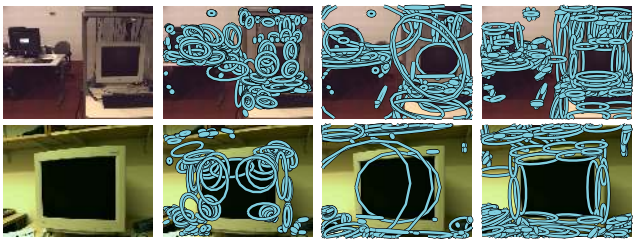

Figure 1. Three types of interest operators applied to two office scenes: Harris–affine corners (left), maximally stable extremal regions (center), and clustered Canny edges (right).

2. Features and Geometry

In Sec. 2.1, we first motivate the three types of features used to represent training and test images. We then formulate the imaging geometry underlying our depth estimates (Sec. 2.2), and develop a robust likelihood function which accounts for outliers in matched stereo features (Sec. 2.3).

2.1. Feature Extraction and Representation

Each training or test image is first converted to a set of regions of interest using three detectors. Harris–affine corners extract local maxima of orientation energy [10], while maximally stable extremal regions are derived from a watershed segmentation algorithm [9].[1] We also build a set of straight edge features by dividing Canny edge sequences at points of high curvature. Fig. 1 illustrates these three regions, showing that they favor complementary aspects of object appearance. In particular, edge features play a critical role in modeling textureless objects like desks.

Following several recent papers [1, 13, 14], we use SIFT descriptors [8] (normalized histograms of orientation energy) to describe the appearance of each region of interest. Interest points provide a lower dimensional representation which focuses on the most repeatable and salient regions, while SIFT descriptors provide some lighting and viewpoint invariance. To facilitate learning, we use K–means clustering to vector quantize the SIFT descriptors from the training set, producing a dictionary of appearance "words." As in [15], we also coarsely encode region shape by dividing features into three groups (circular, horizontal, or vertical) prior to quantization. The i^{th} feature in image j is then described by the closest appearance descriptor w_{ji}, and 2D pixel coordinates (v^x_{ji}, v^y_{ji}). Our current model neglects the scale at which features are detected.

2.2. Binocular Stereo Matching

Let $u = (u^x, u^y, u^z)$ denote the world coordinates of a 3D point, where the z-axis has been chosen to align with the camera's optical axis. Then, indexing pixels (v^x, v^y) from the optical center, the perspective projection of u equals

$$v^x = \xi \frac{u^x}{u^z} \qquad v^y = \xi \frac{u^y}{u^z} \tag{1}$$

where ξ denotes the magnification, in pixels, corresponding to the camera's focal length. Other coordinate systems are easily accomodated by appropriate transformations.

Training images used in this paper are captured by a calibrated stereo camera (the MEGA-D, by Videre Design). As in recent approaches to sparse wide baseline stereo [9], we begin by extracting regions of interest in both the left and right images. For each interest point in the reference (left) image, we then search for the best matching regions along the corresponding epipolar line (see Fig. 2). Match quality is measured via the Euclidean distance between SIFT descriptors [8]. Let v^d denote the disparity, in pixels, corresponding to a candidate pair of matching features. Each match corresponds to some set of world coordinates:

$$u^x = \frac{v^x}{\xi} u^z \qquad u^y = \frac{v^y}{\xi} u^z \qquad u^z = \frac{\xi D}{v^d} \tag{2}$$

Here, D is the baseline distance between cameras (in our case, 89 mm). Note that we have written the world u^x and u^y in terms of the unknown depth u^z, rather than the disparity v^d. This form emphasizes our primary interest in the underlying 3D geometry, and is more easily incorporated with our generative model of scene structure (see Sec. 4).

2.3. Robust Disparity Likelihoods

Because we represent images by a sparse set of interest regions, we must only estimate scene depths at these points. While this problem is simpler than the estimation of dense depth maps, it is still ill–posed based solely on local feature correspondences. In particular, repetitive scene structures and occlusion effects near object boundaries often lead to inaccurate disparity estimates for some features. In Fig. 2, we illustrate the noisy depth estimates produced by local matching in stereo images of office scenes. Wide baseline stereo algorithms typically employ a geometric validation stage to discard such outliers [9]. This approach would work poorly for our application, however, because features near object boundaries are often the most informative for recognition tasks. We instead propose a probabilistic model which robustly converts approximate disparity matches to depth distributions. The learning algorithms developed in Sec. 4 then use geometric models of objects to impose a scene structure which resolves local ambiguities.

Consider a feature which has candidate stereo matches at C different disparities $\{\bar{v}^d_c\}_{c=1}^C$, and let $\bar{v}^s_c$ denote the matching score (distance between SIFT descriptors) for $\bar{v}^d_c$. Features with no matches induce an uniformative likelihood on the underlying scene depth u^z. Otherwise, at most one match can correspond to the true scene depth, and the others must be outliers. Let a be an unobserved random variable

[1] Software provided by the Oxford University Visual Geometry Group: http://www.robots.ox.ac.uk/~vgg/research/affine/

