The sums in eq. (9) only include the feature positions from the corresponding image m. These expected reference positions define a lower bound on the likelihood, which is maximized by the M-step. Given M_ℓ images of object ℓ, the maximizing reference position parameters equal

$$\hat{\zeta}_\ell = \frac{1}{M_\ell} \sum_{m|o_m=\ell} \hat{r}_m \qquad \delta_\ell = \frac{1}{M_\ell + \nu_o + 3} \quad (10)$$

$$\hat{\Phi}_\ell = \delta_\ell \left(\Delta_o + \sum_{m|o_m=\ell} R_m + (\hat{r}_m - \hat{\zeta}_\ell)(\hat{r}_m - \hat{\zeta}_\ell)^T \right)$$

The part position parameters are similarly updated as

$$\hat{\mu}_j = \frac{1}{n_j^P} \sum_{m=1}^{M} \sum_{k|z_{mk}=j} (x_{mk} - \hat{r}_m)$$

$$\hat{\Lambda}_j = \frac{1}{n_j^P + \nu_p + 3} \left(\Delta_p + \sum_{m=1}^{M} \hat{\Lambda}_{jm} \right) \quad (11)$$

$$\hat{\Lambda}_{jm} = \sum_{k|z_{mk}=j} R_m + (x_{mk} - \hat{\mu}_j - \hat{r}_m)(x_{mk} - \hat{\mu}_j - \hat{r}_m)^T$$

Note that the updates of eq. (11) are similar to the moment matching of eq. (6), except that parts are translated by the current expected reference position in each image.

We apply these EM updates between every Gibbs sampling operation. Because the posterior mode is not dramatically changed by the reassignment of a single feature, only a single EM iteration per sample is needed for accurate mode tracking. Conditioned on the parameter estimates produced by the M-step, the reference position r_m has a Gaussian distribution with mean and covariance as in eq. (9). The feature position likelihood then has the following closed form:

$$p\left(x_{mi} \mid z_{mi} = j, \bar{\mathbf{z}}_{mi}, \bar{\mathbf{x}}_{mi}, \mathbf{o}\right) = \mathcal{N}(x_{mi};\, \hat{r}_m + \mu_j, R_m + \Lambda_j) \quad (12)$$

This expression is used in eq. (7) to evaluate the probabilities for each Gibbs sampling operation.

Direct implementation of these EM updates requires $\mathcal{O}(MP)$ operations per iteration due to the coupling between the reference positions and parts. However, we may reduce the cost of each iteration to $\mathcal{O}(P)$ using incremental EM updates [14]. In particular, when sampling a part assignment for image m, we fix the expectations of eq. (9) for all reference positions except r_m. By caching statistics of the other reference position estimates, the M–step (eqs. (10, 11)) may also be performed efficiently. Although we no longer find the exact posterior mode, the dependencies of the reference positions in other images on $\mathbf{z}_m$ are very weak, so this approximation is extremely accurate. Empirically, incremental updates produce dramatic computational gains with negligible loss of sampling accuracy.

3.3. Likelihoods for Object Detection

To use the hierarchical model for detection or recognition, we must compute the likelihood that a test image t, with features $(\mathbf{w}_t, \mathbf{x}_t)$, is generated by each candidate object category o. Because each image's features are independently sampled from a common parameter set, we have

$$p\left(\mathbf{w}_t, \mathbf{x}_t \mid o, \mathcal{M}\right) = \int p\left(\mathbf{w}_t, \mathbf{x}_t \mid o, \Theta\right) p\left(\Theta \mid \mathcal{M}\right) d\Theta$$

In this expression, $\mathcal{M}$ denotes the set of training images, and $\Theta = (\theta, \phi, \mu, \Lambda, \zeta, \Phi)$ the model parameters. The sequence of part assignments produced by the Gibbs sampler provides samples $\mathbf{z}^{(s)}$ approximately distributed according to $p\left(\mathbf{z} \mid \mathcal{M}\right)$. Given a set of S samples, we approximate the test image likelihood as

$$p\left(\mathbf{w}_t, \mathbf{x}_t \mid o, \mathcal{M}\right) \approx \frac{1}{S} \sum_{s=1}^{S} p\left(\mathbf{w}_t, \mathbf{x}_t \mid o, \hat{\Theta}^{(s)}\right) \quad (13)$$

where $\hat{\Theta}^{(s)}$ denotes the approximate modes of the posterior distribution over parameters computed using $\mathbf{z}^{(s)}$ in eqs. (4, 5, 6, 10, 11).

When the reference position is neglected, the image features are independent conditioned on the model parameters:

$$p\left(\mathbf{w}_t, \mathbf{x}_t \mid o, \hat{\Theta}^{(s)}\right) = \prod_{i=1}^{N_t} \sum_{j=1}^{P} \hat{\theta}_o(j)\, \hat{\phi}_j(w_{ti})\, \mathcal{N}\Big(x_{ti};\, \hat{\mu}_j, \hat{\Lambda}_j\Big) \quad (14)$$

This expression calculates the likelihood of N_t features in $\mathcal{O}(N_t P)$ operations. To account for the reference position, we first run the Gibbs sampling updates on the test image features. The EM estimates of Sec. 3.2 then provide a reference position estimate which can be combined with the likelihood of eq. (12) to evaluate eq. (14).

4. Object Categorization Experiments

To explore the advantages of sharing parts among objects, we consider a collection of 16 categories with noticeable visual similarities. Fig. 2 shows images from each category, which can be divided into three groups: seven animal faces, five animal profiles, and four wheeled vehicles. As object recognition systems scale to hundreds or thousands of categories, the inter–category similarities exhibited by this dataset will become increasingly common.

4.1. Learning Shared Parts

Given 30 training examples from each of the 16 categories, we constructed a feature appearance dictionary with $F = 600$ words, and used Gibbs sampling (Sec. 3.1) to fit a model with 32 shared parts. Because the database images had been manually aligned, the EM likelihood updates