A Hierarchical Part-Based Model for Visual Object Categorization

Guillaume Bouchard and Bill Triggs
Guillaume.Bouchard@inria.fr, Bill.Triggs@inrialpes.fr
LEAR, GRAVIR-INRIA, 655 av. de l'Europe, 38330 Montbonnot, France

Abstract

We propose a hierarchical generative model for coding the geometry and appearance of visual object categories. The model is a collection of loosely connected parts containing more rigid assemblies of subparts. It is optimized for domains where there are relatively large numbers of somewhat informative subparts, such as the features returned by local feature methods from computer vision. The model is learned quickly by an E-M procedure. Some experiments on real images show the its ability to fit complex natural object classes.

1 Introduction

In object categorization from digital images, existing geometrical models are typically very specific to a particular object class (for example 3D human body models). There is a need for generic models that are suitable for more general object categories. "Part" or "fragment" based models that combine local image features or regions into loose geometric assemblies offer one possible solution to this [9, 11, 4, 3, 8]. One such approach is the method of Fergus, Perona & Zisserman ("FZP") [4]. This is a joint probabilistic model for multiple parts distributed normally in appearance and location space. One of its major limitations is the fact that it is requires an explicit enumeration over possible matchings of model features to image ones. This optimal, but combinatorially expensive, step limits the model to relatively few detected features ('parts'), typically 6 or at most 7. This in turn means that a good deal of the available image information must often be ignored, especially in cases where the objects have many parts, either naturally, or because fine grained local visual features are being used to characterize them. Indeed, such structural approaches often fail to compete with geometry-free "bag of features" style approaches because the latter make better use of the available image information [9, 10, 1]. Hence it is useful to investigate structural models that can handle models with hundreds of local features efficiently.

Secondly, many natural object categories (humans and animals, man made classes with variable forms) have relatively rigid local shape, but significant large scale shape variability, so that nearby object features have strongly correlated positions while more distant ones are much more weakly correlated. Another advantage of part-based models is that they can easily represent this kind of covariance structure. But to do this well, it is natural to include at least two levels of part hierarchy, with loosely connected parts containing more

Parameters Coding Model Structure	
ρ_r	the part r average position relative to the center of gravity
α_{kr}	the k^{th} subpart average position relative to the center of the part r
τ_{kr}	the probability for the k^{th} subpart to be assigned to the part r
π_k	the probability of observing the k^{th} subpart
σ_r^2	the variances of the parts around their mean
Random Variables depending on the Image i	
g_i	the center of gravity of the object
s_i	the scale of the object
h_{ir}	the position of the part r in the image
ℓ_{ik}	the observed position of the part k
$\Gamma_{ik} \in \{1, \ldots, R\}$	the index of the part to which the subpart k is assigned
$O_{ik} \in \{0, 1\}$	the observation variable equal to one is the subpart k is observed, 0 if is is an outlier

Table 1: A summary of the parameters and variables used in our model.

tightly connected subparts. Hence the overall model becomes a tree-structured graphical model [7].

In this paper, we propose such a two-level model that is capable of handling hundreds of subparts efficiently (albeit slightly suboptimally). Compared to FZP, we use a two layer model rather than a single layer one, and we have simplified the correspondence problem by using greedy nearest-neighbour matching in location-appearance space . . . but as a result we can deal with many more subparts. The model is learned by using E-M over the hidden structure variables.

Below, we first present the probabilistic model and the initial matching procedure. Then the learning method is explained, and finally we show some experiments on real images.

2 Part-based Model

In this paper we consider only a two-layer model containing R parts and K subparts, however additional hierarchical layers can easily be added if required. Table 1 defines the parameters and variables used.

2.1 Appearance and Location Clustering

As input, our method requires a set of local features detected in each image. In the experiments below, we followed a current computer vision trend and used SIFT descriptors [9, 10] calculated over scale invariant salient regions returned by Kadir & Brady's entropy-based local region detector [6]. However, any kind of local features with appearance descriptors in some space $\mathcal{A}$ and some kind of local position variables can be used. To construct the inputs to the model, we need a first guess of the position of the subparts in the image. This is done by grouping the detected points in the joint local position / appearance space $\mathbb{R}^2 \times \mathcal{A}$ into K clusters, using K-means with the norm $\|\ell\|^2 + q\|a\|^2$. Each cluster defines a corresponding initial class of subparts, coded by the cluster centre $\mu_k = (\lambda_k, \mu_k)'$. The constant q is set to make the influence of the appearance similar in magnitude to that of the location.