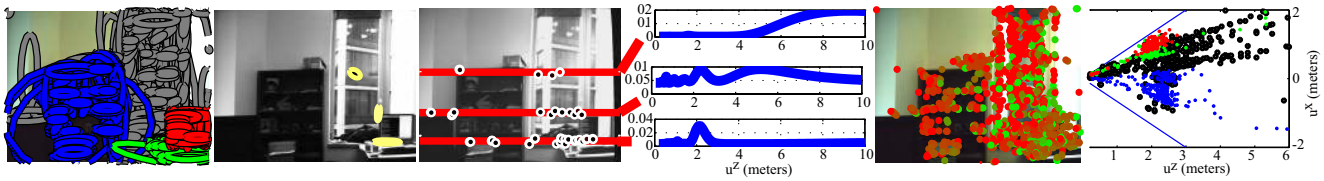

Figure 2. An office scene depicting a computer screen, desk, and bookshelf (color–coded, left). For three features, we show matches along epipolar lines in the right stereo image, and corresponding depth likelihoods. Depth estimates (right) are independently chosen for each feature. In the frontal view, close features are green and distant red. The overhead view colors features like their associated object.

indicating which of the C matches is *not* an outlier, and take $a = 0$ if all matches are outliers. Neglecting possible correlations due to scene structure, we assume that inlier and outlier matches are independently sampled as

$$p\left(\{\bar{v}_c^d, \bar{v}_c^s\}_{c=1}^C \mid u^z\right) = \sum_{a=0}^{C} p\left(\{\bar{v}_c^d, \bar{v}_c^s\}_{c=1}^C, a \mid u^z\right)$$
$$\propto \sum_{a=0}^{C} p(a) \prod_{c=1}^{C} p\left(\bar{v}_c^d \mid a, u^z\right) p\left(\bar{v}_c^s \mid a\right) \quad (3)$$

Let ϵ denote the prior probability that all observations are outliers ($a = 0$), so that all other outlier hypotheses have equal probability $(1 - \epsilon)/C$. We assume that correct matches are corrupted by Gaussian noise, while outlier disparities are sampled uniformly over a range determined by the camera geometry:

$$p\left(\bar{v}_c^d \mid a = c, u^z\right) = \mathcal{N}\left(\bar{v}_c^d; \frac{\xi D}{u^z}, \sigma_d^2\right) \quad (4)$$
$$p\left(\bar{v}_c^d \mid a \neq c, u^z\right) = \mathcal{U}\left(\bar{v}_c^d; d_{\min}, d_{\max}\right) \quad (5)$$

We also assign the inlier and outlier matching scores $\bar{v}_c^s$ log–normal densities with differing mean and variance.

To estimate the parameters of this likelihood function, we collected disparity matches for 16,000 monitor and bookshelf features from the stereo training images used in Sec. 5. Because each selected object was approximately orthogonal to the optical axis, the median depth of each instance's raw stereo matches provides an accurate estimate of true depth for all features. We may then compute maximum likelihood parameter estimates by extending standard EM updates for mixture models [4]. The E–step averages over possible outlier hypotheses a, producing a lower bound on the likelihood which is maximized in the M–step.

From our training set, we estimated the probability that all matches are outliers to be $\epsilon = 0.22$, and the noise level for correct matches as $\sigma_d = 2.4$ pixels. Fig. 2 illustrates depth likelihoods correponding to three sample features. Intuitively, matches with small disparities lead to greater depth uncertainty, due to the inversion induced by the perspective projection of eq. (2). When there are many conflicting matches, the likelihood becomes uniform.

3. Hierarchical Models for 3D Scenes

In this section, we develop a transformed Dirichlet process (TDP) model for multiple object scenes. We begin by reviewing Dirichlet processes (Sec. 3.1), and previous work describing spatial data via transformations (Sec. 3.2). In Sec. 3.3, we extend these approaches to learn part–based models of 3D object structure and appearance.

3.1. Counting with Dirichlet Processes

Graphical models describe the statistical structure of a fixed set of random variables. Many machine vision applications, however, must deal with uncertainty in the *number* of parts composing a particular object, or objects present in a particular scene. We address this issue using the *Dirichlet process* (DP), a flexible nonparametric prior which has been widely applied in Bayesian statistics [2, 11].

Consider a collection of spatial data, such as 3D points extracted from a visual scene. Let $\theta = (\mu, \Lambda)$ denote the mean and covariance parameters of a Gaussian distribution, and H be a prior measure on the space of Gaussian distributions Θ. A Dirichlet process with concentration parameter γ, denoted by $\mathrm{DP}(\gamma, H)$, then defines a prior distribution over *infinite* Gaussian mixtures $G \sim \mathrm{DP}(\gamma, H)$:

$$G(\theta) = \sum_{\ell=1}^{\infty} \beta_\ell \delta(\theta, \theta_\ell) \qquad \theta_\ell \sim H \quad (6)$$

$$\beta_\ell = \beta'_\ell \prod_{m=1}^{\ell-1} (1 - \beta'_m) \qquad \beta'_m \sim \mathrm{Beta}(1, \gamma) \quad (7)$$

This *stick–breaking construction* [16] defines the mixture weights $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots)$ using beta random variables. We use $\boldsymbol{\beta} \sim \mathrm{GEM}(\gamma)$ to denote this process. For moderate concentrations γ, all but a small random subset of the mixture weights will be nearly zero.

Given $G \sim \mathrm{DP}(\gamma, H)$, each observation u_i is generated by independently sampling mean and covariance parameters $\bar{\theta}_i \sim G$, and then choosing $u_i \sim \mathcal{N}(\bar{\theta}_i)$ from the corresponding Gaussian:

$$p\left(u_i \mid \boldsymbol{\beta}, \theta_1, \theta_2, \ldots\right) = \sum_{\ell=1}^{\infty} \beta_\ell \mathcal{N}(u_i \mid \theta_\ell) \quad (8)$$

More generally, Θ could parameterize any family with a complementary prior measure H. The stick–breaking pro-

cess induces a clustering bias, and leads to efficient Monte Carlo methods which automatically learn the number of clusters underlying a particular set of observations [2, 11].

