J. Vis. Commun. Image R. 25 (2014) 1299–1307

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

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

Bird breed classification and annotation using saliency based graphical model

Chao Huang *, Fanman Meng, Wang Luo, Shuyuan Zhu

School of Electronic Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China

ARTICLE INFO

Article history:
Received 19 April 2013
Accepted 8 May 2014
Available online 20 May 2014

Keywords:
Graphical model
Saliency
Annotation
Classification
Bird breed
Local context
GMS
Segmentation

ABSTRACT

Due to the variations among the birds, bird breed classification is still a challenging task. In this paper, we propose a saliency based graphical model (GMS), which can precisely annotate the object on the pixel level. In the proposed method, we first over-segment the image into several regions. Then, GMS extracts the object and classifies the image based on the local context, global context and saliency of each region. In order to achieve a high precision of classification, we use SVM to classify the image based on the features of the annotated bird. Finally, we employ posterior probability distribution obtained by GMS and SVM to perform the image classification. Experiments on the Caltech-UCSD Birds dataset show that the proposed model can achieve better results compared with existing bird breed classification methods based on graphical model.

1. Introduction

With the wide spread of camera, the images on the internet grow quickly. Due to the large time consuming of manual work, automatic image classification and annotation becomes more important and necessary to support scene understanding and image retrieval. A variety of image classification methods have been developed [1–3]. In general, traditional methods of image classification can be divided into three steps. The first step is to extract features [4–6] from the images. Then, the bag-of-words (BOW) [7] is used to represent the image based on the clustering algorithm. Finally, the category of the image is obtained by using the classifier such as LDA [8], SVM [9]. In recent years, fine-grained image classification has attracted lots of attentions which brings a challenging task for the traditional methods.

For fine-grained image, there are several objects which are similar with each other. To classify the fine-grained image, we need the details of the objects which are easy to be mixed up with the noise induced by the background. As shown in the first row of Fig. 1, the background induces noise for classification. In order to solve this problem, we argue and demonstrate that using features extracted from the object will enhance the performance of fine-grained image classification. In the second row of Fig. 1, we show the annotated results which indicate the good performance of our proposed method.

In this paper, we focus on classifying and annotating the fine-grained bird image (first reported in [32]). Bird breed is used since bird is one of the most photographed animal. There are many photos of birds in the internet, which include all kinds of birds with different pose, color and shape. The success of bird classification can help the research of species identification.

Bird breed classification is a very challenging and difficult task due to the variation of shape, color and texture. To precisely annotate the bird on the pixel level and classify the bird breed, we construct a novel graphical model considering saliency (GMS). In this model, we use the local context, global context and saliency information to annotate the image on the pixel level. To enhance the performance of classification, we employ SVM to classify the image using the features extracted from the annotated bird. Then, the posterior probability distribution obtained by GMS and SVM is employed to obtain the category of the image. Experiments evaluated on the Caltech-UCSD Birds dataset demonstrate the efficiency of our proposed method.

There are mainly two contributions in this paper. Firstly, we construct a novel graphical model, which can annotate the fine-grained image on the pixel level and classify the image; Secondly, a new method of over-segmenting the image is proposed, which uses the strong boundary to correct the error regions.

This paper is organized as follow. Section 2 reviews the related work briefly. Section 3 introduces the framework of our proposed

* Corresponding author. Fax: +86 028 61830064.
E-mail address: huangchao_uestc@yahoo.cn (C. Huang).

http://dx.doi.org/10.1016/j.jvcir.2014.05.002
1047-3203/

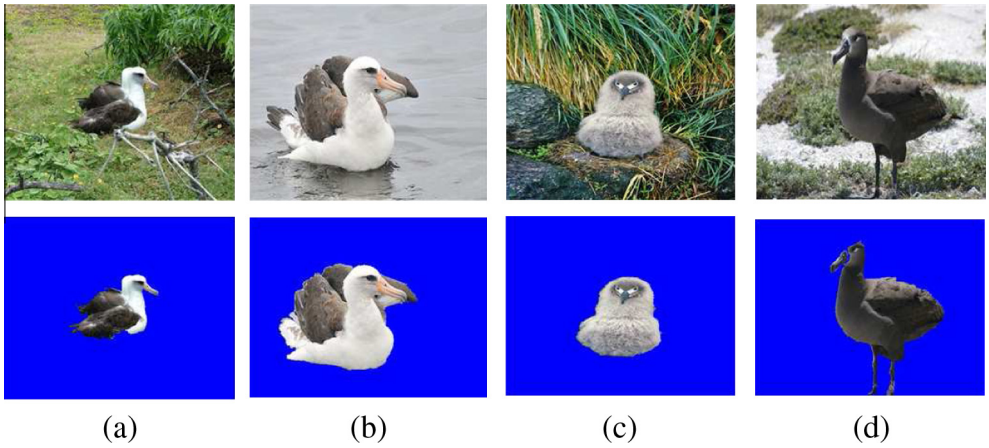

Fig. 1. The first row shows background induces noise which will mix up with important details for bird classification. The objects extracted by our graphical model are shown in the second row.

method. Experiment on bird breed classification will be presented in Section 4, which is followed by the conclusion of our work.

