A Generative Model for Multi Class Object Recognition and Detection

Ilkay Ulusoy

METU, Electrical and Electronics Eng. Department, Ankara Turkey
ilkay@metu.edu.tr

Abstract. In this study, a generative type probabilistic model is proposed for object recognition. This model is trained by weakly labelled images and performs classification and detection at the same time. When test on highly challenging data sets, the model performs good for both tasks (classification and detection).

1 Introduction

In recent years object recognition is being approached using machine learning techniques based on probability theory. From basic decision theory [2] we know that the most complete characterization of the solution is expressed in terms of the set of posterior probabilities $p(k|\mathbf{X})$ where k is one of the classes and $\mathbf{X}$ is the vector which describes the image. Once we know these probabilities it is straightforward to assign the image $\mathbf{X}$ to a particular class to minimize the expected loss.

Since detailed hand-segmentation and labelling of images is very labour intensive, learning object categories from 'weakly labelled' data is studied in recent years. Weakly labelled data means that training images are labelled only according to the presence or absence of each category of object. A major challenge presented by this problem is that the foreground object is accompanied by widely varying background clutter, and the system must learn to distinguish the foreground from the background without the aid of labelled data.

A key issue in object recognition is the need for predictions to be invariant to a wide variety of transformations of the input image. Any member of a category of objects should be recognized in spite of wide variations in visual appearance due to variations in the form and colour of the object, occlusions, geometrical transformations (such as scaling and rotation), changes in illumination, and potentially non-rigid deformations of the object itself. As a soltion to these variations, many of the current approaches rely on the use of local features obtained from small patches of the image. Informative features selected using some information criterion versus generic features were compared in [11] and although the informative features used were shown to be superior to generic features when used with a simple classification method, they are not invariant to scale and orientation. By contrast, generic interest point operators such as saliency [6], DoG (Difference of Gaussians) [8] and Harris-Laplace [9] detectors are invariant to location, scale and orientation, and some are also affine invariant [8] [9] to some degree.

In the hierarchy of object recognition problems, one upper level is the localization of the objects in the view. Most of the methods are good either in classification [3] or

F.A. Savacı (Ed.): TAINN 2005, LNAI 3949, pp. 32–40, 2006.

detection [7] but not both. In [7] an implicit shape model is provided for a given object category and this model consists of a class specific alphabet of local appearances that are prototypical for the object category and of a spatial probability distribution which specifies where each codebook entry may be found on the object. Fergus et al. [5] learn jointly the appearances and relative locations of a small set of parts. Since their algorithm is very complex, the number of parts has to be kept small. [4] tried to find out informative features (i.e. object features) without considering spatial relationship between the features based on information criteria. However, in this supervised approach, hundreds of images were hand segmented. Finally, Xie and Perez [12] extended the GMM based approach of [4] to a semi-supervised case. A multi-modal GMM was trained to model foreground and background features where some uncluttered images of foreground were used for the purpose of initialization.

In this study we will also follow a probabilistic approach for object classification using weakly labelled data set and details of our model is given in Section 2. Our model also has the ability to localize the objects in the image. We do not consider a spatial model but we will label each feature as belonging to the object or background. We will focus on the detection of objects within images by combining information from a large number of patches of the image.In this study we will use DoG interest point detectors, and at each interest point we extract a 128 dimensional SIFT feature vector [8]. Following [1] we concatenate the SIFT features with additional colour features comprising average and standard deviation of (R, G, B), (L, a, b) and $(r = R/(R+G+B), g = G/(R+G+B))$, which gives an overall 144 dimensional feature vector [10]. In Section 3 we will provide experiments we have performed and discuss our results.