3.2. Transformed Dirichlet Processes

As we demonstrate later, the DP mixture of eq. (8) leads to effective part–based models for the internal geometry of rigid objects. More generally, we expect multiple object scenes to share local features, but differ significantly in their global spatial structure. The *hierarchical Dirichlet process* (HDP) [16] was developed to address the related problem of partially sharing topics among text documents. Applied to spatial data, the HDP chooses a globally shared mixture $G_0 \sim \mathrm{DP}(\gamma, H)$ as in eqs. (6, 7). Each image is then assigned a mixture $G_j \sim \mathrm{DP}(\alpha, G_0)$, reusing the same Gaussian clusters θ_ℓ in different proportions:

$$G_j(\theta) = \sum_{\ell=1}^{\infty} \pi_{j\ell}\delta(\theta, \theta_\ell) \qquad \boldsymbol{\pi}_j \sim \mathrm{DP}(\alpha, \boldsymbol{\beta}) \tag{9}$$

The global mixture weights $\boldsymbol{\beta}$ determine the expected cluster proportions $\boldsymbol{\pi}_j$, while α specifies the variability from image to image. This construction assumes that images differ only in the proportion of observed features for each spatial cluster, rather than the location and shape of those clusters. Because objects are not observed in consistent locations relative to the camera, a standard HDP would thus not adequately generalize to novel visual scenes.

Motivated by these difficulties, we consider a family of *transformations* $\tau(\theta; \rho)$ of the global mixture components θ, indexed by ρ. For spatial data, we associate these transforms [6] with shifts of the mean location of global clusters:

$$\tau(\mu, \Lambda; \rho) = (\mu + \rho, \Lambda) \tag{10}$$

The *transformed Dirichlet process* (TDP) [15] generalizes the HDP to more flexibly share spatial structure among images. The TDP is derived from *distributions over transformations* $q(\rho \mid \phi)$, indexed by $\phi \in \Phi$. Let R denote a prior measure on the space of transformation distributions Φ, which we later constrain to be zero–mean Gaussians.

We begin by augmenting the Dirichlet process stick–breaking construction of eq. (7) to define a global measure describing both parameters θ and transformations ρ:

$$G_0(\theta, \rho) = \sum_{\ell=1}^{\infty} \beta_\ell \delta(\theta, \theta_\ell) q(\rho \mid \phi_\ell) \qquad \begin{matrix} \theta_\ell \sim H \\ \phi_\ell \sim R \end{matrix} \tag{11}$$

As before, $\boldsymbol{\beta} \sim \mathrm{GEM}(\gamma)$. Note that each cluster θ_ℓ has a different transformation distribution $q(\rho \mid \phi_\ell)$. We then independently sample $G_j \sim \mathrm{DP}(\alpha, G_0)$ for each image. Because samples from DPs are discrete with probability one, this joint measure can be written as

$$G_j(\theta, \rho) = \sum_{\ell=1}^{\infty} \pi_{j\ell}\delta(\theta, \theta_\ell) \left[\sum_{t=1}^{\infty} \lambda_{j\ell t}\delta(\rho, \rho_{j\ell t}) \right] \tag{12}$$

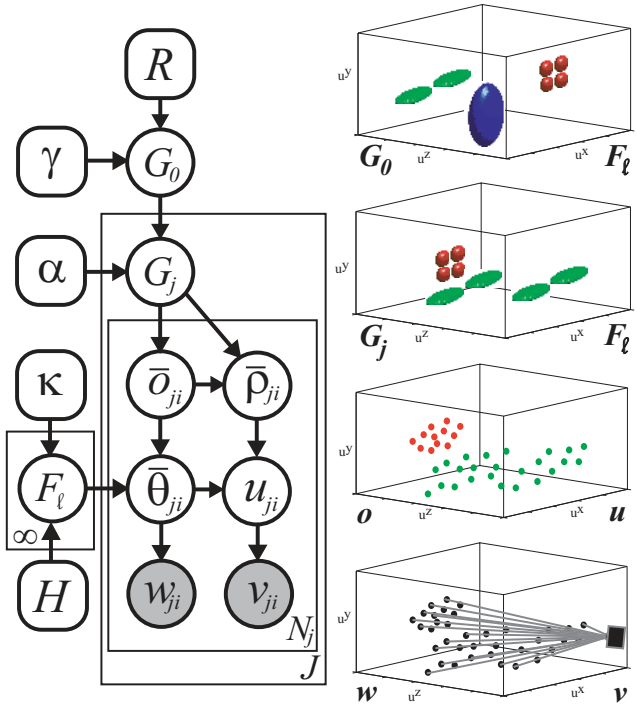

Figure 3. TDP model for 3D scenes (left), and cartoon illustration of the generative process (right). Global mixture G_0 describes the expected frequency and location of visual categories, whose internal structure is represented by part–based appearance models $\{F_\ell\}_{\ell=1}^{\infty}$. Each image mixture G_j instantiates a randomly chosen set of objects at transformed locations ρ. 3D feature positions u_{ji} are sampled from transformed parameters $\tau(\bar{\theta}_{ji}; \bar{\rho}_{ji})$ corresponding to parts of object $\bar{o}_{ji}$. The camera observes projections v_{ji} of these features, with part–dependent appearance w_{ji}.

where $\sum_t \lambda_{j\ell t} = 1$. As $G_0(\theta, \rho)$ only has support at a discrete set of cluster parameters, $G_j(\theta, \rho)$ will associate many different transformations $\rho_{j\ell t}$ with each distinct θ_ℓ.

