CATEGORY LEVEL OBJECT SEGMENTATION
Learning to Segment Objects with Latent Aspect Models

Diane Larlus and Frédéric Jurie
LEAR Group, INPG-CNRS, INRIA Rhône-Alpes, France

Keywords: Object segmentation, Latent aspect models.

Abstract: We propose a new method for learning to segment objects in images. This method is based on a latent variables model used for representing images and objects, inspired by the LDA model. Like the LDA model, our model is capable of automatically discovering which visual information comes from which object. We extend LDA by considering that images are made of multiple overlapping regions, treated as distinct documents, giving more chance to small objects to be discovered. This model is extremely well suited for assigning image patches to objects (even if they are small), and therefore for segmenting objects. We apply this method on objects belonging to categories with high intra-class variations and strong viewpoint changes.

1 INTRODUCTION

The problem of image segmentation and labeling image region is one of the key problems of computer vision. It consists in separating or grouping image pixels into consistent parts, brought to be elements that humans consider as individual objects or distinct object parts. This problem received a huge amount of attention in the past, and was originally addressed as an unsupervised problem. Many different methods have been developed, using various image properties such as color, texture, edges, motion, etc. (Haralick and Shapiro, 1985). It eventually turned out that image segmentation and image understanding were two closely related problems which cannot be solved independently. After being abandoned for a while, image segmentation came back into favor recently, taking advantage of recent advances of machine learning.

The goal addressed here is the segmentation of objects belonging to a given category (the so-called *figure-ground segmentation* problem) assuming the category is defined by a set of training images. This is illustrated in Figure 1 for the "bicycle" category which is a very challenging category. The overall objective is to classify image pixels as being *figure* or *ground*. Objects can appear in any size and any position in the image. They can occur with widely varying appearances.

Figure 1: We show an image of the "bike" category, with its corresponding hand-made segmentation masks. Our goal is to design algorithms able to compute automatically this segmentation.

In such conditions, object segmentation is strongly linked to object detection and recognition. Indeed, segmenting objects requires learning object models from training images, as well as to search for occurrences of these models in images.

For this paper we will focus on difficult real-condition images where the objects can present extreme appearance variations (see Figure 1).

1.1 Previous Work

The method proposed in this paper is inspired by several related recent works, summarized below.

Leibe and Schiele (Leibe and Schiele, 2003) were among the first authors proposing to learn how to segment objects. Their method generates object hypotheses, without prior segmentation, that can be exploited

Larlus D. and Jurie F. (2007).
CATEGORY LEVEL OBJECT SEGMENTATION - Learning to Segment Objects with Latent Aspect Models.
In *Proceedings of the Second International Conference on Computer Vision Theory and Applications - IU/MTSV*, pages 122-127

to obtain a category-specific figure-ground segmentation. Training images are used to build a visual vocabulary of interest points, containing information about their relative positions as well as their corresponding segmentation masks.

Borenstein *et al.*(Borenstein et al., 2004) use the same idea of selecting informative patches from training images and then use their segmentation masks on new unseen images. They combine bottom-up and top-down approaches into a single process. The top-down approach uses object representation learned from examples to detect an object in a new image and provides an approximation to its segmentation. The bottom-up approach uses image-based criteria to define coherent groups of pixels that are likely to belong to the same part. The resulting combination benefits from both approaches.

Several approaches propose to use Conditional Random Field (CRF) for part-based detection (Quattoni et al., 2004) or segmentation (Kumar and Hebert, 2006). The previous authors extend the notion of CRFs to Discriminative Random Fields (DRFs) by exploiting probabilistic discriminative models instead of the generative models generally used with CRF.

Kumar *et al.* (Kumar et al., 2005) propose another methodology for combining top-down and bottom-up cues with CRFs. They combine CRFs and pictorial structures (PS). The PS provides good priors to CRFs for specific shapes and provides much better results.

None of the previous approaches is able to cope with occlusion. Win and Shotton (Winn and Shotton, 2006) were the first to address specifically this problem using an enhanced CRF. Their approach allows the relative layout (above/below/left/right) of parts to be modeled, as well as the propagation of long-range spatial constraints.

1.2 Description of our Approach

Our approach shares many common features with the previously mentioned approaches. First, it combines bottom-up and top-down strategies.

The bottom-up process consists in sampling, and normalizing in size, dense image patches (small square image sub-windows), as in (Kumar and Hebert, 2006; Winn and Shotton, 2006), represented for subsequent processing by SIFT descriptors (Lowe, 2004). These descriptors are then vector quantized into a discrete set of labels so called *visual words*. Each patch is described by the word of the nearest centroid. This process is illustrated Figure 2. From this stage, images are seen as sets of visual words occurrences. As the process assigns figure/ground labels to patches, the pixel level segmentation requires

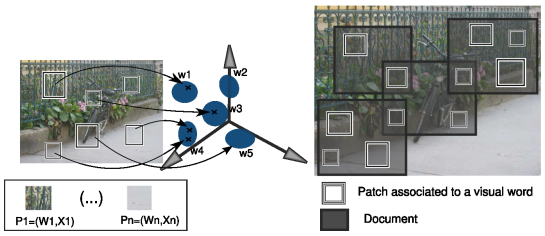

Figure 2: The visual vocabulary is obtained by vector quantizing a set of image patches descriptors.Images are modeled as sets of overlapping documents, each document being a set of patches.

an additional process, responsible for combining labels carried by patches into pixel hypotheses.

