Neurocomputing 172 (2016) 9–18

Contents lists available at ScienceDirect

Neurocomputing

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

Automatic tag saliency ranking for stereo images

Yang Cao [a], Kai Kang [a], Shijie Zhang [a], Jing Zhang [a], Zengfu Wang [b,*]

[a] Department of Automation, University of Science and Technology, China
[b] Institute of Intelligent Machines, Chinese Academy of Science, China

ARTICLE INFO

Article history:
Received 25 November 2013
Received in revised form
23 July 2014
Accepted 24 September 2014
Available online 9 May 2015

Keywords:
Tag ranking
3D saliency detection
Stereo image
Multi-instance learning

ABSTRACT

With the rapid advances in 3D capture and display technology, tag ranking for stereo images will be a potential application on web image retrieval. Directly extending the existing approaches to stereo images is problematic as it may neglect the representative 3D elements. In this paper, a novel automatic tag saliency ranking algorithm for stereo images is presented. Specifically, a novel method of interacting with stereo images is proposed to segment the two images into regions simultaneously. Then tags annotated on the image-level are propagated to the region-level via an improved multi-instance learning algorithm. In the next, a new 3D saliency detection algorithm is proposed using occlusion cues along with the contrast in color and disparity. And finally, tags are re-ranked according to the saliency values of the corresponding regions. Moreover, to evaluate the performance of our approach, an annotated stereo image dataset is set up. The experimental results on this dataset demonstrate the effectiveness of our approach.

1. Introduction

Stereo images and videos can provide an immersive 3D viewing experience, and accordingly have been introduced into many multimedia applications, such as 3D TV, free-viewpoint video, immersive teleconferencing and video game systems. With the development and maturation of 3D capture and display technology in recent years, an abundance of stereo images is being generated. Retrieving stereo images from enormous collections becomes an important research topic and practical problem. Although the technique of Content-based Image Retrieval (CBIR) has been widely investigated [1,2], directly extending the existing 2D image retrieval methods to stereo images is problematic as it may neglect the representative elements in the extra dimension of depth. Consequently, it will lead to some misleading results. Typical examples are shown in Fig. 1. The image shown in Fig. 1 is annotated with tags "water, dog, grass, bird". The image shows both high visual representative to the tag "water" and "dog" from the perspective of 2D viewing. However, from the perspective of depth perception, the 3D salient regions of the image in general attract more attentions. So the tag "dog" should be ranked in front of "water" to facilitate the image retrieval.

The depth information can be estimated according to the stereo correspondence between the left and the right view of the stereo images. However, it is not given even with stereo correspondence. Occluded regions and areas off the side of an image have no corresponding regions in the other view. Moreover, since a perfect stereo matching algorithm has yet to be invented, errors will exist in the estimated depth map. Due to this, the depth cues cannot be directly used for image retrieval.

This paper focuses on the task of tag ranking for stereo images, a fundamental step for retrieving stereo images on a large amount of collections. In developing the tag ranking algorithm, a primary objective is to rank the existing tags according to the relevance scores to the visual content of the given image [3]. This means measuring the visual representative degree of the tags with respect to the corresponding contents of the stereo images. Tag ranking, which aims to rank image tags according to the semantic relevance with respect to the image content, has emerged as a hot topic recently [4]. Existing tag ranking methods can be roughly classified into two categories, i.e., tag relevance ranking and tag saliency ranking [5]. Tag relevance ranking has attracted more attentions among the earliest works. Li et al. [6] introduced a neighborhood voting method which learns tag relevance by accumulating votes from visual neighbors. Recently, Liu et al. [7] applied Kernel Density Estimation (KDE) to estimate the relevance of each tag individually, and further performed a random-walk based refinement to boost tag ranking performance. Tag saliency ranking, in which the annotated tags are re-ordered according to the saliency property of the corresponding visual content, is firstly proposed in [8]. This method integrates visual attention model to measure the importance of regions of the given image. Therefore, it can provide more comprehensive information and is much consistent with the human perception.

Intuitively, tag saliency ranking method is more suitable for stereo images, since most of stereo images contain distinct objects

* Corresponding author.
E-mail address: zfwang@iim.ac.cn (Z. Wang).

http://dx.doi.org/10.1016/j.neucom.2014.09.097
0925-2312/

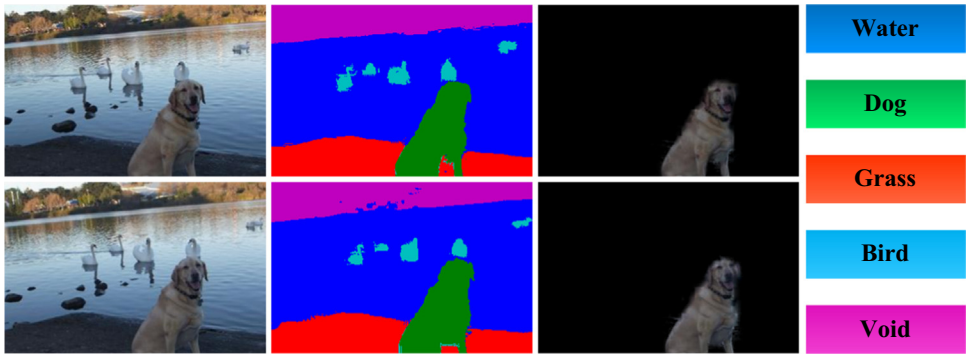

Fig. 1. The first column is an exemplar stereo image pair. The second column is the consistent segmentation result, where each region is labeled with the assigned colors acting as tags list shown in the last row. The third column shows the 3D salient regions of the given stereo images.