In the simplest case, each 3D feature in image j is now generated by sampling $(\bar{\theta}_{ji}, \bar{\rho}_{ji}) \sim G_j$, and then choosing $u_{ji} \sim \mathcal{N}(\tau(\bar{\theta}_{ji}; \bar{\rho}_{ji}))$ from a transformed Gaussian [15]. Intuitively, global mixture components θ_ℓ define object geometry in a "canonical" coordinate frame, while the random set of transformations ρ determine the object instances within a particular scene. Critically, the TDP allows uncertainty in the *number* of objects depicted by each image. For instance, in the toy example of Fig. 3, the green object appears twice, while the blue does not appear at all.

3.3. Part–Based Object Appearance Models

Applied directly, the TDP model of eqs. (11, 12) describes the geometry of each global object cluster by a single Gaussian. This representation poorly captures the complex structure of many real objects, and does not model local variations in object appearance. In this section, we show how Dirichlet processes may also be adapted to learn richer,

part–based descriptions of visual categories.

Extending the DP mixture of eq. (8), we associate parts with clusters of features that have distinctive appearances w_{ji} and 3D positions u_{ji}. Each part $\theta_{\ell k} = (\eta_{\ell k}, \mu_{\ell k}, \Lambda_{\ell k})$ of object ℓ is then defined by a Gaussian position distribution $\mathcal{N}(\mu_{\ell k}, \Lambda_{\ell k})$, and a multinomial appearance distribution $\eta_{\ell k}$. Letting H denote a prior measure on part parameters $\theta_{\ell k} \in \Theta$, we take $F_\ell \sim \mathrm{DP}(\kappa, H)$ as the potentially infinite set of parts underlying the ℓ^{th} category:

$$F_\ell(\theta) = \sum_{k=1}^{\infty} \varepsilon_{\ell k} \delta(\theta, \theta_{\ell k}) \qquad \begin{array}{l} \boldsymbol{\varepsilon}_\ell \sim \mathrm{GEM}(\kappa) \\ \theta_{\ell k} \sim H \end{array} \tag{13}$$

Generalizing the earlier TDP construction, we allow infinitely many potential visual object categories o, and define a prior on their probabilities and associated transformations:

$$G_0(o, \rho) = \sum_{\ell=1}^{\infty} \beta_\ell \delta(o, \ell) q(\rho \mid \phi_\ell) \qquad \begin{array}{l} \boldsymbol{\beta} \sim \mathrm{GEM}(\gamma) \\ \phi_\ell \sim R \end{array} \tag{14}$$

Finally, each image is based on a set of randomly transformed objects $G_j \sim \mathrm{DP}(\alpha, G_0)$, analogously to eq. (12).

As illustrated in Fig. 3, given these infinite discrete measures, the i^{th} feature in image j is independently sampled in three stages. First, a visual category $\bar{o}_{ji}$ and transformation $\bar{\rho}_{ji}$ are chosen from G_j, selecting a particular object instance. Second, parameters $(\bar{\eta}_{ji}, \bar{\mu}_{ji}, \bar{\Lambda}_{ji}) = \bar{\theta}_{ji} \sim F_{\bar{o}_{ji}}$ corresponding to one of that objects' parts are selected, and a 3D feature position $u_{ji} \sim \mathcal{N}(\tau(\bar{\mu}_{ji}, \bar{\Lambda}_{ji}; \bar{\rho}_{ji}))$ sampled relative to that instance's position. Finally, we observe a 2D feature with appearance $w_{ji} \sim \bar{\eta}_{ji}$, and position v_{ji} determined by the perspective projection of eq. (1).

The hierarchical, TDP scene model of Fig. 3 employs three different stick–breaking processes, allowing uncertainty in the number of visual categories ($\mathrm{GEM}(\gamma)$), parts composing each category ($\mathrm{GEM}(\kappa)$), and object instances depicted in each image ($\mathrm{GEM}(\alpha)$). It thus generalizes the parametric model of [14], which assumed fixed, known sets of parts and objects. In the limit as $\kappa \to 0$, each category uses a single part, and we recover a 3D extension of [15].

4. Learning Object Geometry and Appearance

We now describe a Gibbs sampling algorithm for learning our 3D TDP scene model's parameters from training data, extending related methods developed for other Dirichlet process models [2, 11, 15, 16]. For each observed feature (w_{ji}, v_{ji}), we resample the corresponding 3D depth u^z_{ji}, as well as the assignments (t_{ji}, k_{ji}) of that feature to object instances t and parts k. Then, for each instance t in image j, we jointly resample assignments o_{jt} to visual categories with corresponding transformations ρ_{jt} and part assignments $\{k_{ji} \mid t_{ji} = t\}$. Iterating these steps, we approximately sample from the model's posterior distribution over scene interpretations, simultaneously recognizing objects and reconstructing 3D geometry. For simplicity, Sec. 4.1 and 4.2 assume fixed values for the parameters of parts $\boldsymbol{\theta}$ and transformations $\boldsymbol{\phi}$; we discuss their learning in Sec. 4.3.

