Signal Processing 93 (2013) 1597–1607

Contents lists available at SciVerse ScienceDirect

Signal Processing

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

Fast multi-view segment graph kernel for object classification

Luming Zhang, Mingli Song *, Xiao Liu, Jiajun Bu, Chun Chen

College of Computer Science, Zhejiang University, China

ARTICLE INFO

Article history:
Received 30 October 2011
Received in revised form
9 May 2012
Accepted 11 May 2012
Available online 23 May 2012

Keywords:
Object classification
Graph kernel
Multi-view segment graph

ABSTRACT

Object classification is an important issue in multimedia information retrieval. Usually, we can use images from multiple views (or multi-view images) to describe an object for classification. However, two issues remain unsolved. First, exploiting the spatial relations of local features from different view images for object classification. Second, accelerating the multi-view object classification process. To solve these two problems, we propose fast multi-view segment graph kernel (FMSGK). Given a set of multi-view images for an object, we segment each of them in terms of its color intensity distribution. And inter- and intra-view segment graphs are constructed to describe the spatial relations of the segments between and within view images respectively. Then, these two types of graphs are integrated into a so-called multi-view segment graph. And the kernel between objects is computed by accumulating all matchings' of walk structures between their corresponding multi-view segment graphs. Since computing the kernel directly is highly time-consuming, an accelerating algorithm is derived. Finally, a multi-class support vector machine (SVM) (Duda et al., 2000 [19]; Wang et al., 2008 [32]; Dai and Mai, 2012 [6]) is trained based on the computed kernels for object classification. The experimental results on three data sets validate the effectiveness of our approach.

1. Introduction

Object classification is an important issue for many multimedia applications, such as scene recognition and surveillance. To classify an object, we can either describe the object by a single-view image and classify this image into an object category or, describe the object by a set of multi-view images and classify the set of images into an object category. Obviously, the second scheme is more robust because it contains richer information for object classification. For instance, in the single-view case, some discriminative information may be occluded, while in the multi-view case, the occluded information is recovered. However, it is still a challenge to deal with the multi-view object classification successfully due to two factors: on one hand, the components in the multi-view images and their spatial relations are complex and unstable, which makes it difficult to extract features discriminative enough for classification; on the other hand, the huge number of components and their bilateral relations bring challenges to computer to be processed efficiently. Therefore, more discriminative and efficient features are becoming more and more important for multi-view object classification.

In the evolution of image analysis, many features have been proposed and they can be categorized into two groups: global features and local features. Global features, e.g., eigenspace [38], represent the entire image by a single vector and are hence tractable for conventional classifiers, such as SVM. Usually, we use dimensionality reduction algorithms [39,1,25] to increase the discrimination of this single vector. However, global features are sensitive to occlusions and clutters, which result in poor

---

* Corresponding author.
E-mail addresses: zglumg@cs.zju.edu.cn (L. Zhang), brooksong@zju.edu.cn (M. Song), ender_liux@cs.zju.edu.cn (X. Liu), bjj@cs.zju.edu.cn (J. Bu), chenc@cs.zju.edu.cn (C. Chen).

0165-1684/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.sigpro.2012.05.012

classification accuracy. In contrast to global features, local features, e.g., scale invariant feature transformation (SIFT) [20], are extracted at interest points and are robust to image deformations. Different images may produce different number of local features. In order to be tractable for conventional classifiers, these local features are often integrated into an orderless bag-of-features representation. Unfortunately, as a non-structural representation, the bag-of-features representation ignores the geometric property of images, i.e., the spatial of local features, which prevents it from being discriminative.

