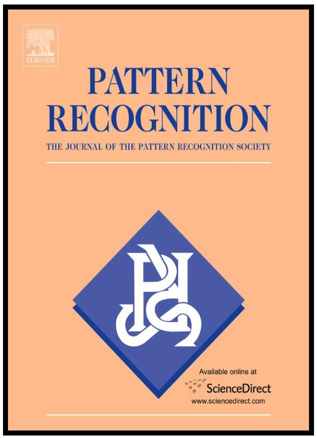

www.elsevier.com/locate/pr

Category co-occurrence modeling for large scale scene recognition

Xinhang Song, Shuqiang Jiang, Luis Herranz, Yan Kong, Kai Zheng

PII: S0031-3203(16)00040-6
DOI: http://dx.doi.org/10.1016/j.patcog.2016.01.019
Reference: PR5621

To appear in: *Pattern Recognition*

Received date: 30 July 2015
Revised date: 19 January 2016
Accepted date: 20 January 2016

Cite this article as: Xinhang Song, Shuqiang Jiang, Luis Herranz, Yan Kong anc Kai Zheng, Category co-occurrence modeling for large scale scene recognitior *Pattern Recognition,* http://dx.doi.org/10.1016/j.patcog.2016.01.019

Category co-occurrence modeling for large scale scene recognition

Xinhang Song[a], Shuqiang Jiang*[a], Luis Herranz[a], Yan Kong[b], Kai Zheng[c]

[a] *The Institute of Computing Technology of the Chinese Academy of Sciences, Beijing, China*
[b] *The Institute of Automation of the Chinese Academy of Sciences, Beijing, China*
[c] *School of Computer Science, Soochow University, China*

Abstract

Scene recognition involves complex reasoning from low-level local features to high-level scene categories. The large semantic gap motivates that most methods model scenes resorting to mid-level representations (e.g. objects, topics). However, this implies an additional mid-level vocabulary and has implications in training and inference. In contrast, the semantic multinomial (SMN) represents patches directly in the scene-level semantic space, which leads to ambiguity when aggregated to a global image representation. Fortunately, this ambiguity appears in the form of scene category co-occurrences which can be modeled a posteriori with a classifier. In this paper we observe that these patterns are essentially local rather than global, sparse, and consistent across SMNs obtained from multiple visual features. We propose a co-occurrence modeling framework where we exploit all these patterns jointly in a common semantic space, combining both supervised and unsupervised learning. Based on this framework we can integrate multiple features and design embeddings for large scale recognition directly in the scene-level space. Finally, we use the co-occurrence modeling framework to develop new scene representations, which experiments show that outperform previous SMN-based representations.

Keywords:
scene recognition; co-occurrence modeling; semantic space; feature embedding; multiple feature combination; large scale image recognition

1. Introduction

Visual understanding is essentially a complex process of abstraction, from purely local visual information to abstract semantic entities such as objects and scenes. The conventional visual recognition strategy has consisted of extracting local visual features[1], and encoding them into a global representation of the image using some variation of the bag-of-words (BOW) model[2, 3, 4, 5]. While very effective for object recognition, scenes often require more abstract representations composed of other lower-level semantic entities, such as objects or themes, which appear in the scene in a loose layout, contrasting with the much more rigid structure of parts in objects. Thus, it may be difficult to model scene categories directly from low-level visual descriptors, due to a larger semantic gap.

An intermediate abstraction level can represent the presence of local concepts (e.g. *sky*, *rock*, *street, car*)[6] (see Figure 1a and c), and then scene categories (e.g. *coast*, *inside city*, *kitchen*) are recognized based on this intermediate representation. Thus, the semantic gap is reduced by performing the abstraction gradually in two steps. Note that now several problems arise related with the mid-level representation. First, it requires selecting a set of mid-level vocabulary of local concepts. In addition, training intermediate classifiers[6, 7, 8] requires images with regions annotated with these mid-level concepts, which is much more costly than annotating just one scene label. Some works avoid this problem by using latent topics for the intermediate representation[9, 10, 11], where topics are discovered during learning. However, these methods often have limited performance due to poor supervision[12], and are often based on complex generative models difficult to scale to large datasets.

Alternatively, the *semantic multinomial (SMN)*[13] represents the probability that a given patch belongs to each scene category, being a local but not mid-level representation. Image-SMNs are obtained by aggregating patch-

*Corresponding author.
Email addresses: xinhang.song@vipl.ict.ac.cn (Xinhang Song), sqjiang@ict.ac.cn (Shuqiang Jiang*), luis.herranz@vipl.ict.ac.cn (Luis Herranz), kongyanwork@gmail.com (Yan Kong), zhengkai@suda.edu.cn (Kai Zheng)