4.1. Inferring Feature Depths

Intuitively, the most likely depth u^z_{ji} for a particular feature is strongly dependent on the 3D object instance t_{ji}, and corresponding part k_{ji}, generating that feature. For adequate convergence of the Gibbs sampler, we thus employ blocked sampling updates of $(t_{ji}, k_{ji}, u^z_{ji})$. Let $\mathbf{t}_{\setminus ji}$ denote all instance assignments except t_{ji}, and define $\mathbf{k}_{\setminus ji}$ similarly. The Markov properties of the TDP then imply that

$$\begin{aligned} p\left(t_{ji} = t, k_{ji} = k, u^z_{ji} \mid v_{ji}, w_{ji}, \mathbf{t}_{\setminus ji}, \mathbf{k}_{\setminus ji}, \mathbf{o}, \boldsymbol{\rho}, \boldsymbol{\theta}\right) \\ \propto p\left(t \mid \mathbf{t}_{\setminus ji}\right) p\left(k \mid \mathbf{k}_{\setminus ji}, \mathbf{t}, \mathbf{o}\right) \eta_{o_{jt}k}(w_{ji}) \\ \cdots \times p\left(v^x_{ji}, v^y_{ji}, u^z_{ji} \mid \theta_{o_{jt}k}, \rho_{jt}\right) p\left(v^d_{ji} \mid u^z_{ji}\right) \end{aligned} \tag{15}$$

The first term corresponds to the prior clustering bias of the Dirichlet process [2, 11, 16], which takes a convenient form:

$$p\left(t_{ji} \mid \mathbf{t}_{\setminus ji}\right) \propto \sum_{t=1}^{T_j} N^{-i}_{jt} \delta(t_{ji}, t) + \alpha \delta(t_{ji}, \bar{t}) \tag{16}$$

Here, N^{-i}_{jt} denotes the number of other features currently assigned to each of the T_j existing object instances, and $\bar{t}$ allows for the creation of new object instances. Similarly, the second term encodes the part clustering bias of eq. (13):

$$\begin{aligned} p\left(k_{ji} \mid t_{ji} = t, o_{jt} = \ell, \mathbf{k}_{\setminus ji}, \mathbf{t}_{\setminus ji}, \mathbf{o}_{\setminus jt}\right) \\ \propto \sum_{k=1}^{K_\ell} B^{-i}_{\ell k} \delta(k_{ji}, k) + \kappa \delta(k_{ji}, \bar{k}) \end{aligned} \tag{17}$$

In this case, $B^{-i}_{\ell k}$ denotes the number of other features assigned to the K_ℓ current parts of object ℓ. Computationally, we cache these statistics in a dynamically resized list of instantiated parts [2, 11]. The infinitely many equivalent, unoccupied parts are then tractably represented by $\bar{k}$.

The appearance likelihood $\eta_{o_{jt}k}(w_{ji})$ of eq. (15) is directly determined by the chosen part k_{ji} of the visual category associated with instance t_{ji}. However, the position likelihood for feature (v^x_{ji}, v^y_{ji}) is complicated by the imaging process. In particular, each candidate depth u^z_{ji} selects a different 3D point $\tilde{v} u^z_{ji}$ along a ray $\tilde{v}$ defined by eq. (1). The fourth term of eq. (15) is then the probability that the transformed 3D Gaussian $\mathcal{N}\left(\mu_{o_{jt}k} + \rho_{jt}, \Lambda_{o_{jt}k}\right)$ corresponding to part k of instance t (see Fig. 3) assigns to this point. Letting $\tilde{\mu}_{tk} = \mu_{o_{jt}k} + \rho_{jt}$ denote the 3D position of this part, and conditioning this Gaussian to the projection ray $\tilde{v}$, we recover a *scaled* 1D Gaussian distribution in depth:

$$\begin{aligned} p\left(u^z_{ji} \mid k_{ji} = k, t_{ji} = t, o_{jt} = \ell\right) \propto \omega_{tk} \mathcal{N}\left(u^z_{ji}; \zeta_{tk}, \chi_{tk}\right) \\ \chi^{-1}_{tk} = \tilde{v}^T \Lambda^{-1}_{\ell k} \tilde{v} \qquad \chi^{-1}_{tk} \zeta_{tk} = \tilde{v}^T \Lambda^{-1}_{\ell k} \tilde{\mu}_{tk} \\ \log \omega_{tk} = \frac{1}{2} \log \frac{\chi_{tk}}{|\Lambda_{\ell k}|} - \frac{1}{2} (\tilde{v}\zeta_{tk} - \tilde{\mu}_{tk})^T \Lambda^{-1}_{\ell k} (\tilde{v}\zeta_{tk} - \tilde{\mu}_{tk}) \end{aligned} \tag{18}$$