Given an image i, to get the initial values of (ℓ_i, a_i), we match each of the K cluster centres μ_k to the detected image point that is closest to it under the above position-appearance

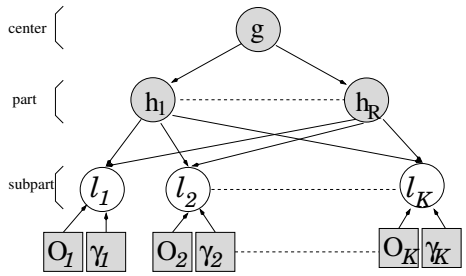


Figure 1: Left: graphical model of the image random variables. The gray nodes represent hidden variables.

norm:

$$(\ell_{ik}, a_{ik}) \;=\; \arg \min_{j \in \{1,\ldots,n_i\}} \|x^{\ell}_{ij} - \lambda_k\|^2 + q\,\|x^{a}_{ij} - \mu_k\|^2$$

where n_i is the number of detections in the image and x^a_{ij}, x^ℓ_{ij} are the detected appearance and location vectors.

Finally, λ_k is used as initial value for the quantity $\sum_k \alpha_{kr} + \rho_r$. Note that we do not re-estimate the values of the appearance means μ^a_k in the learning step, as we consider that we have chosen K large enough to code the full set of possible feature appearances and locations reasonably well.

2.2 Generative model

We now define the generative model for an image i, where the object occurs at position g_i and scale s_i. The position of each of the r parts is assumed to have a Gaussian distribution around its rescaled mean offset:

$$h_{ir} \,|\, g_i, \rho_r \;\;\sim\;\; \mathcal{N}(g_i + s_i\rho_r, s_i^2\sigma_r^2 I_d).$$

The observed subparts can be generated by any part or the background. The parent of subpart k is part r with probability τ_{kr}. Given this assignment, the position of the subpart relative to the part has a local Gaussian distribution:

$$\ell_{ik} \,|\, \Gamma_{ik}{=}r, h_{ir}, \alpha_{kr} \;\;\sim\;\; \mathcal{N}(h_{ir} + s_i\alpha_{kr}, s_i^2 I_d).$$

Only the positions ℓ_{ik} and appearances a_{ik} of the K subparts are directly observable. The positions of the parts and the overall centre and scale are hidden variables that must be estimated anew in each image.

Putting all of these components together, the posterior probability for the complete model is:

$$\mathrm{P}(g, \rho, h, \alpha, \pi, \tau | \ell, a) \;\;\propto\;\; \mathrm{P}(\ell, a|*)\mathrm{P}(h|g,\rho)\mathrm{P}(g)\mathrm{P}(\rho)\mathrm{P}(\alpha) \qquad (1)$$

where $\mathrm{P}(x|*)$ denotes the distribution of x conditional on all of the other variables, $g, \rho, h, \alpha, \pi, \tau$.

Priors: The distribution $\mathrm{P}(g)$ is assumed to be uniform on the image range. The prior on ρ is also assumed uniform. Parts are encouraged to be spatially compact by putting a Gaussian prior on α with zero mean and variance η^2. A very small η would force all of the subparts to be located at one of the part centers. Such a constrained model could be viewed as simple extension of the rigid model of Weber *etal* [12], in which each subpart is a mixture of possible appearances.

Part and subpart models: As above, the distribution of the parts relative to the object frame is

$$\mathrm{P}(h_{ir}|g_i,\rho_r) \;=\; \Phi(h_{ir};g_i+\rho_r,s_i^2\sigma_r^2 I_d),$$

where $\Phi(m,\Sigma)$ is the 2-D Gaussian distribution with mean m and covariance Σ. To ensure a unique parametrization, we add the natural constraint that the object frame is centred on the mean of the part positions ρ_r, *i.e.* $\sum_r \rho_r = 0$.