In order to exploit the spatial relations of local features for multi-view object classification, several methods have been proposed. Lazebink et al. [15] developed the spatial pyramid matching (SPM) by partitioning an image into increasingly fine grids and computing histograms of local features inside each grid cell. However, SPM requires nonlinear classifier, which is highly time-consuming, to achieve good classification accuracy. Towards a more efficient classification model, Yang et al. [36] proposed SC-SPM, which encodes image local descriptors based on sparse coding [16]. Further in [28], Wang et al. proposed LLC-SPM, which improves conventional SPM by utilizing the locality constrains to encode each image local descriptor. Inspired by the local coordinate coding (LCC) [37], Zhou et al. [40] applied a so-called super-vector encoding of image local descriptors to represent each image. Empirical results show that [36,28,40] perform well under a linear SVM. Although these SPM variants have many advantages in object classification, they are designed for single-view object classification only. For instance, they fail to exploit spatial relations of segments between different view images. In [5], latent dirichlet allocation (LDA) [3] is used to model the geometric property of the scene images. Specifically, each scene image is represented by a set of codewords, which are independently generated by the corresponding latent topics. However, as empirically demonstrated in [23], LDA affects adversely on scene classification. In [26,7], each image is modeled as a tree and image matching is formulated into tree matching. Unfortunately, compared to general graphs, the capability of modeling regions' relations by trees is limited Felzenszwalb et al. [8] modeled the relation of different parts of an object as a spring; however [8] relies heavily on optimal background subtraction. In [12], Hedau et al. defined a new measure of pairwise regions based on the overlaps between regions; but just region overlaps are too simple to capture the complicated spatial relations between regions. Keselaman et al. [14] defined a graph, called least common abstraction (LCA), to represent the spatial relations of components of an object; however, LCA cannot be an output to conventional classifier, e.g., SVM [19], directly. By exploring the complementary property of different types of features, multi-view spectral embedding [33] and multi-view stochastic neighbor embedding [34] obtain a physical meaningful embedding of the multi-modal features. However, [33,34] fail to consider the geometric information of each view image.

To solve or at least reduce the aforementioned problems, an efficient kernel named FMSGK, which exploits the spatial relations of local features between and within view image, is proposed for object classification. Given a set of multi-view images for an object, we first segment each view image into a number of regions in terms of its color intensity distribution. And two types of graphs, the inter- and the intra- view segment graphs are constructed to model these multi-view images. Then, we integrate these two types of graphs into a so-called multi-view segment graph. Finally, by constructing a product graph, the kernel between objects is computed efficiently by accumulating all matchings of walk structures between the corresponding multi-view segment graphs. Based on the computed kernel, a SVM classifier is trained for object classification. A graphical illustration of the flowchart of our approach is given in Fig. 1.

The contributions of this paper are as follows: (1) FMSGK, a new method to build the representation of multi-view images, is presented for object classification; (2) inter- and intra- view segment graph to represent the spatial relations of segments between and within multi-view images; (3) an efficient kernel between multi-view segment graphs.

The rest of this paper is organized as follows. In Section 2 we present the construction of the inter- and intra- view segment graph. Section 3 introduces the integration of multi-view segment graph and Section 4 presents an efficient multi-view segment graph kernel. Section 5 presents the algorithm of our approach. We give the empirical results in Section 6 and Section 7 concludes.

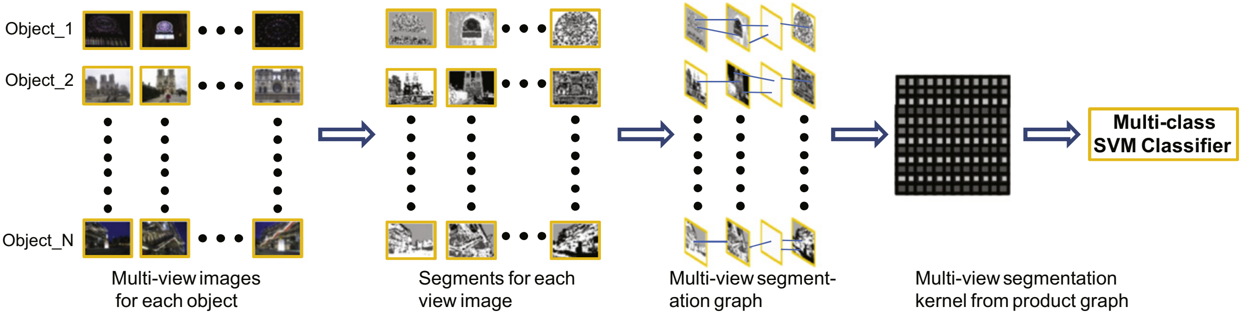

Fig. 1. The flowchart of our approach.