Neurocomputing 148 (2015) 578–586

Contents lists available at ScienceDirect

Neurocomputing

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

Scene and place recognition using a hierarchical latent topic model

Jinfu Yang [a,*], Shanshan Zhang [a], Guanghui Wang [b], Mingai Li [a]

[a] Department of Control Science and Engineering, Beijing University of Technology, Beijing 100124, PR China
[b] Department of Electrical Engineering & Computer Science, University of Kansas, Lawrence, KS 66045-7608, USA

ARTICLE INFO

Article history:
Received 13 January 2014
Received in revised form
2 May 2014
Accepted 3 July 2014
Communicated by: Yongdong Zhang
Available online 15 July 2014

Keywords:
Probabilistic topic model
Highlighted Latent Dirichlet Allocation
Fast variational inference
Place recognition

ABSTRACT

Place classification and object categorization are necessary functions of vision-based robotic systems. In this paper, a novel latent topic model is proposed to learn and recognize scenes and places. First, each image in the training set is characterized by a collection of local features, known as codewords, obtained by unsupervised learning, and each codeword is represented as part of a topic. Then, the codeword distribution of detected local features from the training images is learned by performing a *k*-means algorithm. Next, a modified Latent Dirichlet Allocation model is employed to highlight the significant features (i.e., the codewords with higher frequency in the codebook). The *Highlighted Latent Dirichlet Allocation* (HLDA) improves the efficiency of learning procedure. Finally, a fast variational inference algorithm for HLDA is proposed to reduce the computational complexity in parameter estimation. Experimental results using natural scenes, indoor and outdoor datasets show that the proposed HLDA method performs better than other counterparts in terms of accuracy and robustness with the variation of illumination conditions, perspectives, and scales. The Fast HLDA is order of magnitudes faster than the HLDA without obvious loss of accuracy.

1. Introduction

Localization and mapping are critical underlying factors for mobile robot navigation in large environments. The process of simultaneously tracking the position of a mobile robot and building a map of the environment is known as Simultaneous Localization and Mapping (SLAM). As described in [1], accurate localization is a necessary condition of building a reliable map; in return, a precise map is helpful for accurate localization.

There are different types of sensor modalities for localization, such as sonar, laser range finders, and vision. Many early robot localization approaches are based on laser or sonar sensors, as vision is more computational expensive. For example, Nourbakhsh et al. [2] adopted sonar configuration to allow DERVISH robot to navigate in office buildings. However, sonar inevitably suffers from both an array of inherent weaknesses and a lack of richness. Due to the fact that vision can provide semantic information of a scene through understanding its visual appearance, vision-based localization is becoming more and more popular in autonomous mobile systems.

Place recognition is prerequisite for mobile robot localization, as a mobile robot system must be able to categorize where it is when it navigates in an environment. In recent years, studies on place recognition for robot localization and navigation have received considerable attention. Torralba et al. [3] presented a context-based method for place recognition. The algorithm utilized global image features to predict the scene that can be used as a prior for the local detectors. Ulrich and Nourbakhsh [4] used a panoramic vision system to explore the environment in which the color images can be categorized in real-time based on nearest-neighbour learning, image histogram matching, and a simple voting scheme. Gaspar et al. [5] proposed a visual-based navigation method for a mobile robot in indoor environments, using a single catadioptric camera. The algorithm combines appearance based methods and visual servoing upon some environment features to travel long distances, which does not require knowledge of the exact position of the robot, and adopts eigenspace as a representation of images to achieve local and precise navigation. Se et al. [1] used scale-invariant image features as landmarks in unmodified dynamic environments for mobile robot localization and mapping. The 3D landmarks are localized and robot egomotion is estimated by matching them, taking into account the feature viewpoint variation. Tamimi and Zell [6] used Kernel Principal Component Analysis (KPCA) to extract features from the visual scene of a mobile robot for localization. It is applied only to local features so as to guarantee better computational performance as well as translation invariance.

Visual sensing can provide the robot extensive information about its environment. However, the visual information tends to be very complex and difficult to analyze due to changes of

* Corresponding author. Tel.: +86 10 67396309.
E-mail address: jfyang@bjut.edu.cn (J. Yang).

http://dx.doi.org/10.1016/j.neucom.2014.07.005
0925-2312/

illumination and transformation. These problems bring many challenges to the representation of vision-based localization. Local descriptors computed for interest regions have proven to be very successful in object recognition, image retrieval, and robot localization. The main idea is to detect image regions covariant to a class of transformations, and these regions are then used as support regions to compute invariant descriptors. Different descriptors have been proposed in the literatures [7–11]. In previous work, SIFT is one of the most widely used local descriptors [12–14]. In addition, taxonomic methods have also been proposed to address the problem of the changes of illumination and transformation. Pronobis et al. [15] used rich global descriptors and support vector machines as discriminative classifier to cope with illumination and pose changes in place recognition. Luo et al. [16] proposed a discriminative incremental learning approach using a version of the fixed-partition incremental SVM to place recognition, which allows controlling the memory growth as the system updates its internal representation and achieved close recognition performances as other batch algorithms.

