A Bayesian Hierarchical Model for Learning Natural Scene Categories
Paper Critique
CS886 - Fall 2010
Karim Ali

1 Overview

The paper starts off by explaining why supervised learning and recognition of natural scene categories is no longer a desirable approach. The authors' motivation is that supervised learning depends on the manual annotation of intermediate properties of natural scenes to be able to come up with a categorization scheme for the images under analysis. Of course, manual image annotation is both tedious and expensive. In addition, expert-defined labels are some what arbitrary and possibly sub-optimal.

The main contribution of the paper is providing a Bayesian hierarchical model to learn and recognize natural scene categories without supervision. The authors extend the notion of a **texton** (i.e. codeword) from texture and material literature and apply it to images. The algorithm starts by manually assigning a category label to the image. The image is then divided into codewords that are added to the specified category. The process goes on until all the training images are analyzed. Then a k-means algorithm is executed over the codewords of each category to find out what are the codewords that represent each category the most. The algorithm can then be used to decide on the category of a given test image by comparing the likelihood of the image given each category based on the image codewords. The algorithm would then pick the correct category label by maximizing this likelihood.

2 Significance and Originality

The paper can be considered as an extension/application of the idea of using **textons**, in the context of natural scene categorization. The novelity of the approach is that it reduces the dependence on human effort compared to previous methods that follow a supervised learning approach.

The citation count of the paper (around 500) also suggest that the paper is heavily referred to in the field. This certainly shows how significant the contributions of the paper are.

3 Evaluation

The authors motivate their work by assuming that manual annotation of the intermediate properties of natural scenes is **some what arbitrary and possibly sub-optimal** (Section 1, page 2). However, there is no explanation for why this is the case. They do not provide references to back up this claim either. I can understand how this process might need extra work (mainly human effort), but I could not reach the conclusion that this manual intervention might lead to a sub-optimal solution.

The authors show in Figure 10 how the performance of their algorithm is affected by increasing the number of training samples. Nonetheless, they do not explain how well their algorithm will scale when used to categorize a very large database of images. In addition, the complexity of their algorithm is not discussed.

The good aspect of the experimental evaluation in the paper is that the authors admit that their algorithm poorly performs for certain categories of images (indoor images). However, they do not further explain why they think this is the case. I think it might be attributed to two factors:

1. Indoor scenes usually have more objects in them (and probably less light). Since they are using grayscale images, determining the boundaries for objects (therefore, affecting the value for each pixel) might not be as accurate as outdoor scenes.
2. The 4 indoor categories mentioned in the paper (bedroom, kitchen, office, livingroom) can have infinite designs, patterns (based on people's taste). Therefore, it is very hard to put all the designs of, for example, kitchens in one category, same goes for other indoor categories. It might be a good idea to further categorize those categrize based on the design (i.e. modern, classic, american, etc.).

4 Related Work

Although the core of the authors' algorithm is based on the key insights of previous work (Section 1, page 2), they provide a shallow discussion for such work. Moreover, they only compared their approach with previous approaches in Table 2 which was mentioned in the summary section of the paper (where they are not supposed to add new information to the paper).

5 Readability

The paper suffers from a poor choice of figure placements. For example, in the first paragraph in Section 2, the authors refer to Figure 4 which appears 4 pages later. In addition, Figure 7 is placed right under the heading of Section 4 (Results), which is confusing because I did not know why the flow of the text suddenly changed from explaining the algorithm to discussing the results. The authors mentioned in Section 2.1.3 that they will soon publish a technical note with the detailed derivations for their algorithm on their website. However, they did not provide a reference/link for that website. Finally, I could only catch one typo in Section 2.2; *Codeswords* should be *Codewords*.

6 Comments

Although the authors claim that the learning process is an unsupervised/automatic process, they still depend on human efforts to assign the category labels for the training images. This step is crucial for the algorithm, otherwise the algorithm will not be able to decide which set of codewords it should mark as **relevant**.

7 Suggestions

An improvement on the way of assigning a category label for the training images can be achieved by following the same logic the algorithm uses to generate the codebook. In other words, each training image can be divided into patches (i.e. codewords) and then a clustering algorithm (e.g. k-means) can be run over all the patches from all images to know the optimal number of clusters (i.e. categories) and their structure. However, I can see why the authors did not resort to that solution, since the clusters will not have descriptive labels for them as opposed to the case of having a human categories the training data.