Log-Linear Mixtures for Object Recognition

Tobias Weyand[1]
Tobias.Weyand@rwth-aachen.de
Thomas Deselaers[12]
deselaers@vision.ee.ethz.ch
Hermann Ney[1]
ney@informatik.rwth-aachen.de

[1] Computer Science Department
RWTH Aachen University
Aachen, Germany

[2] Computer Vision Laboratory
ETH Zurich
Zurich, Switzerland

We present the log-linear mixture model as a fully discriminative approach to object category recognition which can, analogously to kernelised models, represent non-linear decision boundaries. This model is applied to the problem of recognising object classes in natural images, which is one of the most fundamental and best researched problems in computer vision. Similarly to many recent approaches our method uses local image descriptors and learns an object model from weakly annotated training data (i.e. only class labels).

The use of local image descriptors has become a de-facto standard, because it has several advantages:

- simplicity of feature representation and extraction
- robustness w.r.t. changes in object position, scale, and orientation
- robustness w.r.t. occlusion

Probably the most common approach to using local descriptors for image categorization is the bag-of-visual-words approach which consists of three steps:

1. local descriptors are extracted from all images,
2. a vector quantisation technique is applied in order to obtain low-dimensional (often in the order of a few thousand) histograms of local features,
3. a (commonly discriminative) model is used as a classifier to determine whether the object of interest is present in the images.

The bag-of-visual-words approach, although it often leads to good results in practice has several shortcomings:

- During the creation of the histograms, most appearance information is deliberately discarded. This problem can be partly relieved using a soft-quantisation but this only superficially avoids this problem.
- It is difficult to incorporate spatial information into the bag-of-visual words approach.

In our previous work [2], we used Gaussian mixtures to avoid these two problems, resulting in a generative model that implicitly models different object parts using mixtures of Gaussians. The obtained model works reasonable but we found that a discriminative refinement after the training was finished led to a signficant performance boost. Unfortunately, this takes away some of the elegance. Minka [3] noted that discriminative training of generative models should not be done. The model proposed here is fully discriminative and thus avoids this issue entirely.

Note that most bag-of-visual-words models also effectively fuse generative and discriminative models: the quantisation step often applies a generative model (e.g. Gaussian mixture densities obtained using k-means) in order to obtain a fixed length feature vector which is then classified using a discriminative model.

The log-linear mixtures which we present here avoid several of the problems described above and have many desirable properties:

- no hard vector quantisation
- fully discriminative modelling
- easy fusion of different information cues such as appearance and position information

Log-linear mixtures only consist of few parameters which indicates that they should be robust w.r.t. overfitting. We also show that log-linear mixtures can be considered to be the discriminative counterpart of Gaussian mixtures but that they are easier to implement and are numerically more stable to train and to evaluate.

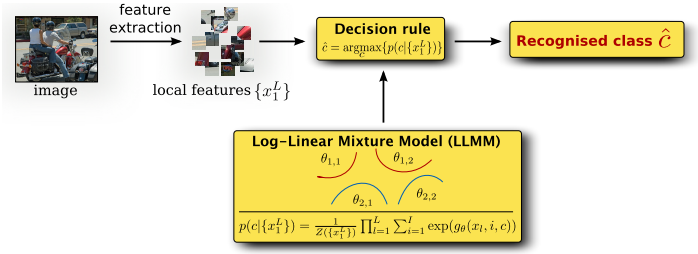

Figure 1. Image object categorisation using log-linear mixtures

Given our log-linear mixtures, the class-posterior for class c of an image, represented by the set of local features $\{x_1^L\} = \{x_1, \ldots, x_L\}$ is defined as

$$p(c|\{x_1^L\}) \quad = \quad \frac{1}{Z(\{x_1^L\})} \prod_{l=1}^{L} \sum_{i=1}^{I} \exp(g_\theta(x_l, i, c)),$$

where g_θ is a linear or quadratic discriminant function with parameters θ, i is a hidden variable denoting the mixture components, and $Z(\{x_1^L\}) = \sum_{c'} \prod_l \sum_i \exp(g_\theta(x_l, i, c'))$ is the normalisation term.

Given a set of training images, the model parameters are trained according to the maximum mutual information criterion using gradient descent. To efficiently train the model, we apply an alternating optimisation scheme, where the assignment of local features to mixture components and the model parameters θ are optimised in turn.

It is easily possible to fuse multiple cues in the proposed model. Assuming we have several different cues, e.g. different types of local features, we can easily extend the model to incorporate all the cues, without explicitly having to account for their scaling and localisation in feature space. Note that in this model the descriptors of the individual cues do not have to be extracted from the same interest points but that they can be entirely unrelated.

We experimentally evaluate our model on the PASCAL VOC 2006 data (see Table 1) and the results compare favourably well to the state-of-the-art despite the model consisting of an order of magnitude fewer parameters, which suggests excellent generalisation capabilities.

Table 1. Comparison of the results of log-linear mixtures on the PASCAL VOC 2006 task with other approaches. The AUC scores are averages over the 10 PASCAL VOC 2006 tasks.

approach	AVG_{10}
(single) log-linear model	0.79
log-linear mixture model (baseline)	0.82
+ second order features	**0.85**
+ additional appearance features	**0.86**
+ spatial layout information	0.82
GMDs (tuned + spatial + disc) [2]	0.86

[1] M. Everingham, A. Zisserman, C. K. I. Williams, and L. Van Gool. The PASCAL Visual Object Classes Challenge 2006 (VOC2006) Results. Technical report, PASCAL Network of Excellence, 2006.

[2] Andre Hegerath, Thomas Deselaers, and Hermann Ney. Patch-based object recognition using discriminatively trained gaussian mixtures. In *British Maschine Vision Conference*, volume 2, pages 519–528, Edinburgh, UK, September 2006.

[3] Tom Minka. Discriminative models, not discriminative training. Technical Report TR-2005-144, Microsoft Research Cambridge, Cambridge, UK, October 2005.