The subpart distribution $\mathrm{P}(\ell,a|*)$ is somewhat more complicated as it depends on the hidden variables O and Γ. Conditioning on the observation variable O allows us to separate location and appearance:

$$\mathrm{P}(\ell,a|*) \;=\; \prod_{k=1}^{K}\left\{\pi_k\,\mathrm{P}(\ell_k,a_k|O_k=1)+(1-\pi_k)\,\mathrm{P}(\ell_k,a_k|O_k=0)\right\} \tag{2}$$

The appearance and location of the subpart are independent given O, and we assume that the appearance does not depend on the assigned parent part r, so:

$$\mathrm{P}(\ell_{ik},a_{ik},\Gamma_{ik}=r|O_{ik}=j,*) \;=\; \begin{cases} p_{ik}^{app}\,p_{ikr}^{loc} & \text{if } j=1 \\ \frac{1}{\chi^{app}\chi^{loc}} & \text{if } j=0 \end{cases} \tag{3}$$

with

$$p_{ik}^{app} \;=\; \mathrm{P}(a_{ik}|O_{ik}=1,*) \;=\; \Phi(a_{ik};\mu_k,\sigma^{app}) \tag{4}$$

$$p_{ikr}^{loc} \;=\; \mathrm{P}(\Gamma ik=r)\mathrm{P}(\ell_{ik}|\Gamma ik=r,h,\alpha,O_{ik}=1) \tag{5}$$

$$\;=\; \tau_{kr}\Phi(\ell_{ik};h_{ir}+\alpha_{kr},I_d). \tag{6}$$

Assuming uniform distributions for outliers (unassigned subparts), χ^{app} and χ^{loc} are constants. χ^{loc} is set to the area of the image domain that we search for possible matches. χ^{app} is the volume of the appearance-space domain in which two descriptors are considered to be similar (which should be set by preliminary studies with the image descriptor that is being used).

We use the E-M algorithm to make maximum likelihood estimates of our model. The full E-M update equations are given in the appendix.

3 Experiments

Datasets: We used five different image classes from the Caltech database: motorbikes (200 images), aeroplanes (362 images) backgrounds (430 images), leaves (186 images) and faces (435 iamges). These datasets have already been used by several groups, *e.g.* [12, 5, 3, 2]. Half of the images in each class were held out for testing.

Feature Detection: For our underlying features we used the Kadir & Brady entropy-based scale invariant salient region detector [6], but any other fairly well localized class of features could be used. We used $K = 200$ clusters to characterize the subparts.

Training: The current implementation assumes that the position, orientation and scale of the objects in the training images is approximately known (although both of these are estimated on-line in the test images). For motorbikes images, the approximate bounding box of the object location was initialized by hand. We set the subpart coherence parameter η to 2.0, a value that separates parts well in the images, while still allowing most of the subparts to contribute actively to the fit.

Experimentally, the EM algorithm converges in around 30 iterations. On these datasets, about one minute was needed to learn the parameters, as compared to many hours for the FZP model.