2 The Generative Model

We used a probabilistic generative model and model the joint distribution of image label and image $p(\mathbf{t}, \mathbf{X})$. Here $\mathbf{X}$ denotes the image and $\mathbf{t}$ denotes the image label vector with independent components $t_k \in \{0, 1\}$ in which $k = 1, \ldots K$ labels the class. Each class can be present or absent independently in an image. $\mathbf{X}$ denotes the observation for image. $\mathbf{X}$ comprises a set of J patch vectors $\{\mathbf{x}_j\}$ where $j = 1, \ldots, J_n$. τ_{j} denotes the patch label vector for patch j with components $\tau_{jk} \in \{0, 1\}$ denoting the class of the patch. These are mutually exclusive, so that $\sum_{k=1}^{K} \tau_{jk} = 1$.

By decomposing the joint distribution into $p(\mathbf{t}, \mathbf{X}) = p(\mathbf{X}|\mathbf{t})p(\mathbf{t})$ we model the two factors separately. The prior probability $p(\mathbf{t})$ is specified in terms of K parameters ψ_k where $0 \leqslant \psi_k \leqslant 1$ and $k = 1, \ldots, K$:

$$p(\mathbf{t}) = \prod_{k=1}^{K} \psi_k^{t_k} (1 - \psi_k)^{1-t_k}. \tag{1}$$

The ψ_k parameters can be taken directly as the real world frequencies.

The remainder of the model, $p(\mathbf{X}|\mathbf{t})$, is specified in terms of the conditional probabilities $p(\mathbf{X}|\mathbf{t}) = p(\boldsymbol{\tau}|\mathbf{t})p(\mathbf{X}|\boldsymbol{\tau})$. The probability of generating a patch from a particular class is governed by a set of parameters π_k, one for each class, such that $\pi_k \geqslant 0$, constrained by the subset of classes actually present in the image. Thus

$$p(\boldsymbol{\tau}|\mathbf{t}) = \left(\sum_{l=1}^{K} t_l \pi_l\right)^{-1} \prod_{k=1}^{K} (t_k \pi_k)^{\tau_k} \tag{2}$$

Note that there is an overall undetermined scale to these parameters, which may be removed by fixing one of them, e.g. $\pi_1 = 1$.

For a given class, the distribution of patch feature vector $\mathbf{x}$ is governed by a separate mixture of Gaussians which we denote by

$$p(\mathbf{x}|\boldsymbol{\tau}) = \prod_{k=1}^{K} \phi_k(\mathbf{x}; \boldsymbol{\theta}_k)^{\tau_k} \tag{3}$$

where the model $\phi_k(\mathbf{x}; \boldsymbol{\theta}_k)$ is given by

$$\phi_k(\mathbf{x}; \boldsymbol{\theta}_k) = \sum_{m=1}^{M} \rho_{km} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_{km}, \boldsymbol{\Sigma}_{km}).$$

If we assume N independent images, and for image n we have J_n patches drawn independently, then the joint distribution of patch labels and patch feature vectors is

$$\prod_{n=1}^{N} p(\mathbf{t}) \prod_{j=1}^{J_n} p(\mathbf{x}_{nj}|\boldsymbol{\tau}_{nj}) p(\boldsymbol{\tau}_{nj}|\mathbf{t}_n). \tag{4}$$

Since the $\{\boldsymbol{\tau}_{nj}\}$ are unobserved we use the EM algorithm. The expected complete-data log likelihood is given by

$$\sum_{n=1}^{N} \sum_{j=1}^{J_n} \left\{ \sum_{k=1}^{K} \langle \tau_{njk} \rangle \ln \left[t_{nk} \pi_k \phi_k(\mathbf{x}_{nj}) \right] - \ln \left(\sum_{l=1}^{K} t_{nl} \pi_l \right) \right\}. \tag{5}$$