The top-down process embeds object models and uses them to obtain a global coherence, by combining local information provided by the bottom-up process. Most of the models previously used in this context cannot be used here, because of the strong variation of object's appearance. Geometric models such as the Pictorial Structure (Kumar et al., 2005) or the Implicit Shape Model (Leibe and Schiele, 2003) would require a huge number of training images in order to capture the large variability of appearance. Approaches based on characteristic edge patches (Borenstein et al., 2004) are only usable when object outlines are sufficiently stable. As a consequence, it appears that a more flexible model is required to address such object categories.

For the recognition of complex object categories, the *bag-of-words* model (Csurka et al., 2004) has been shown to be one of the most effective. It was inspired from text classification; the text framework becomes applicable for documents corresponding to images, once images have been transformed into sets of visual words. More recently, techniques based on latent aspects were developed on top of this unordered word based representation and were applied first for text classification (Griffiths and Steyvers, 2004), and then for image classification (Sivic et al., 2005). Such models are usually coming from the probabilist Latent Semantic Analysis (pLSA) (Hofmann, 2001) or its Bayesian form, the Latent Dirichlet Allocation (LDA) (Blei et al., 2002). Visual words are considered as generated from latent aspect (or topics) and images are combination of specific distribution of topics.

Using this latent aspect based framework for segmenting images is appealing for several reasons. First because object appearances (topics) can be automatically discovered and learned, limiting the amount of supervision required. Second, the flexibility of such a framework can handle large variations in appear-

The estimation is done according to the maximum likelihood criterion: We collect N images and observe the set of patches $(x_1, w_1), \ldots, (x_N, w_N)$. We want to compute θ and ϕ maximizing $P((x_1, w_1), \ldots, (x_N, w_N)|\theta, \phi, \alpha, \beta)$.

Since the integral of the observation generation probability (equation 1) makes the direct optimization of the likelihood intractable, we estimate variables of interest by an approximate iterative technique called Gibbs sampling.

During this process we estimate topic affectations (hidden variables of the model) jointly with θ and ϕ. θ and ϕ are never explicitly estimated, but instead the posterior distribution over the assignments of words to topics $P(z|w)$ is considered. This process has been proposed by (Griffiths and Steyvers, 2004) for the LDA model. Gibbs sampling works as follows: documents are initialized with equiprobable distribution over topics, then we iterate to estimate the posterior distribution $P(z|w)$.

In practice, the model includes only two topics, one for describing foreground patches, the other for background patches. However in a totally unsupervised framework, as we presented so far, foreground objects might not be automatically chosen as a topic of the model. The multi-documents model might capture other more frequent aspects of the image.

This is the reason why some extra supervision is added during the learning stage. A set of training images with possibly different levels of supervision is used to estimate foreground topic distribution over words. This is done with a standard LDA model (Blei et al., 2002) which is able to capture efficient topic distributions even with small supervision. This learned distribution is then used as a prior on the ϕ distribution for the test images.

It should be noted that for making the estimation possible we only process one test image at a time. We typically have thousands of documents per image. Processing all these images simultaneously would be infeasible. As a consequence, documents of different images become independent.

The hyper-parameters α and β play an important role as they allow to control topics and visual words distribution. For α, a small scalar value has been taken as in (Griffiths and Steyvers, 2004) in order to produce sparse and therefore specialized topics distribution. For β the knowledge acquired during the training stage has been used. We used values proportional to the number of patches affected to topics and words. This prior has large values as the knowledge acquired by LDA model is strong, but during the estimation the ϕ distribution can still be adapted to a particular test image.

2.3 From Patches to Segmentation

At the end of the estimation process, all patches have a probability of being generated by one of the foreground/background topics. These patches correspond to the squared sub-window's pixels used to build visual words. To compute the probability for a pixel p belonging to an object class (corresponding to topic z), we accumulate the knowledge on patches $\mathcal{P}$ containing the pixel. This is modeled by a mixture model, where weights (probability of a pixel to have been generated by a patch $P(p|\mathcal{P})$) are functions of the distance between the pixel and the center of the patch. We have $P(class(p) = z) \propto \sum_{\mathcal{P}_i \ni p} P(t_i = z) P(p|\mathcal{P}_i)$ where t_i stands for the topic of patch $\mathcal{P}_i$.

This can be seen as a summary of all labels provided for the same pixel. In regions where neighboring patches disagree, the confidence will be low; in contrast if neighboring patches agree, the probability for the pixel to belong to the object becomes higher.

3 EXPERIMENTS

We tried to evaluate the soundness of our method by comparing segmentation it produces with hand ground truth segmentation. We chose to use the bike class of the Graz-02 dataset[1] because of its high complexity. It contains images with high intra-class variability on highly cluttered backgrounds. The ground truth is available for 300 images. It is given in terms of pixel segmentation masks (one example is shown Figure 1). These masks will be used to evaluate the quality of our segmentation.

Methodology We have shown in section 2.3 that our algorithm computes the probability for each image's pixel to belong to an object of a given category. On the other hand, we know ground truth pixels labels, given by the provided segmentation masks. It is therefore natural to evaluate the performance by computing a ROC curve for each image. The ROC curve represent the true positive rate (TP) against the false positive rate (FP), *i.e.,* the rate of correct classification for the category of interest against the rate of object pixels misclassified. The true positive rate at equal error rate (EER) is the true positive rate at the curve point where $TP = 1 - FP$.

Experimental settings Patches are chosen at different scales according to a dense grid. This sampling is dense enough for all pixels to belong to several

[1] available at http://www.emt.tugraz.at/pinz/data/GRAZ_02/