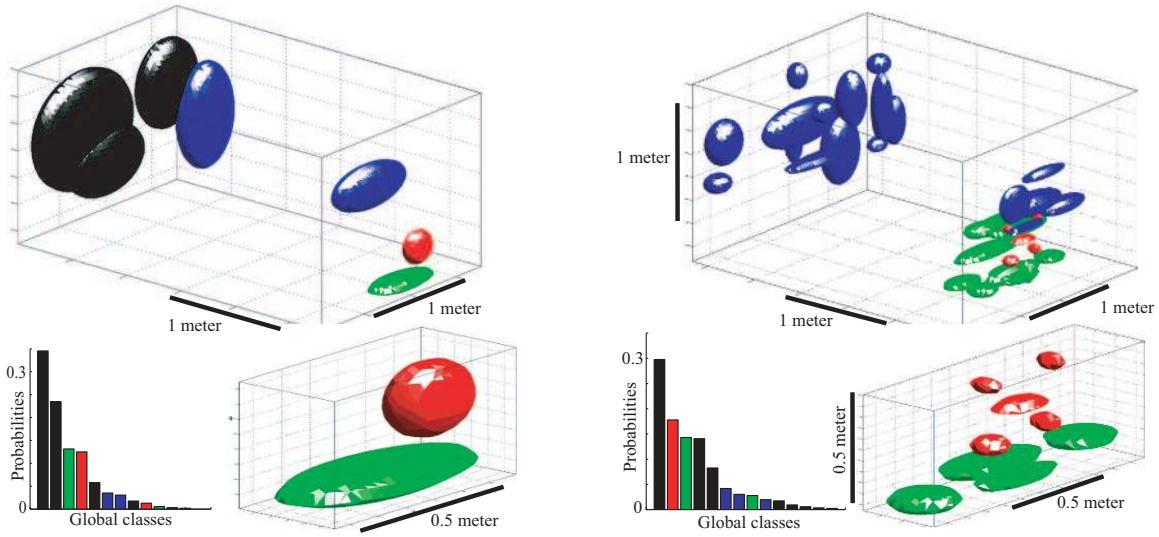

Figure 4. Visual object categories learned from stereo images of office scenes containing computer screens (red), desks (green), bookshelves (blue), and background clutter (black). Covariance ellipses model 3D part geometry, and are positioned at their mean transformed location. Bar charts show posterior probabilities for all instantiated global categories. *Left:* Single part TDP, as in Sec. 3.2. We show the seven visual categories with highest posterior probability (top), and a close–up view of the screen and desk models (bottom). *Right:* Multiple part TDP, as in Sec. 3.3. For clarity, we show the most likely parts (those generating 85% of observed features) for the five most frequent non–background categories (top). The close–up view shows a five–part screen model, and a four–part desk model (bottom).

Note that transformed parts whose mean is farther from the projection ray are given lower overall weight ω_{tk}. To evaluate the likelihood of new object instances $\bar{t}$, we integrate over potential transformations $\rho_{j\bar{t}}$, and evaluate eq. (18) with an appropriately inflated 3D covariance.

The final term of eq. (15) is the depth likelihood corresponding to stereo–based disparity matches. For monocular images, we jointly resample $(t_{ji}, k_{ji}, u^z_{ji})$ by using the prior clustering bias of eqs. (16, 17), and appearance likelihood, to reweight the Gaussian mixture of eq. (18). For stereo training images, we evaluate the likelihood learned in Sec. 2.3 on a uniformly spaced grid determined by the largest expected scene geometry. We then evaluate eq. (18) on the same grid for each candidate instance and part, and resample from that discrete distribution. Given Z depths, and T_j object instances with (on average) K parts, this resampling step requires $\mathcal{O}(ZT_jK)$ operations.

4.2. Inferring Object Categories

In the second phase of each Gibbs sampling iteration, we fix feature depths u^z and object assignments $\mathbf{t}$, and consider potential reinterpretations of each instance t using a new global object category o_{jt}. Because parts and transformations are defined with respect to particular categories, blocked resampling of $(o_{jt}, \rho_{jt}, \{k_{ji} \mid t_{ji} = t\})$ is necessary. Suppose first that $o_{jt} = \ell$ is fixed. Given ρ_{jt}, part assignments k_{ji} are conditionally independent:

$$p\left(k_{ji} = k \mid w_{ji}, u_{ji}, t_{ji} = t, o_{jt} = \ell, \mathbf{k}_{\setminus ji}, \mathbf{t}_{\setminus ji}, \mathbf{o}_{\setminus jt}\right) \propto p\left(k \mid \mathbf{k}_{\setminus ji}, \mathbf{t}, \mathbf{o}\right) \eta_{\ell k}(w_{ji}) \mathcal{N}\left(u_{ji}; \mu_{\ell k}, \Lambda_{\ell k}\right) \quad (19)$$

Here, the first term is as in eq. (17). Alternatively, given fixed part assignments ρ_{jt} has a Gaussian posterior:

$$p\left(\rho_{jt} \mid o_{jt} = \ell, \left\{k_{ji}, u_{ji} \mid t_{ji} = t\right\}\right) \propto \mathcal{N}(\rho_{jt}; \phi_\ell) \prod_{k=1}^{K_\ell} \prod_{i \mid k_{ji} = k} \mathcal{N}\left(u_{ji} - \rho_{jt}; \mu_{\ell k}, \Lambda_{\ell k}\right) \quad (20)$$

The Gaussian transformation prior $\mathcal{N}(\phi_\ell)$ is specific to the visual category (see eq. (14)), while the posterior mean and covariance follow standard equations [4, 14]. Note that our use of continuous, Gaussian position densities avoids an expensive discretization of 3D world coordinates.

For each candidate visual category o_{jt}, we first perform a small number of auxiliary Gibbs sampling iterations using eqs. (19, 20). Given the resulting transformations, the part assignments of eq. (19) may be directly marginalized to compute the likelihood of o_{jt}. The stick–breaking construction of eq. (14) also induces a clustering prior:

$$p\left(o_{jt} \mid \mathbf{o}_{\setminus jt}\right) \propto \sum_{\ell=1}^{L} M_\ell^{-t} \delta(o_{jt}, \ell) + \gamma \delta(o_{jt}, \bar{\ell}) \quad (21)$$

Here, M_ℓ^{-t} denotes the number of object *instances* assigned to the L current categories, and $\bar{\ell}$ indicates a new visual category. Combining these terms, we resample o_{jt}, and conditionally choose $(\rho_{jt}, \{k_{ji} \mid t_{ji} = t\})$ via eqs. (19, 20).

