J. Vis. Commun. Image R. 27 (2015) 44–56

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

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

Object recognition via contextual color attention ☆

Jie Zhu [a,b], Jian Yu [a,*], Chaomurilige Wang [a], Fan-Zhang Li [c]

[a] Beijing Key Lab of Traffic Data Analysis and Mining, School of Computer and Information Technology, Beijing Jiaotong University, Beijing, China
[b] Department of Information Management, The Central Institute for Correctional Police, Baoding, China
[c] School of Computer Science and Technology, Soochow University, Suzhou, China

ARTICLE INFO

Article history:
Received 11 December 2013
Accepted 1 January 2015
Available online 9 January 2015

Keywords:
Object recognition
Color attention
Discriminative color
Strong patch
Weak patch
False weak patch
Contextual color attention
Optimal threshold of contextual color attention

ABSTRACT

Visual attention is effective in differentiating an object from its surroundings. Color is used to guide attention via a top-down category-specific attention map in the top-down color attention (CA) method. To uniformly highlight the entire object, our color attention map is reconstructed based on the estimated object patches. The object patches consist of strong patches and false weak patches whose contextual color attention values are beyond the optimal threshold of class-specific contextual color attention. The color attention map constructed by the object color histogram is then used to weight the local shape for object recognition. Extensive experiments are conducted to show that our method can provide state-of-the-art results on several challenging datasets.

1. Introduction

Image classification is an important topic in computer vision. Many methods rely on local features to represent the local appearances [1,2] of images, but the number of extracted local features is often large. To overcome this problem, focus is placed on the most relevant features rather than all the features.

Object recognition in computer vision is the task of finding and identifying objects in an image or video sequence. Visual attention is effective in differentiating an object from its surroundings. For humans, attention is facilitated by a retina that has evolved a high-resolution central fovea and a low-resolution periphery. Visual attention guides this anatomical structure to important parts of the scene to gather more detailed information, but the computational mechanisms underlying this guidance remain unclear [3]. Attention maps are widely used in many computer vision applications [4–8], but few studies focus on their usefulness for object categorization. Bag-of-words (BOW)-based [50] image representation is one of the most successful approaches to object recognition, whereby images are represented by a histogram of visual words. Khan et al. [6] used top-down color attention (CA) to modulate the impact of the shape words in an image during the histogram construction. The color attention of a patch was dependent on the occurrence frequency of the patch color in a category; therefore, object patches with different colors had different attention values. In our previous work [23], an object was recognized as a collection of class-specific discriminative colors. As a result, the attention map could not highlight object patches whose colors were not discriminative. To uniformly highlight the entire object, we have proposed a method to estimate the object patches and assign the same attention value to these patches in the attention map.

The object patches are divided into two types: strong patches and false weak patches. The strong patches include the class-specific discriminative colors, while the other patches are the weak patches. Some of the weak patches are also on the object and are called false weak patches. A diagram of our method is shown in Fig. 1. First, the local shape and features are extracted from the images. Next, we find the strong patches whose colors are discriminative. We then compute the optimal threshold of the class-specific contextual color attention to distinguish the false weak patches from the weak patches. We subsequently update the attention map using the class-specific object color histogram. Finally, the attention maps are used to modulate the weights of BOWs, and the final image representation is obtained by concatenating the class-specific histograms.

In our previous work [23], we selected the class-specific discriminative colors based on the ranking of mutual information

☆ This paper has been recommended for acceptance by Yehoshua Zeevi.

* Corresponding author.
E-mail addresses: arthurzhujie@gmail.com (J. Zhu), jianyu@bjtu.edu.cn (J. Yu).

http://dx.doi.org/10.1016/j.jvcir.2015.01.003
1047-3203/

and the inter-class color dissimilarity. To identify false weak patches, we propose the use of contextual color attention to distinguish false weak patches from other weak patches. If the contextual color attention value of a weak patch is beyond the optimal threshold of the class-specific contextual color attention, we consider this weak patch to be a false weak patch. The shapes of strong patches and false weak patches are assigned a uniform high weight for the final image representation, whereas other patches are assigned lower weights.

An example of our attention maps is illustrated in Fig. 2. Here, the original images are shown in Fig. 2(a)–(c) are the attention maps computed by the CA method and our method, respectively. It is obvious that our attention maps highlight more object patches than do the CA attention maps. For example, blue is the discriminative color in the Chelsea category (first row), and blue patches are the strong patches. All the blue patches are assigned high attention values. In contrast, many object patches are assigned low attention values by the CA method, including the false weak patches on the logos and outfits. Similar cases can also be found in the Tigerlily category (second row), where orange patches are the strong patches. The attention map of the CA can only highlight the strong patches, and false weak patches, such as black spots, are not highlighted as strong patches. In our proposed method, the attention maps can uniformly highlight both of these patches.

