vector quantize these descriptors, producing a finite dictionary of F appearance patterns. This feature set provides some invariance to lighting and pose changes, and was more effective than features based on unnormalized pixel patches [21] in our experiments.

Given this feature dictionary, the i^{th} interest point in image m is described by its position x_{mi} and the best matching descriptor w_{mi}. Let $\mathbf{w}_m$ and $\mathbf{x}_m$ denote the appearance and position, respectively, of the N_m features in image m. Examples of features detected in this way are shown in Fig. 2.

2.2. Using Parts to Generate Objects

The representation of objects as a collection of spatially constrained parts has a long history in vision [8]. In the graphical model of Fig. 1, parts $\mathbf{z}$ are formalized as clusters of features that appear in similar locations, and have similar appearance. Object categories are in turn defined by a probability distribution θ specifying which parts are most likely to produce corresponding visual features.

Consider the generative process for an image of object o_m containing N_m features $(\mathbf{w}_m, \mathbf{x}_m)$. All feature positions are defined relative to an image–specific coordinate frame, or reference position, r_m. Each object category has its own Gaussian prior over reference positions:

$$p\left(r_m \mid o_m\right) = \mathcal{N}(r_m;\, \zeta_{o_m}, \Phi_{o_m}) \tag{1}$$

To generate the i^{th} feature, we first independently sample a part z_{mi} according to an object–specific multinomial distribution θ_{o_m} over the P possible parts. Then, conditioned on the chosen part index z_{mi}, we independently sample an appearance w_{mi} and position x_{mi}:

$$p\left(w_{mi}, x_{mi} \mid z_{mi} = j, r_m\right) = \phi_j(w_{mi}) \times \mathcal{N}(x_{mi};\, r_m + \mu_j, \Lambda_j) \tag{2}$$

Each part z is defined by a multinomial distribution ϕ_z over the F possible appearance descriptors, as well as a Gaussian distribution over feature positions. Because the mean of this Gaussian is shifted relative to r_m, we may recognize objects whose spatial translation varies from image to image.

Although we assume the collection of objects is known, the probability distributions defining this generative model must be learned from training data. The hierarchical structure allows information to be shared in two distinct ways: parts combine the same features in different spatial configurations, and objects reuse the same parts in different proportions. The learning process, as described in Sec. 3, is free to give each object category its own parts, or "borrow" parts from other objects, depending on which better explains the observed images. As we show in Sec. 4, this sharing can significantly improve detection performance.

When learning statistical models from small data sets, prior distributions play an important regularizing role [4]. To simplify the learning process, we assume that these priors have a conjugate form [9]. In particular, the multinomial distributions θ and ϕ are assigned independent, symmetric Dirichlet priors with hyperparameters α and β, respectively. The covariance matrices Λ_z of the Gaussian part position densities have inverse–Wishart priors with scale Δ_p and ν_p degrees of freedom, while the means μ_z are given noninformative priors. Similarly, the reference position's covariance prior is inverse–Wishart with hyperparameters Δ_o and ν_o.

2.3. Related Models

The graphical model of Fig. 1 was partially inspired by recently proposed models of text documents. In particular, if position variables are neglected, we recover a variant of the *author–topic model* [15], where objects correspond to authors, features to words, and parts to the latent topics underlying a given corpus. The generative aspect model, or *latent Dirichlet allocation (LDA)* [2, 10], is in turn a special case in which each document has its own topic distribution, and authors are not explicitly modeled.

LDA has been previously adapted to discover object categories from images of single objects [17], categorize natural scenes [6], and (with a slight extension) parse presegmented captioned images [1]. However, following an initial stage of low–level feature extraction [6, 17] or segmentation [1], these models ignore spatial information, treating the image as an unstructured *bag of words*. In contrast, our introduction of a reference position allows us to explicitly model the spatial locations of detected features. This extension raises additional computational issues, which we address using the EM algorithm (Sec. 3.2), and leads to improved performance in detection and recognition tasks.

When modeling a single object category, our model also shares many features with constellation models [8], particularly recent extensions which use Bayesian priors when learning from few examples [4, 5]. The principal difference is that their likelihood assumes that each part generates at most one observed feature, creating a combinatorial data association problem for which greedy approximations are needed to learn more than a few parts [11]. In contrast, our association of objects with distributions over parts leads to simple learning algorithms which scale linearly with P. In addition, by sharing parts when learning multiple object categories, we can improve generalization performance.

3. Learning Objects with Shared Parts

In this section, we derive a Gibbs sampling algorithm for learning the parameters of the hierarchical model of Fig. 1. We begin in Sec. 3.1 by assuming that all objects occur at roughly the same position in each image, so that the reference position r_m can be neglected. Many standard object recognition datasets, as well as systems which use cues such as motion to focus attention, satisfy this assumption. In Sec. 3.2, we extend the Gibbs sampler by using the EM