4.3. Inferring Part and Transformation Parameters

The preceding sections assumed fixed values for the parameters $\theta_{\ell k} = (\eta_{\ell k}, \mu_{\ell k}, \Lambda_{\ell k})$ defining part appearance and position, as well as category–specific transformation

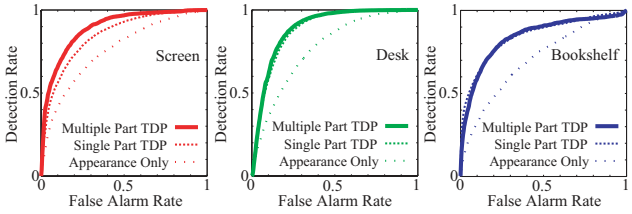

Figure 5. ROC curves for the segmentation of features corresponding to computer screens (red), desks (green), and bookshelves (blue). Using stereo test images, we compare the single and multiple part TDPs of Fig. 4 to a classifier based on feature appearance.

distributions ϕ_ℓ. Given fixed values for all assignments $(\mathbf{k}, \mathbf{t}, \mathbf{o})$, depths $\mathbf{u}^z$, and transformations $\boldsymbol{\rho}$, parameters could be independently resampled via standard methods [4]. For efficiency, we instead analytically marginalize these parameters, replacing all conditional likelihoods by Rao–Blackwellized predictive likelihoods [11, 16]. These integrals are made tractable by our use of conjugate, normal-inverse-Wishart distributions for position densities, and a Dirichlet prior for part appearance densities [4]. Empirically, the inferred model is fairly insensitive to the hyperparameters H and R. For greater robustness, we place vague gamma priors on the DP concentration parameters, and resample them via an auxiliary variable method [2, 16].

5. Analyzing Office Scenes

We consider a dataset of stereo office scenes containing four labeled objects: computer screens, desks, bookshelves, and background clutter. With 120 training images segmented as in Fig. 2, we used the Gibbs sampler of Sec. 4 to learn TDP scene models. Fig. 4 shows visual categories for the full TDP (Fig. 3), and the simpler single part model of Sec. 3.2, after 100 sampling iterations. While the single part TDP captures coarse geometric relationships, parts allow more accurate descriptions of object structure. For example, the screen model defines parts for each of its four corners. Note that the number of parts associated with each category is inferred *automatically*.

During training, we distinguish the four manually labeled *object categories* from the *visual categories* G_0 discovered by the TDP. We restrict the Gibbs sampler from assigning different objects to the same visual category, but multiple visual categories may be used to describe different forms of a particular object. For example, both models learn (without supervision) two shapes for bookshelves, one horizontal and the other vertical. Note that our allowance for transformations causes the TDP to model scaling via 3D translations, rather than multiple visual categories.

Fig. 6 shows typical test image interpretations for the part–based TDP scene model. For stereo test images, TDP depth estimates consistently improve on the raw estimates of Fig. 2. In addition, as shown by the ROC curves of Fig. 5, the TDP more accurately segments features into object categories than a histogram model based solely on feature appearance. Parts improve segmentation performance for monitors, but not for the less structured desk and bookshelf categories. For monocular test images, we detect monitors at multiple scales, and thus approximately infer scene geometry via the presence of familiar objects.

6. Discussion

We have developed an integrated, hierarchical model for the 3D geometry and appearance of multiple object scenes. Applied to a dataset of office scenes, we show that geometry improves feature categorization, and that inferred objects provide coarse depth estimates from monocular images. We believe that the 3D structure of our generative model will provide an exciting framework to explore unsupervised object discovery, recognition from multiple viewpoints, and the incorporation of contextual cues.

Acknowledgments

Funding provided by the National Geospatial-Intelligence Agency NEGI-1582-04-0004, the National Science Foundation NSF-IIS-0413232, the ARDA VACE program, and a BAE Systems grant.

References

[1] G. Csurka et al. Visual categorization with bags of keypoints. In *ECCV Workshop on Statistical Learning in Computer Vision*, 2004.

[2] M. D. Escobar and M. West. Bayesian density estimation and inference using mixtures. *J. Amer. Stat. Assoc.*, 90(430):577–588, June 1995.

[3] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *CVPR*, volume 2, pages 264–271, 2003.

[4] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman & Hall, 2004.

[5] D. Hoiem, A. A. Efros, and M. Hebert. Geometric context from a single image. In *ICCV*, vol. 1, pages 654–661, 2005.

[6] N. Jojic and B. J. Frey. Learning flexible sprites in video layers. In *CVPR*, volume 1, pages 199–206, 2001.

[7] S. Lazebnik, C. Schmid, and J. Ponce. A maximum entropy framework for part-based texture and object recognition. In *ICCV*, volume 1, pages 832–838, 2005.

[8] D. G. Lowe. Distinctive image features from scale–invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[9] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide baseline stereo from maximally stable extremal regions. In *BMVC*, pages 384–393, 2002.

[10] K. Mikolajczyk and C. Schmid. Scale and affine invariant interest point detectors. *IJCV*, 60(1):63–86, 2004.

[11] R. M. Neal. Markov chain sampling methods for Dirichlet process mixture models. *J. Comp. Graph. Stat.*, 9(2):249–265, 2000.