Two points differentiate our attention map from other attention maps. First, our attention map is only used for image representation, and we are interested in the object patches rather than an accurate outline of the object. The images are treated as collections of independent patches, which are selected using keypoint detectors because of its convenience as an image representation similar to BOW. Some methods choose to use segmentation methods [9,33] or salient pixels [11,16] to find the accurate object regions. As our method is based on CA, we use the patches obtained by DoG [38] and the Harris-Laplace keypoint detector [39] to find the object regions. The patches with high attention values are the attended patches and are likely to be the object patches. Our color attention map consists of many patches; therefore, we cannot obtain an accurate location of the object, unlike some color-based segmentation methods. However, our interest is all the patches on the object, not all the pixels.

Second, we use color to modulate the importance of shape features, which are represented by SIFT [38]. In contrast, color is only used to estimate better object regions in most existing saliency maps [9,11,16].

The paper is organized as follows. In Section 2, the related work is introduced. In Section 3, we review the top-down color attention method for object recognition. In Section 4, we compute a new attention map based on the contextual color attention to modulate the impact of the shape words in the image for object recognition. The experimental setup and results are presented in Section 5, and Section 6 provides the conclusions of the study.

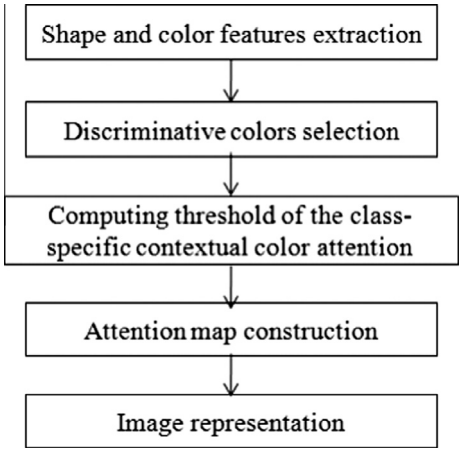

Fig. 1. Flow diagram of our method.

2. Related works

Visual attention has been extensively studied in many fields, including psychology and computer vision. Previous studies have shown that visual attention follows two main procedures: bottom-up and top-down. In bottom-up attention, the selection of visual attention depends on low-level features of images [9–11], and the term "salient" is often considered in the context of bottom-up computations [3]. In contrast, proponents of top-down attention claim that object information dominates over attention selection because humans focus on familiar object entities rather than regions with salient low-level features [12,6]. The work presented in our paper utilizes the top-down visual attention mechanism, and our goal is to recognize the object category.

Saliency maps are widely used for object detection. Cheng et al. [11] proposed a regional-contrast-based saliency extraction algorithm, which simultaneously evaluates global contrast differences and spatial coherence. Wang et al. [13] partitioned image into patches and estimated the saliency in each patch relative to a large dictionary of un-annotated patches from the remaining images. Achanta et al. [10] introduced a method for salient region detection that outputs full-resolution saliency maps having well-defined boundaries of salient objects. Li and Ngan [9] introduced a method to detect co-saliency from an image pair that may have some objects in common. These methods improve the precision of the object recognition process but do not consider the contextual information. Chen et al. [22] used saliency for the hierarchical clustering of the encoded local features. They supposed that the saliences of different objects were different but those of pixels of the same object were the same, meaning that saliency could be used to distinguish different objects.

Using contextual information to enable a better image understanding, whereby the contextual information can be obtained from the nearby image data or objects, has received significant attention in recent research. Parikh et al. [14] explored the use of context to determine which low-level appearance cues in an image are salient or representative of an image's contents. Perko and Leonardis [15] presented a framework for visual context-aware object detection based on a sparse coding of contextual features based on geometry and texture. Goferman et al. [16] presented a saliency detection method based on four basic principles of context-aware saliency, which were supported by psychological evidence, and they suggested that the regions surrounding the foci can convey the context and draw our attention; thus, they are salient.

Top-down factors play an important role in directing attention. Khan et al. [6] constructed a top-down color attention map according to the probability of the category given the corresponding color word, and the attention map was used to weight the shape feature. Yang and Yang [17] proposed a novel top-down saliency model that jointly learned a conditional random field (CRF) and a discriminative dictionary. Kanan et al. [18] presented a "top-down" knowledge (appearance)-based saliency model derived in a Bayesian framework. In [19], a model of attention guidance based on a global scene configuration was proposed, and it was shown that the statistics of low-level features across the scene image determine where a specific object should be located. Bottom-up attention information and top-down attention information were combined in [20,21] for the computation of the saliency map.

Our method can also be recognized as a feature fusion method. In [31], shape words and color words were weighted by a logistic regression-based fusion method. Gehler and Nowozin [32] studied