Signal Processing 112 (2015) 129–136

Contents lists available at ScienceDirect

Signal Processing

journal homepage: www.elsevier.com/locate/sigpro

Semantic embedding for indoor scene recognition by weighted hypergraph learning

Jun Yu [a], Chaoqun Hong [b], Dapeng Tao [c,d,*], Meng Wang [e]

[a] School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou, China
[b] Faculty of Computer Science, Xiamen University of Technology, Xiamen, China
[c] Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, China
[d] The Chinese University of Hong Kong, Hong Kong, China
[e] School of Computer Science and Information Engineering, Hefei University of Technology, Hefei, Anhui 230009, China

ARTICLE INFO

Article history:
Received 27 January 2014
Received in revised form
29 July 2014
Accepted 30 July 2014
Available online 9 August 2014

Keywords:
Semantic information
Indoor scenes classification
Attributes learning
Weighted hypergraph

ABSTRACT

Conventional methods for indoor scenes classification is a challenging task due to the gaps between images' visual features and semantics. These methods do not consider the interactions among features or objects. In this paper, a novel approach is proposed to classify scenes by embedding semantic information in the weighted hypergraph learning. First, hypergraph regularization is improved by optimizing weights of hyperedges. Second, the connectivity among images is learned by statistics of objects appearing in the same image. In this way, semantic gap is narrowed. The experimental results demonstrate the effectiveness of the proposed method.

1. Introduction

A critical ability for an agent is to recognize surrounding environments. The holistic concepts activate top-down knowledge that leads the visual analysis in scene interpretation tasks. To achieve this purpose, the Gist features are proposed in [1], which describes the spatial envelope of a scene. Many scene recognition approaches have been proposed based on this general idea, and they work well for outdoor scenes recognition. However, these methods perform badly in recognizing indoor scenes. The local feature-based approaches describe scene images with detected interest points based on specified descriptors. For instance, bag-of-features is adopted to describe a scene image. The local detectors include Harris's corner detector [5], scale invariant feature transform (SIFT) [6], Harris–Laplace detector [7]. These detectors find points and regions in exploring basic features in scene images. Since local feature-based methods neglect spatial information, some methods have been proposed to combine global Gist information with local object information to achieve better performance on recognizing indoor scenes. For instance, Quattoni et al. [2] exploited local and global discriminative information to propose a prototype based indoor scene recognition model that can successfully combined these two sources. Yu et al. [3] described scene images by finding a group of features and explored their complementary characteristics. Pandey and Lazebnik [30] adopted weakly supervised object localization in scene recognition. Wu and Rehg [31] proposed a new visual descriptor named CENTRIST for recognizing scene categories. Besides, attributes learning [8] has been commonly adopted in describing the semantics of objects or images through vectors of semantic attributes. Farhadi et al. [9] proposed a method, which adopted a group of visual semantic attributes to describe the objects. In addition, several approaches have been proposed to describe the entire image through visual semantic attributes. Vogel and Schiele [10] adopted visual attributes to describe the semantics of local regions in

* Corresponding author.
E-mail address: dapeng.tao@gmail.com (D. Tao).

http://dx.doi.org/10.1016/j.sigpro.2014.07.027
0165-1684/

outdoor scenes, and integrated these local semantics into a global image representation. Torresani et al. [11] adopted the outputs of a large number of object category classifiers to describe images and showed good performances in both image classification and image retrieval. Furthermore, Li et al. [4] proposed a high-level image representation, called the Object Bank, in which the image was described as a scale-invariant response map of a large number of pre-trained generic object detectors. The indoor recognition model was built based on high-level image representation. Ramalingam et al. [35] considered the problem of detecting junctions and using them for recovering the spatial layout of an indoor scene. Pero et al. [36] proposed a method for understanding the 3D geometry of indoor environments while simultaneously identifying objects in the scene. Espinace et al. [37] adopted object classifiers to associate low-level visual features to objects, and at the same time, they used contextual relations to associate objects to scenes.