[12] A. Saxena, S. H. Chung, and A. Y. Ng. Learning depth from single monocular images. In *NIPS 18*. MIT Press, 2006.

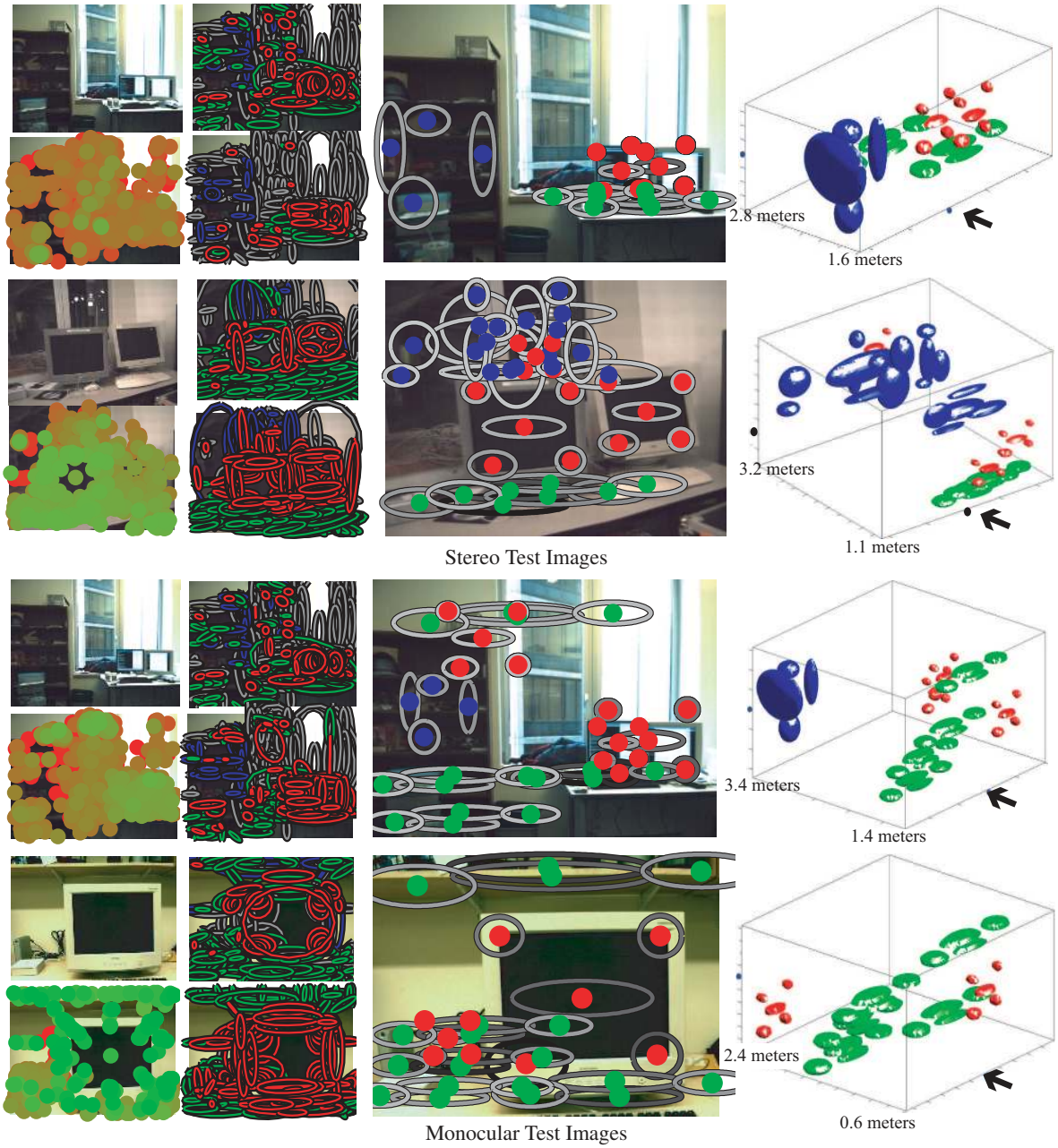

Figure 6. Analysis of stereo (top) and monocular (bottom) test images using the 3D, part–based TDP model of Fig. 4. For each result group, we show (left, clockwise) the test image, a segmentation based solely on feature appearance, a TDP segmentation, and corresponding TDP depth estimates (green features are near, red far). We also show transformed 3D parts corresponding to non–background object instances inferred by the TDP (right), and overlay perspective projections of these parts on the test image (center).

[13] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman. Discovering objects and their location in images. In *ICCV*, volume 1, pages 370–377, 2005.

[14] E. B. Sudderth, A. Torralba, W. T. Freeman, and A. S. Willsky. Learning hierarchical models of scenes, objects, and parts. In *ICCV*, volume 2, pages 1331–1338, 2005.

[15] E. B. Sudderth, A. Torralba, W. T. Freeman, and A. S. Willsky. Describing visual scenes using transformed dirichlet processes. In *NIPS 18*. MIT Press, 2006.

[16] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. Technical Report 653, U.C. Berkeley Statistics, Oct. 2004.

[17] A. Torralba and A. Oliva. Depth estimation from image structure. *IEEE Trans. PAMI*, 24(9):1226–1238, 2002.

[18] Z. Tu, X. Chen, A. L. Yuille, and S. C. Zhu. Image parsing: Unifying segmentation, detection, and recognition. In *ICCV*, volume 1, pages 18–25, 2003.