The expected values of τ_{nkj} are computed in the E-step using

$$\begin{aligned} \langle \tau_{njk} \rangle &= \sum_{\{\boldsymbol{\tau}_{nj}\}} \tau_{njk} p(\boldsymbol{\tau}_{nj}|\mathbf{x}_{nj}, \mathbf{t}_n) \\ &= \frac{\displaystyle\sum_{\{\boldsymbol{\tau}_{nj}\}} \tau_{njk} p(\boldsymbol{\tau}_{nj}, \mathbf{x}_{nj}|\mathbf{t}_n)}{\displaystyle\sum_{\{\boldsymbol{\tau}_{nj}\}} p(\boldsymbol{\tau}_{nj}, \mathbf{x}_{nj}|\mathbf{t}_n)} \\ &= \frac{t_{nk} \pi_k \phi_k(\mathbf{x}_{nj})}{\displaystyle\sum_{l=1}^{K} t_{nl} \pi_l \phi_l(\mathbf{x}_{nj})}. \end{aligned} \tag{6}$$

Notice that the first factor on the right hand side of (2) has cancelled in the evaluation of $\langle \tau_{njk} \rangle$.

Setting the derivative with respect to one of the parameters equal to zero and rearranging the equality, the re-estimation equations are obtained and following the EM algorithm all parameters are estimated. For details of the method please refer to [10].

This model can be viewed as a generalization of that presented in [12] in which a parameter is learned for each mixture component representing the probability of that component being foreground. This parameter is then used to select the most informative N components in a similar approach to [4] and [11] where the number N is chosen heuristically. In our case, however, the probability of each feature belonging to one of the K classes is learned directly.

Given all patches $\mathbf{X} = \{\mathbf{x}_j\}$ from an image, for the inference of the patch labels, the posterior probability of the label $\boldsymbol{\tau}_j$ for patch j can be found by marginalizing out all other hidden variables

$$\begin{aligned} p(\boldsymbol{\tau}_j|\mathbf{X}) &= \sum_{\mathbf{t}} \sum_{\boldsymbol{\tau}/\boldsymbol{\tau}_j} p(\boldsymbol{\tau}, \mathbf{X}, \mathbf{t}) \\ &= \sum_{\mathbf{t}} p(\mathbf{t}) \frac{1}{\left(\sum_{l=1}^{K} \pi_l t_l\right)^J} \prod_{k=1}^{K} \left(\pi_k t_k \phi_k(\mathbf{x}_j)\right)^{\tau_{jk}} \prod_{i \neq j} \left[\sum_{k=1}^{K} \pi_k t_k \phi_k(\mathbf{x}_i)\right] \end{aligned} \tag{7}$$

where $\boldsymbol{\tau} = \{\boldsymbol{\tau}_j\}$ denotes the set of all patch labels, and $\boldsymbol{\tau}/\boldsymbol{\tau}_j$ denotes this set with $\boldsymbol{\tau}_j$ omitted. Note that the summation over all possible $\mathbf{t}$ values, which must be done explicitly, is computationally expensive.

Inference of image label needs almost as much computation as patch label inference where the posterior probability of image label $\mathbf{t}$ can be computed using

$$p(\mathbf{t}|\mathbf{X}) = \frac{p(\mathbf{X}|\mathbf{t})\, p(\mathbf{t})}{p(\mathbf{X})} \tag{8}$$

where $p(\mathbf{t})$ is computed from the data set, $p(\mathbf{X})$ is the normalization factor and $p(\mathbf{X}|\mathbf{t})$ is calculated by integrating out patch labels

$$\begin{aligned} p(\mathbf{X}|\mathbf{t}) &= \sum_{\boldsymbol{\tau}} \prod_{j=1}^{J} p(\mathbf{X}, \boldsymbol{\tau}|\mathbf{t}) \\ &= \prod_{j=1}^{J_n} \frac{\sum_{k=1}^{K} t_k \pi_k \phi_k(\mathbf{x}_j)}{\sum_{l=1}^{K} t_l \pi_l}. \end{aligned} \tag{9}$$

3 Results and Discussion

In this study, we have used four test sets (motorbikes, bicycles, people and cars) in which the objects vary widely in terms of number, pose, size, colour and texture. Some examples from each set are given in Figure 1. The sets are obtained from PASCAL challenge[1] and are formed by combining some other sets. We used ”train and validation” set for training and ”test” set for testing for each category. Test and train image

[1] Detailed information can be reached at $(http\ :\ //www.pascal - network.org/challenges/VOC/)$.