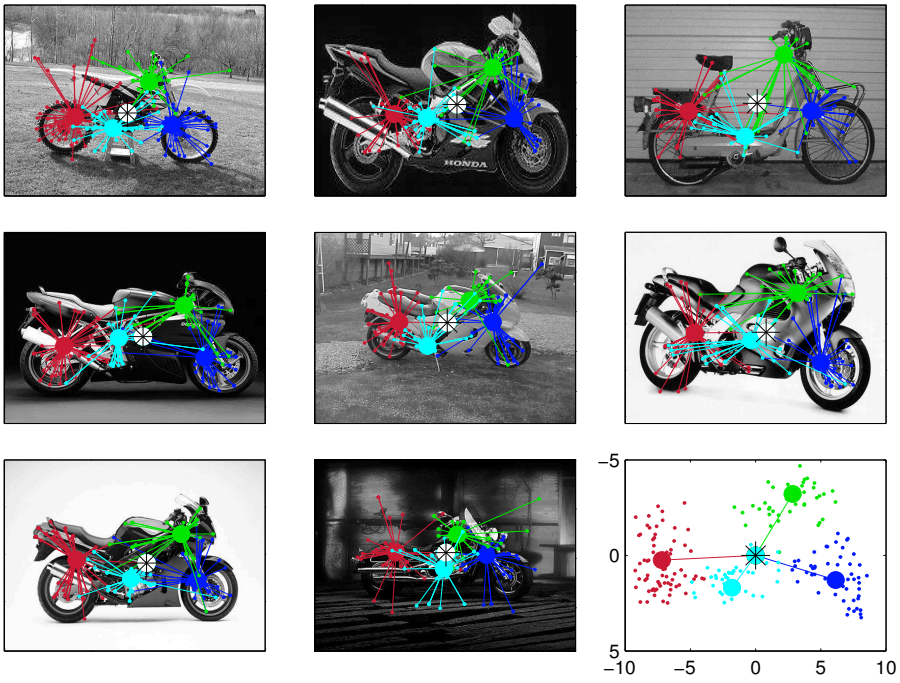
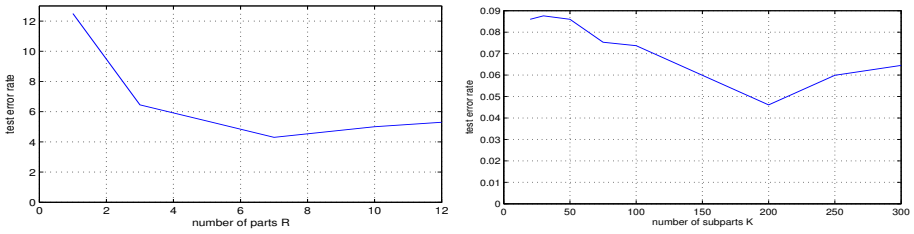


Figure 2: Some examples of our motorbike model with $R = 4$ parts applied to the motorbike dataset. The final image shows the canonical structure of the model. Each point on this figure gives the average position of a subpart over the training set: $x_k = \sum_r \tau_{kr}$.

The models estimated for the aeroplane and motorbike datasets are illustrated in figures 3 and 2. It can be seen that the relative locations of the parts and subparts have adjusted relatively well to the forms of the objects, and that the final models have sufficient flexibility to adapt to a considerable range of shape variation in each case.

To test whether the models had really managed to learn the most important appearance parameters and spatial interrelationships, and whether they were sufficiently selective for a given object category, we assessed their discriminative power by fitting true-class and false-class models to unseen test images, and using their fitted log likelihood ratios as decision variables. The decision thresholds were set to give equal error rates for false positives and false negatives. The resulting error rates for $R = 1$ and 3 parts models are given in table 2. The basic rigid model ($R = 1$) is already strongly discriminative for these data sets. Using a model with 3 parts reduced the error rate by a factor of two. The following graphs plot the test error rate of the leaves/faces classifier against R and K:



we see that the actual number of parts is not very critical, but overfitting starts to worsen the results at around 8-10 parts. It is also clear that a large number of subparts is needed

Figure 3: Our $R = 3$ part aeroplane model to the aeroplane dataset. Note the range of viewpoints and visibility of the wings, and the fact that the 'part' degrees of freedom allow the model to adjust to a considerable range of variation in the fuselage length.

	3 parts models				one-part model			
model	aero.	motos	bg.	leaves	arero.	motos	bg.	leaves
motorbikes	1.66				2.0			
backgrounds	1.4	0.0			2.79	0.0		
leaves	3.31	0.0	2.15		4.97	0.0	8.6	
faces	1.10	0.0	0.0	6.45	2.3	1.0	0.9	12.9

Table 2: Test set error rates in % for binary probabilistic classifiers based on true-model versus false-model likelihood ratios, for $R = 3$ and $R = 1$ parts.

for optimal results – about 200 in this case.