to achieve better 3D viewing experience. Therefore, the saliency values of the objects in the images can be utilized as the importance measures to facilitate the tag ranking. Accordingly, we propose in this paper a tag saliency ranking approach for stereo images. Specifically, the two stereo images are segmented simultaneously in the first. Then tags annotated on the image-level are propagated to the region-level via an improved multi-instance learning algorithm. In the next, the 3D saliency map extracted from stereoscopies is used to measure the importance of regions of the given image. And finally, tags are ranked according to the saliency values of the corresponding regions. To evaluate the performance of our approach, we conduct experiment on our own stereo image dataset. Our dataset is composed of 424 photographs of 16 object classes. We labeled each image pair with the assigned colors acting as tags into the list of object classes. The experimental results demonstrate the effectiveness of our approach. To the best of our knowledge, this is the first time that we propose a tag ranking approach for stereo images.

The rest of the paper is organized as follows. Related work is presented in Section 2. The proposed tag saliency ranking approach is presented in Section 3. The experimental results are shown in Section 4 and conclusions are made in Section 5.

2. Related work

There has been significant previous work in tag refinement, including tag annotation [9–13], tag ranking [4,5], and tag denoising [14]. Although these previous techniques could be applied to multi-view images simply by processing the images, such an approach would neglect the valuable features [15,16] to represent the depth perception or other useful cues. Feng et al. [17] proposed a generic framework for stereo image retrieval where the disparity features extracted from the stereo pairs offer complementary clues to refine the retrieving results provided by the visual features. However, since a perfect stereo matching algorithm has yet to be invented, errors will exist in the estimated disparity. Directly using the disparity features would cause mislead results.

Visual attention model has been proved that it can improve the tag ranking performance for images with multiple tags. Feng et al. [8] introduced the concept of tag saliency, where visual saliency is used to investigate the ranking order information of tags with respect to the given image. Then the annotated tags are re-ordered according to the saliency property of the corresponding visual content. This approach is based on the observation that users usually pay more attention to the visual salient regions of image. Therefore, it is much consistent with the human perception. Moreover, this approach can provide more comprehensive information when an image is relevant to multiple tags, such as those describing different objects in the image. Particularly, tag saliency ranking approach is well suited for stereo images. To achieve better 3D viewing experience, stereo images tend to contain distinct objects instead of cluttered scene. Therefore, the corresponding salient properties can clearly emphasize the content of the given images.

In this paper, we aim at extending the tag saliency ranking approach to stereo images. To accomplish this task, the annotated tags should be firstly propagated from image-level to region-level. This indeed includes two steps, the stereo image segmentation and the region labeling. There is only limited literature about stereo image segmentation. Although the previous work on single images segmentation [18–20] could be applied to stereo image pairs simply by joining the two images into an image volume, such an approach cannot handle the fact that corresponding pixels may have large disparities. A recent study [21] proposes details of interactive selecting objects in the stereo image pairs via graph cut. However, this method requires a lot of user input. Levin et al. propose a closed-form solution for natural image matting [22]. This method can effectively extract a foreground object from the input image with a small amount of user input. In our work, we extend this matting method to stereo images by introducing stereo consistency on the mattes of each view. The region labeling is indeed a weakly supervised learning problem since tags are usually associated with images instead of individual regions. Multi-instance learning (MIL) has been proved to be an effect method to model such problems [23]. In MIL, an individual example is called an instance and a set of instances is called a bag. Training labels are associated with bags rather than instances. A bag is labeled positive if at least one of its instances is positive; otherwise, the bag is negative [24]. A lot of work has employed MIL for content-based image retrieval and automatic image annotation [23,25–27]. In this paper, we also utilize MIL to accomplish the label propagation from image-level to region-level. In our work, each segmented region is treated as an instance and the segmented regions which form an image are grouped as a bag of instances. The tags are initially labeled on the bag-level instead of instance-level. Then, an image is annotated by keyword w if at least one region in the image has the semantic meaning of w.

3D saliency detection technique plays an important role in our task. Although a wide variety of computational methods have been developed to detect saliency using 2D features extracted from image, such as color, shape, orientation, and texture, these methods show their limits in 3D saliency detection, when salient objects do not exhibit visual uniqueness with respect to one or more of these features. Li and Ngan applied a linear combination of the single-image and the multi-image saliency map to detect co-saliency from an image pair [28]. Their method exploited the local appearance, e.g., color and texture properties to construct a co-multilayer graph, which described the similarity between two SISMs. Since their method did not consider the geometry correspondence between image pair, its performance for dealing with the stereoscopies was not very well. Niu et al. computed the contrast in disparity to identify the saliency region