2. Related work

Visual recognition attracts much attention recently as a result of the increasing demand for classifying and annotating image. Image classification and annotation using topic model brings out a great deal of researches due to its good performance. One of the most important model called Latent Dirichlet Allocation (LDA) [8] was firstly used for text analysis. The good performance of text analysis demonstrated its efficiency and the basic idea of LDA is to cluster co-occurring words into topics.

However, there are several problems if we directly use the language model to solve the visual problem. First, there are not words for images while text documents are composed of discriminant words naturally. In order to fill their gap, features are extracted from patches of the images and quantified to visual words [3]. These patches can be evenly scattered or around interesting points which can be found by using SIFT [6]. By using this method, images can be represented with a set of visual words. Secondly, LDA treats a document as a bag of words, so the spatial and temporal structure among visual words are ignored. In order to use these information, many researchers have proposed a variety of algorithms [2,10–13] based on LDA. Generally speaking, these methods can be roughly divided into the following two groups.

The first group uses local context to construct topic model. Wang and E. Grimson [14] proposed spatial LDA that added spatial information. Based on the assumption that the visual words near in space should have larger probability to be assigned to the same topic, Gaussian distribution was used to analyze the spatial relationship between visual words and the referring neighbors. Better performance can be obtained by the method in [14] compared with LDA. In the work of Cao and Li [1], spatial-LTM was proposed. The approach [1] can be divided into three phases. Firstly, the over-segmented regions [15] were obtained based on super-pixel algorithm. Secondly, features extracted from the patches of each region were quantized to the visual words. Finally, topics were assigned to each region which assumed that the visual words in the same over-segmented region should share the same topic. The approach [1] enhanced the classification accuracy greatly and can be used for image classification and annotation via supervised or unsupervised way. Based on this work [1], Li et al. [16] proposed a hierarchical generative model which added the tags' information. There are two kinds of tags in this model, one is visual (e.g. human) and the other is un-visual (e.g. wind). After the learning phase, the model [16] can classify the category of image, segment the objects and annotate the image with a number of tags. It is the first model which can solve these problems together and sets a great milestone for image understanding using the generative model.

In the second group, global context is used for model construction. Niu et al. [3] proposed a generative model which used the location information. In the model [3], it assumed that the location information can be very useful for annotation. For instance, the visual words on the top of the image were more likely to be assigned to sky. But the local context [14,1] was not used in the model [3]. So Niu et al. proposed a CA-TW model [17] which combined the local context with global spatial layout of object. Several experiments demonstrated the advantage of this model. But if spatial layout of elements in different scenes varies widely, the performance of this model will drop dramatically.

In addition, there are some approaches focused on improving the performance of image classification which considered spatial saliency information. For instance, Sharma et al. [18] proposed a method to learn the classifier considering the spatial saliency information [19–21]. In their work, the saliency map is used to weight the visual features.

As pointed out in [22], it is helpful to utilize part location in fine-grained image classification. In the approach, exemplar-based geometric and appearance models are built for dog breed. The experiment of classifying dog demonstrated its efficiency.

3. Proposed framework

Motivated by the success of topic model [16,1,3], we construct a model considering saliency information, the appearance of each region and spatial structure of patches. Saliency [18], local context [1] and global context [3] are integrated into a unified framework to learn a reliable classifier. The proposed framework of bird classification and annotation can be divided into four phases.

In the first phase, we over-segment the image into several regions and obtain the saliency map. Then, GMS annotates and classifies the image using the features and saliency information of each region. In order to further improve the classification accuracy, SVM is employed to obtain the posterior probability of category using the features extracted from the annotated bird. Finally, we combine the results of SVM and GMS. MAP measure is used to determine the category of the image. The framework of our proposed method is shown in Fig. 2.

3.1. Over-segmenting the image

To use the local context, the image is over-segmented to several regions assuming that patches of the same region share the same