Except for providing novel feature descriptors, varied learning algorithms have been proposed in the field of classification, such as decision tree [12], k-nearest neighbor, and support vector machines (SVMs) [13,14]. These supervised learning algorithms were conducted based on labeled data. Since image labeling is labor intensive and time consuming, image classification frequently suffers from the insufficiency problem of training data. Transductive learning [15–18] is a popular approach to handle this problem. It explores not only labeled but also unlabeled data. Among existing transductive learning methods, graph-based learning method achieves promising performance. This type of learning is operated on a graph, in which vertices are samples and edge weights indicate the similarity between two samples. A regularization framework is typically formulated by integrating two terms, i.e., a regularizer term that forces the smoothness of the classification results on the graph and a loss term that ensures a low level of training error. These methods intrinsically differ in the graph construction and the definition of the loss term. According to elaborations in [19], graph-based learning methods consider only pairwise relationships between two samples, and they neglect the relationship in higher order. Modeling the high-order relationship among samples will significantly improve classification performance. Hypergraph learning [20,32–34,39–41] have successfully addressed this problem. Unlike a graph that has an edge between two vertices, a set of vertices is connected by a hyperedge in a hypergraph. Each hyperedge is assigned a weight. In hypergraph learning, the weights of the hyperedges are empirically set according to certain rules. For example, the weight of each hyperedge is simply set to 1 in [20]. In [21], the weight of a hyperedge is calculated by summing up pairwise affinities within the hyperedge. Therefore, weighting or selecting hyperedges will help improve classification performance.

Scenes are always consists of many objects and their layouts are quite complicated. Some scenes may contain similar objects. However, the features describing the scenes are usually independent and cannot reflect the above information. Conventional semi-supervised feature selection methods are almost either filter-based methods including spectral analysis [22], or search based methods including forward search [23]. These methods did not consider the interactions among features and the interaction between the feature selection heuristics and the corresponding classifier. Instead, we propose a novel hypergraph learning based feature selection method for indoor scene classification, which is named High-level Attributes Modeling (HAM). In the paper, the proposed semi-supervised feature selection method works in an embedded way: the feature selection process is integrated to the semi-supervised classifier by taking advantage of hypergraph regularization. Therefore, it takes good care of the correlation among features and the integration between the features and the semi-supervised classifiers. Furthermore, the semantic connectivity of objects is also taken in to consideration. In our method, the objective function is composed of three loss terms: (1) labeling of samples that are highly connected in the hypergraph; (2) labeling of training samples with known outcomes; and (3) weighting of the hyperedges associated with the connections in the hyperedge interaction network. We obtain the solution by minimizing the weighted sum of the three loss terms.

The rest of this paper is organized as follows. Section 2 introduces some related works on attributes learning and hypergraph learning. Section 3 specifically describes the proposed semantic indoor scene classification by weighted hypergraph learning. Section 4 shows experiments on practical indoor scene datasets. Section 5 concludes the paper.

2. Related work

In this part, we detail related work in the methods of attributes learning used in scene classification. Besides, we provide some preliminaries of hypergraph learning, which are used in our proposed method.

2.1. Attributes learning

In attributes learning, attributes are visual qualities of objects. Images treat attributes as patterns of image segments, repeatedly sharing some characteristic properties [24]. In this way, the correspondence between specific attributes and image classes can be found, and it narrows the semantic gap. Attribute learning has been widely used in image classification.

The typical application of attributes learning in scene classification is using objects as attributes. The appearance of particular objects indicates the type of current scene. A representation which is predictive of typicality is the so-called “attribute score” [25]. That is, items that are most typical have attributes that are very common in the category. In this approach each attribute is weighted in order to take into account their respective importance for the category. Vogel and Schiele proposed a semantic typicality measure for natural scene classification [26]. In their method, it was the local semantic concepts that act as scene category attributes. Li et al. described complex real-world scenes by collecting the responses of many object detectors [27,28].