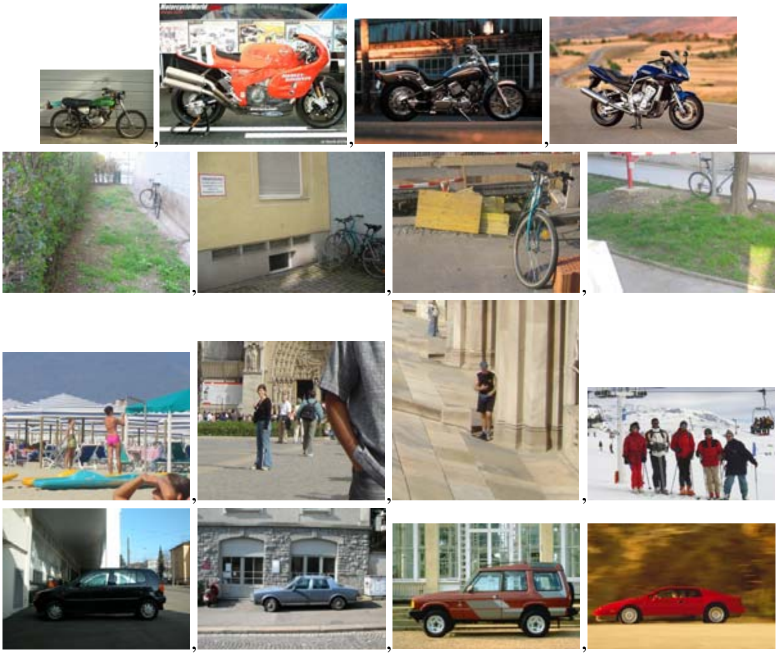

Fig. 1. Example images from test and training sets of motorbikes (first row), bicycles (second row), people (third row) and cars (last row)

Table 1. Train and test set image numbers for each set

Set	Train	Test
Motorbikes	216	216
Bicycles	114	114
People	84	84
Cars	272	275

numbers are given in Table 1. As can be observed from the images, the sets are highly challenging.

An image is labelled by the objects present in the image and also the bounding boxes of the objects in the image are given. In this study, we aim to learn the object categories from weakly labelled images so we do not intend to use bounding box information. The bounding box information is used to test the localization success of the model.

Our model can handle multiple categories to be present in the same image. However, since the train sets (but not the test sets) we used in this study have objects from a single category only, we construct a model for each category. Each model has a separate Gaussian mixture for foreground object and background, each of which has 10 components with diagonal covariance matrices.

Initial results showed that with random initialization of the mixture model parameters it is incapable of learning a satisfactory solution. We conjectured that this is due to the problem of multiple local maxima in the likelihood function (a similar effect was found by [12]). Thus we used one-tenth of training images with their bounding box information for initialization purposes. Feautures inside the bounding box in an image are labelled as foreground and other features are labelled as background. Features belonging to each class were clustered using the K-means algorithm and the component centers of a class mixture model were assigned to cluster centers of the respective class. The mixing coefficients were set to the number of points in the corresponding cluster divided by the total number of points in that class. Also, covariance matrixes were computed using the data points assigned to the respective center.

After training a model for each category, we test it for classification and detection. By classification we mean predicting presence/absence of an example of that class in the test image as been defined by The PASCAL Visual Object Classes Challenge 2005. The classifier produces a real valued output which is actually the probability of image being

