Pattern Recognition 45 (2012) 1671–1683

Contents lists available at SciVerse ScienceDirect

Pattern Recognition

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

Feature fusion within local region using localized maximum-margin learning for scene categorization

Jianzhao Qin [a,b,*], Nelson H.C. Yung [a,b]

[a] Laboratory for Intelligent Transportation Systems Research, Department of Electrical & Electronic Engineering, The University of Hong Kong, Pokfulam Road, Hong Kong SAR, China

[b] Center for Information Security and Cryptography, Department of Computer Science, The University of Hong Kong, Pokfulam Road, Hong Kong SAR, China

ARTICLE INFO

Article history:
Received 8 November 2010
Received in revised form
24 August 2011
Accepted 20 September 2011
Available online 19 October 2011

Keywords:
Scene categorization
Image recognition
Feature fusion
Similarity-metric learning

ABSTRACT

In the field of visual recognition such as scene categorization, representing an image based on the local feature (e.g., the bag-of-visual-word (BOVW) model and the bag-of-contextual-visual-word (BOCVW) model) has become popular and one of the most successful methods. In this paper, we propose a method that uses localized maximum-margin learning to fuse different types of features during the BOCVW modeling for eventual scene classification. The proposed method fuses multiple features at the stage when the best contextual visual word is selected to represent a local region (hard assignment) or the probabilities of the candidate contextual visual words used to represent the unknown region are estimated (soft assignment). The merits of the proposed method are that (1) errors caused by the ambiguity of single feature when assigning local regions to the contextual visual words can be corrected or the probabilities of the candidate contextual visual words used to represent the region can be estimated more accurately; and that (2) it offers a more flexible way in fusing these features through determining the similarity-metric locally by localized maximum-margin learning. The proposed method has been evaluated experimentally and the results indicate its effectiveness.

1. Introduction

Scene categorization concerns with automatically labeling or classifying a given image to a specific scene category (e.g., coast, forest, highway, office, kitchen, street, sitting room, etc.). Automatic categorization of an image to a scene can be used to manage picture libraries and retrieve images from Internet or in image databases [1–3]. Moreover, being able to recognize the scene category of a place is vital for an intelligent vehicle or robot to locate its position and take appropriate actions under different scenes [4,5]. Furthermore, scene categorization can also provide critical contextual information to many computer vision tasks, such as object recognition, image segmentation, etc. [6,7]. It is also essential for an intelligent video surveillance system in the future, which can help define what abnormal conditions are for detection and tracking (e.g., abnormal objects, abnormal behaviors). For instance, a person running can be considered as abnormal in a 'street' scene, but normal in a 'sport ground' scene.

In the early research work for scene categorization, many global feature based methods [2–4,8] have been proposed. In these methods, an image is taken as a whole, and the distribution(s) of color [2,3] and/or texture [2] and/or gradients [4,8] over the entire image region is (are) employed to describe the scene image. They have achieved certain success, especially in separating outdoor scenes from indoor scenes. However, when they are employed to classify scenes that have similar global properties (e.g., bedroom vs. sitting room; open country vs. coast), they often result in poor success rate. In recent years, local semantic feature based methods [9–13] become more popular because of its robustness towards occlusions, illumination variations and slight geometric deformation. They model a scene image by the co-occurrences of a number of visual components or the co-occurrences of a certain number of visual topics (intermediate representation). One of the most popular and successful models is called 'the bags of visual words' (BOVW) [11–13]. Many variants of this model have been proposed [14–18]. In [14–16], latent variables, which can be taken as a group of visual words, are learned using the techniques called probabilistic Latent Semantic Analysis (pLSA) [19] or Latent Dirichlet Allocation [20]. In [17], Lazebnik et al. proposed a spatial pyramid matching method, which matches the distributions of visual words at different spatial resolution between paired images then used it as a similarity measurement. Qin and Yung [21,18] proposed a scene categorization method based on contextual visual words, in which the contextual information from neighbor

* Corresponding author at: Center for Information Security and Cryptography, Department of Computer Science, The University of Hong Kong, Pokfulam Road, Hong Kong SAR, China. Tel.: +852 28578414; fax: +852 25598738.
E-mail addresses: jzhqin@eee.hku.hk (J. Qin), nyung@eee.hku.hk (N.H.C. Yung).

0031-3203/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2011.09.027

regions and the regions from coarser scales are included to describe the region of interest. We called this model a bag-of-contextual-visual-word model (BOCVW). Li et al. [22] proposed a contextual bag-of-word model for visual recognition including scene categorization. The basic idea of this method and its performance is similar to our proposed contextual visual words in [21,18] but with different implementation; and they have added context from semantic relation. (For the SCENE-15 dataset (Section 9.1), we achieved 85.16% accuracy rate using our proposed contextual visual words while they achieved 85.1% accuracy rate.)

To further enhance the performance of the BOVW based methods, algorithms have been proposed to fuse different types of features [23–25] in the field of object recognition. The methods proposed by Varma et al. [23] and Bosch et al. [25] create several spatial pyramid representations of the BOVW model that correspond to different types of features, and then a multiple-kernel learning (MKL) approach is employed to learn the linear weighting of different kernels that correspond to different spatial pyramid representations. Gehler et al. [26] proposed an enhanced multiple kernel method called 'LPboost' (linear programming boosting), which allows the support vector machines' (SVM) parameters trained for different types of features to be different.