4 Conclusions and Future Work

We have described a two-layered part-based generative model for category-level visual object recognition using large numbers of local features. The model managed to adapt very well to the object categories tested in supervised classification experiments. Reasons for this are its well-graded spatial flexibility, and the fact that it can efficiently incorporate a large number of interest points, each carrying a worthwhile amount of discriminant information. We also showed experimentally that so long as the model uses sufficiently many detected points, the matching of subparts to image features does not need to be very accurate.

Future work: Our priority is to include a matching step into the recursive learning pro-

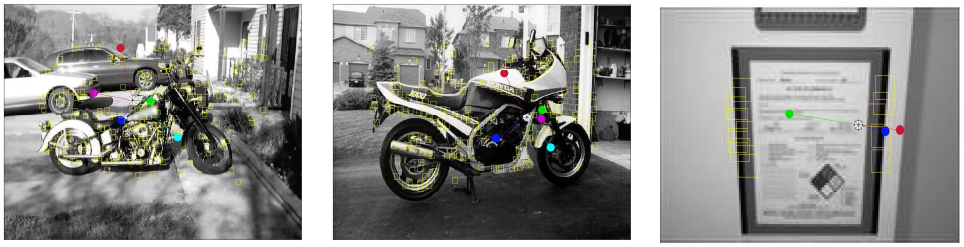


Figure 4: Some examples of incorrectly classified test images, with the detection and the estimated part positions. Most of the incorrect classifications arise because the fitting algorithm has become stuck in a local minimum and hence gives incorrect likelihood estimates. The right image shows the plane model estimated on a badly classified background image.

cedure, to allow automatic localization of the objects during learning. Other obvious extensions are to study deeper hierarchical structures (with parts, subparts, sub-subparts, etc), and to use more flexible models of the parts' positions, including relative scale changes and rotations as well as relative translations.

Acknowledgments

This work was supported in part by the European Union research project LAVA. We thank Navneet Dalal and György Dorkó for their help.

Appendix: E-M Equations

For clarity below, we use scale-normalized positions for the centers, parts and subparts: $\tilde{h}_{ir} = h_{ir}/s_i$, $\tilde{\ell}_{ir} = \ell_{ir}/s_i$ and $\tilde{g}_{ir} = g_{ir}/s_i$. The EM algorithm is used to maximize the likelihood of our model.

E step: We note θ the set of variables $\{\alpha, \rho, \tilde{h}, \tilde{g}, \sigma_r\}$.

$$w_{ikr} = \mathrm{P}(\Gamma_{ik} = r, O_{ik} = 1 | \theta, \ell_{ik}, a_{ik}) \tag{7}$$

$$= \frac{\mathrm{P}(O_{ik} = 1)\mathrm{P}(\ell_{ik}, a_{ik}, \Gamma_{ik} = r | O_{ik} = 1, \theta)}{\sum_{j \in \{0,1\}} \mathrm{P}(O_{ik} = j)\mathrm{P}(\ell_{ik}, a_{ik}, \Gamma_{ik} = r | O_{ik} = j, \theta)} \tag{8}$$

with p^{app}_{ik}, p^{loc}_{ikr}, χ^{app} and χ^{loc} are previously defined.

M step: The completed log-likelihood is:

$$\mathcal{L}_c(\theta) = -\frac{1}{2} \sum_{i,k,r} \left\{ \frac{1}{\eta^2} \|\alpha_{kr}\|^2 + \frac{1}{\sigma_r^2} \|\tilde{h}_{ir} - \tilde{g}_i - \rho_r\|^2 + w_{ikr} \|\tilde{\ell}_{ik} - \tilde{h}_{ir} - \alpha_{kr}\|^2 \right\} \tag{9}$$

To maximize $\mathcal{L}_c$, we set the derivatives equal to zero. Given the variance parameters η and σ_r, the variables α,$\tilde{h}$ and $\tilde{g}$, can be found by solving the system of $KR + R + NR + N$ linear equations:

$$\begin{array}{lccccccc}
\sum_i & (w_{ikr} + \frac{1}{\eta^2})\alpha_{kr} & & + w_{ikr}\tilde{h}_{ir} & & - w_{ikr}\tilde{\ell}_{ik} & = 0 \\
\sum_{i,k} & & \tilde{\rho}_r & - \tilde{h}_{ir} & + \tilde{g}_i & & = 0 \\
\sum_k & w_{ikr}\alpha_{kr} & - \frac{1}{\sigma_r^2}\tilde{\rho}_r & + (w_{ikr} + \frac{1}{\sigma_r^2})\tilde{h}_{ir} & - \frac{1}{\sigma_r^2}\tilde{g}_i & - w_{ikr}\tilde{\ell}_{ik} & = 0 \\
\sum_{k,r} & & & - \tilde{h}_{ir} & + \tilde{g}_i & & = 0
\end{array} \tag{10}$$