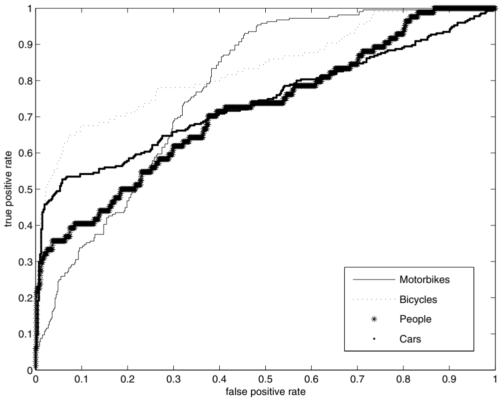

Fig. 2. ROC curves for motorbikes, bicycles, people and cars

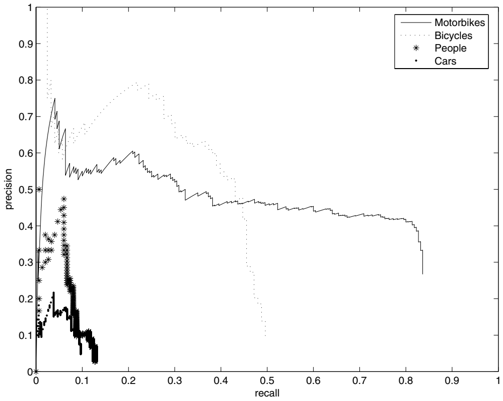

Fig. 3. Precision recall curves for motorbikes, bicycles, people and cars

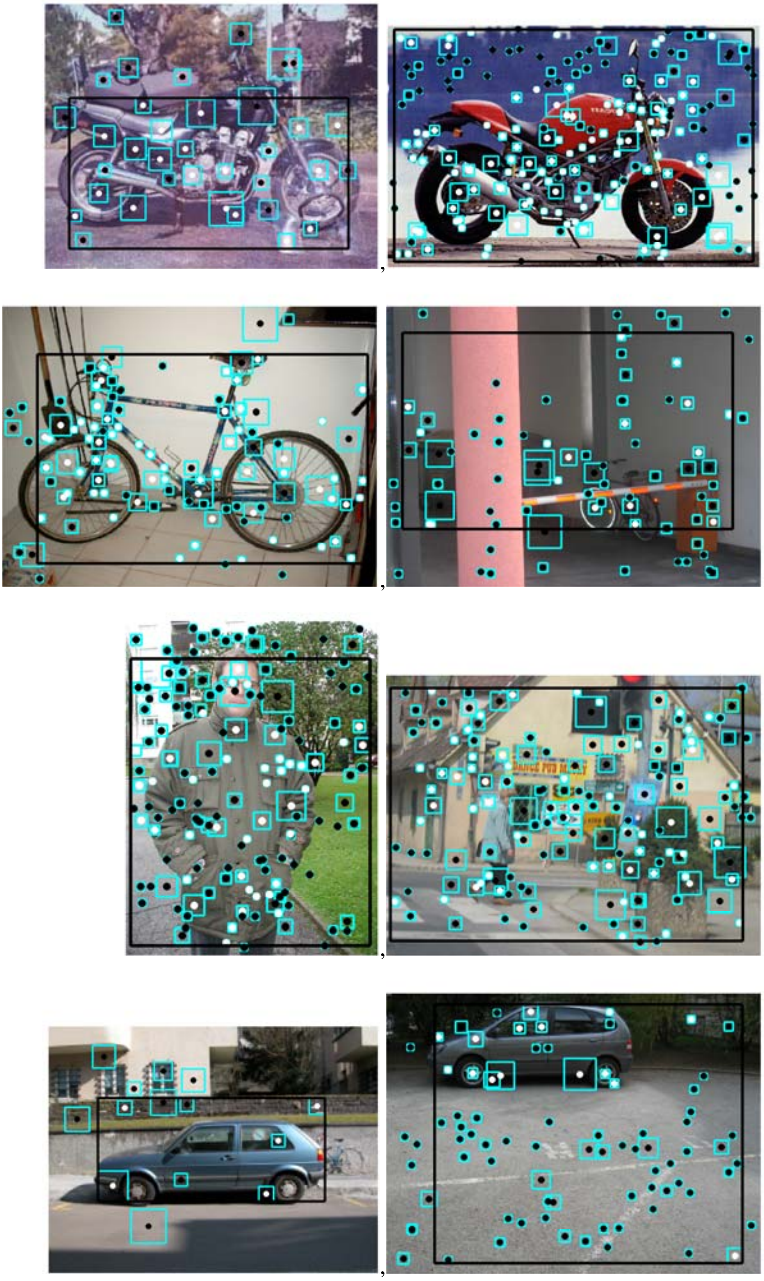

Fig. 4. Example images from test and training sets of motorbikes (first row), bicycles (second row), people (third row) and cars (last row).