From the fusion procedures of the aforementioned methods, we can see that the fusion of the BOVW models of different features occurs after the image have been represented by the BOVW models. In other words, such feature fusion is carried out globally. Fig. 1 depicts the differences between the globally feature fusion method and the proposed feature fusion method within local regions. One of the weaknesses of the globally feature fusion method is that the ambiguity of the local patches caused by the single feature representation would unlikely be resolved by other globally introduced features. This is because as the other features are globally coded, they do not provide information about a specific local image region. For example, just based on the SIFT feature, a region of an image that represents the grass land of the 'Open country' scene may be incorrectly represented by the visual word that represents a part of the sea water from the 'Coast' scene, which may result in incorrect classification of the image from 'Open country' to 'Coast'. Although, combining BOVW model of SIFT and BOVW model of color feature may alter the final classification result ('Open country' may have more green regions while 'Coast' may have more blue regions). However, in some cases, the image of 'Open country' may have an equally large region that represents the blue sky, which is similar to sea water in color. Conversely, 'Coast' may show a large region of green trees. As such, 'Open country' scenes and 'Coast' scenes can share very similar BOVW model of color feature. The other weakness of the global approach is that they can only give fixed weightings to the considered features. However, in practice, in order to differentiate a region from other regions, we may give more weight to a particular feature. For instance, in order to differentiate the shore in the 'Coast' scene from the grass land in the 'Open country' scene, we can put more weights on the color feature. However, if the target is to differentiate the grass land in the 'Open country' scene from the trees in the 'Forest' scene, gradient and texture features should be given more weighting values instead.

Some works in other fields (e.g., finger vein recognition and target recognition) have proposed some approaches to fuse local feature with global feature using canonical correlation analysis (CCA) [27] or kernel canonical correlation analysis (KCCA) [28]. However, since the image in these two works is not modeled by BOVW/BOCVW model, those methods cannot be applied to

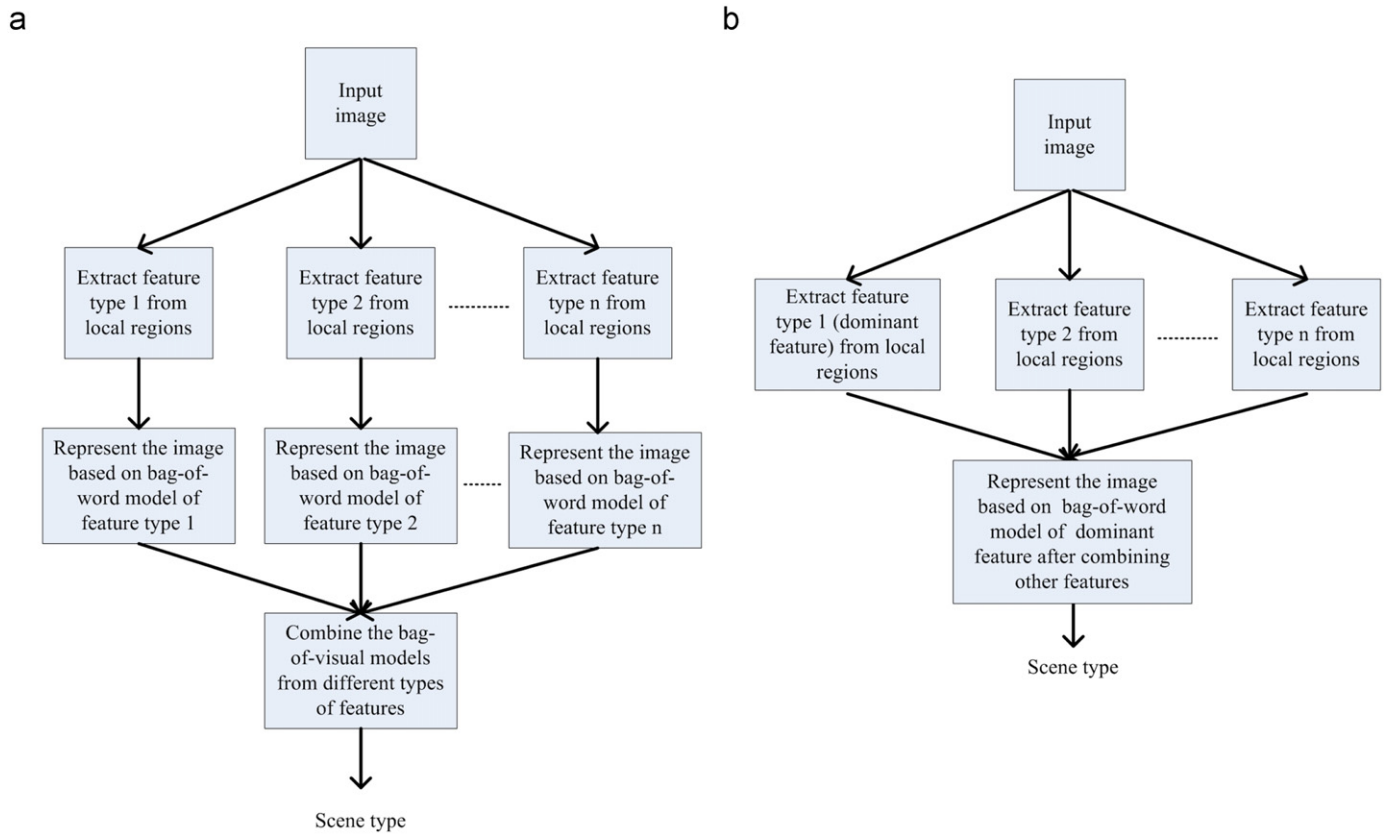

Fig. 1. (a) Globally feature fusion; (b) Feature fusion within local regions based on dominant feature.