Topic models, such as *Latent Dirichlet Allocation* (LDA) proposed by Blei et al. [17] in recent years, have been widely used in computer vision field. Sivic et al. [18] used probabilistic Latent Semantic Analysis (pLSA) and LDA models to detect objects from a collection of images. Russell et al. [19] applied LDA model to find and segment visual topics with an unlabeled collection of images, in which it can obtain object classes automatically. Fei-Fei and Perona [20] used LDA to learn and recognize natural scene categories, and reported satisfactory categorization performance on a large set of 13 categories of complex scenes. Zhu et al. [21] proposed a hierarchical latent topic model in which the codewords were generated by sparse coding and represented with n-dimensional vectors in R^n. Yang et al. [22] provided the Latent Dirichlet Allocation based method for place recognition to cope with the problem of the variation of illumination and transformation. In these applications, LDA model was adopted to cluster low-level visual words into topics with semantic meanings, and demonstrated promising performance in scene and object classification.

However, the traditional LDA model is complex and inefficient since it could not effectively take advantage of the significant features (we refer to a codeword with higher frequency in the codebook as a significant feature) due to chaotic distribution of the frequency vector representation extracted from the images. In this paper, we propose a Highlighted Latent Dirichlet Allocation (HLDA) based on a modified LDA model by explicitly introducing a feature function to highlight the significant features. The algorithm significantly improves the efficiency through learning the significant features and intermediate topics. In order to solve the problem of expensive computation, we also propose a fast learning algorithm for HLDA (Fast HLDA) by applying a fast variational inference [23] to the LDA model. Extensive experiments show that the proposed HLDA model performs better than the traditional LDA model in coping with the problem of changes on illumination, perspective and scale, and the Fast HLDA can significantly accelerate the recognition speed with reasonable recognition performance. The main contributions of this paper are as follows:

- The proposed HLDA model is more efficient than the LDA model by introducing a feature function to take account of the significant features.
- The Fast HLDA method significantly speed up the recognition results without obvious loss of accuracy.
- The Fast HLDA model is more suitable for real-time applications in robot environment perception than other approaches.

The rest of the paper is organized as follows. We introduce the Highlighted Latent Dirichlet Allocation model and its corresponding fast learning algorithm for scene and place recognition in Section 2. Section 3 describes the databases for experiments and the experimental results. Some conclusions are discussed in Section 4.

2. Highlighted Latent Dirichlet Allocation model

In this section, we describe the Highlighted Latent Dirichlet Allocation model and its corresponding learning algorithm, as well as the fast parameter estimation algorithm of HLDA model for place recognition. The procedure of place recognition using the HLDA model is summarized as follows:

First, the local features are extracted from the training images, and a discrete set of clusters, denoted as “codewords”, is formed by clustering all the local features. Each image of the training set can be represented as a word-frequency vector by assigning each feature to the closest cluster. Then, the HLDA model is used to learn topic distributions of the images. Finally, the unknown test images can be recognized according to the similarity of topic distributions.

2.1. Representation of images

It has been shown that local features are more robust to local occlusions and spatial variations than global features [4]. In this paper, we represent each image as a collection of patches, denoted as “codewords” learnt by unsupervised learning. Firstly, a large number of key points are detected from each of the training images using DoG detector [12], and each key point can be represented as a 128-dim SIFT vector. Then, a collection of clusters, commonly known as “codewords”, is obtained by performing the k-means algorithm to cluster all the SIFT features. The set of all codewords is called a codebook. Finally, each SIFT feature of an image is assigned to the closest cluster (“codeword”), as a result, the image can be represented as a frequency vector based on these “codewords”. As described in [17,20], we define the following terms:

- A *codeword* is the basic element, defined to be a patch membership from a codebook. This is analogous to a “word” in document processing.
- An *image* is a sequence of M codewords denoted by $\mathbf{x}=(x_1,x_2,...,x_M)$, where x_m is the mth word in the sequence. This is analogous to a “document” in document processing.
- A *category* is a collection of N images denoted by $\mathbf{D}=\{\mathbf{x}_1,\mathbf{x}_2,...,\mathbf{x}_N\}$. This is analogous to a “corpus” in document processing.

As described above, in the procedure of codebook generation, each SIFT feature extracted from an image is assigned to certain codeword through the clustering process and the image can be represented by the distribution of the codewords. Intuitively, the codewords with higher frequency have more similar feature patches than others. And these similar feature patches will be more helpful to classification tasks than other features assigned to the codewords with lower frequency. Fig. 1 shows the codewords distributions of indoor images used in our experiments (see the details in Section 3). The left one is the codewords distribution of one image randomly selected from the dataset. The right one is the codewords distribution of 50 images randomly selected from the dataset. As shown in Fig. 1, some codewords have higher word-frequency than others. The local features assigned to the codewords with higher frequency are referred as significant features. In Section 3, we introduce a feature function to the traditional LDA model, coming up with HLDA model, to recognize scenes and places using significant features.