labelled as belonging to the category for which the model was trained. By applying different threshold values an ROC curve (true positive versus false positive curves) is plotted for each category. In Figure 2 ROC curves for all categories are plotted. The equal error rates are $70, 75, 65$ and 68% for motorbikes, bicycles, people and cars sets respectively.

By detection we mean predicting the bounding box and label of each object from four target classes in the test image as been defined by The PASCAL Visual Object Classes Challenge 2005. Our model produces a label for each patch as foreground or background. We construct a bounding box as a rectangle which contains all the patches labelled as foreground. Note that in this case we did not consider multiple objects in an image although we have many such images in the test sets. To be considered a correct detection, the area of overlap between the predicted bounding box and ground truth bounding box must exceed 50%. The results of detection are given as precision (true positive over false positive and true positive) recall (true positive over allpositives in the database) curves in Figure 3.

Our model does not have the best classification and detection performances. However, for these challenging sets the model performs well for both localization and classification at the same time.

Both classification and detection are based on patch labelling. Thus patch labelling performance defines the classification and detection performances. In our model, the number of components of Gaussian Mixtures has to be kept small and diagonal matrix instead of full covariance matrix has to be used due to computational problems. Using full covariance matrix brings a very high computational load for 144 dimensional feature vector. Thus we have to use 10 components for each mixture and a diagonal covariance matrix for each component which limit the success of the model. Some patch labelling results for different categories are given in Figure 4. Here patches are shown with squares and each patch centre includes a coloured circle which denotes patch label (white for foreground and black for background). Each row is for a different category (motorbikes, bicycles, people, cars starting from top to bottom). The images in the left column show some good detections whereas right column is for bad detections. A bounding box is shown with a rectangle in each image.

Poor detection is because of our bounding box definition for which we do not consider spatial relationship between patches. All foreground patches are included into the bounding box to give the object location. Any wrong labelled background patch causes wrong detection. For example, the image in the fourth row and right column has a single patch which is labelled as car although it is a background patch. Since this patch is far away from other foreground patches, it causes wrong bounding box placement. Thus bounding box construction should be improved by including spatial relationship between the patches. Also the generative model can be improved by including the spatial relations between the patches.

References

1. K. Barnard, P. Duygulu, D. Forsyth, N. Freitas, D. Blei, and M. I. Jordan. Matching words and pictures. *Journal of Machine Learning Research*, 3:1107–1135, 2003.
2. C. M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, 1995.
3. G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *Workshop on Statistical Learning in Computer Vision, ECCV*, 2004.

4. G. Dorko and C. Schmid. Selection of scale invariant parts for object class recognition. In *ICCV*, 2003.
5. R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale invariant learning. In *CVPR*, 2003.
6. T. Kadir and M. Brady. Scale, saliency and image description. *International Journal of Computer Vision*, 45(2):83–105, 2001.
7. B. Leibe, A. Leonardis, and B. Schiele. Combined object categorization and segmentation with an implicit shape model. In *Workshop on Statistical Learning in Computer Vision, ECCV*, 2004.
8. D. Lowe. Distinctive image features from scale invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.
9. K. Mikolajczyk and C. Schmid. Scale and affine invariant interest point detectors. *International Journal of Computer Vision*, 60:63–86, 2004.
10. I. Ulusoy and C. M. Bishop. Generative versus discriminative methods for object recognition. In *CVPR*, 2005.
11. M. Vidal-Naquet and S. Ullman. Object recognition with informative features and linear classification. In *ICCV*, 2003.
12. L. Xie and P. Perez. Slightly supervised learning of part-based appearance models. In *IEEE Workshop on Learning in CVPR*, 2004.