In the last equation we used the constraint $\sum \tilde{\rho}_r = 0$ to remove the term $\tilde{\rho}_r$. We can reduce (by substitution) the number of equations of this system to either $KR + R$ or $NR + R$, depending on the application. Otherwise, it is possible to apply the reestimation formulae:

$$\alpha_{kr} = \frac{1}{\frac{1}{\eta^2} + \sum_i w_{ikr}} \sum_i w_{ikr}(\tilde{\ell}_{ik} - \tilde{h}_{ir}) \tag{11}$$

$$\tilde{\rho}_r = \frac{1}{NK} \sum_{i,k} \tilde{h}_{ir} - \tilde{g}_i \tag{12}$$

$$\tilde{h}_{ir} = \frac{1}{\frac{K}{\sigma_r^2} + \sum_k w_{ikr}} \left(\frac{1}{\sigma_r^2}(\tilde{g}_i + \rho_r) + \sum_k w_{ikr}(\tilde{\ell}_{ik} - \alpha_{kr}) \right) \tag{13}$$

$$\tilde{g}_i = \frac{1}{KR} \sum_{k,r} \tilde{h}_{ir} \tag{14}$$

References

[1] G. Csurka, C. Dance, L. Fan, J. Williamowski, and C. Bray. Visual categorization with bags of keypoints. In *ECCV'04 workshop on Statistical Learning in Computer Vision*, pages 59–74, Prague, 2004.

[2] Gy. Dorko and C. Schmid. Object class recognition using discriminative local features. IEEE *Transactions on Pattern Analysis and Machine Intelligence*, 2004. submitted.

[3] Li Fei-Fei, Rob Fergus, and Pietro Perona. A Bayesian approach to unsupervised one-shot learning of object categories. In *Proceedings of the 9th International Conference on Computer Vision, Nice, France*, pages 1134–1141, Nice, France, 2003.

[4] R. Fergus, P. Perona, and A.Zisserman. Object class recognition by unsupervised scale-invariant learning. In *Proceedings of the Conference on Computer Vision and Pattern Recognition, Madison, Wisconsin, USA*, 2003.

[5] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *Proceedings of the Conference on Computer Vision and Pattern Recognition, Madison, Wisconsin, USA*, 2003.

[6] T. Kadir and M. Brady. Scale, saliency and image description. *International Journal of Computer Vision*, 45(2):83–105, 2001.

[7] William T. Freeman Kevin Murphy, Antonio Torralba. Using the forest to see the trees: A graphical model relating features, objects, and scenes. In *Neural Info. Processing Systems*, 2003.

[8] B. Leibe, A. Leonardis, and B. Schiele. Combined object categorization and segmentation with and implicit shape model. In *ECCV'04 workshop on Statistical Learning in Computer Vision*, pages 17–32, Prague, 2004.

[9] D. G. Lowe. Local feature view clustering for 3D object recognition. In *Proceedings of the Conference on Computer Vision and Pattern Recognition, Kauai, Hawaii, USA*, pages 682–688, December 2001.

[10] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. In *Proceedings of the Conference on Computer Vision and Pattern Recognition, Madison, Wisconsin, USA*, June 2003.

[11] S. Ullman, E. Sali, and M. Vidal-Naquet. A fragment-based approach to object representation and classification. In *4th International Workshop on Visual Form, Capri, Italy*, May 2001.

[12] M. Weber, M. Welling, and P. Perona. Unsupervised learning of models for recognition. In *Proceedings of the 6th European Conference on Computer Vision, Dublin, Ireland*, pages 18–32, 2000.