Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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

The segmented and annotated IAPR TC-12 benchmark ☆

Hugo Jair Escalante [a,*], Carlos A. Hernández [a], Jesus A. Gonzalez [a], A. López-López [a], Manuel Montes [a], Eduardo F. Morales [a], L. Enrique Sucar [a], Luis Villaseñor [a], Michael Grubinger [b]

[a] National Institute of Astrophysics, Optics and Electronics, Department of Computational Sciences, Luis Enrique Erro # 1, Tonantzintla, Puebla, 72840, Mexico
[b] Victoria University, Australia School of Computer Science and Mathematics P.O. Box 14428, Melbourne, Vic. 8001, Australia

ARTICLE INFO

Article history:
Received 30 April 2008
Accepted 10 March 2009
Available online 20 March 2009

Keywords:
Data set creation
Ground truth collection
Evaluation metrics
Automatic image annotation
Image retrieval

ABSTRACT

Automatic image annotation (AIA), a highly popular topic in the field of information retrieval research, has experienced significant progress within the last decade. Yet, the lack of a standardized evaluation platform tailored to the needs of AIA, has hindered effective evaluation of its methods, especially for region-based AIA. Therefore in this paper, we introduce the segmented and annotated IAPR TC-12 benchmark; an extended resource for the evaluation of AIA methods as well as the analysis of their impact on multimedia information retrieval. We describe the methodology adopted for the manual segmentation and annotation of images, and present statistics for the extended collection. The extended collection is publicly available and can be used to evaluate a variety of tasks in addition to image annotation. We also propose a soft measure for the evaluation of annotation performance and identify future research areas in which this extended test collection is likely to make a contribution.

1. Introduction

The task of automatically assigning semantic labels to images is known as automatic image annotation (AIA), a challenge that has been identified as one of the *hot-topics* in the new age of image retrieval [1]. The ultimate goal of AIA is to allow image collections without annotations to be searched using keywords. This type of image search is referred to as annotation-based image retrieval (ABIR) and is different from text-based image retrieval (TBIR), which uses text that has been manually assigned to images [2].

Despite being relatively new, significant progress has been achieved in this task within the last decade [2–9]. However, due to the lack of a benchmark collection specifically designed for the requirements of AIA, most methods have been evaluated in small collections of unrealistic images [3–9]. Furthermore, the lack of region-level AIA benchmarks lead to many region-level methods being evaluated by their annotation performance at image-level, which can yield unreliable estimations of localization performance [5,10]. Recently, the combination of automatic and manual annotations has been proposed to improve the retrieval performance and diversify results in annotated collections [11]. However, the impact of AIA methods on image retrieval has not yet been studied under realistic settings.

Thus, in order to provide reliable ground-truth data for benchmarking AIA and the analysis of its benefits for multimedia image retrieval, we introduce the segmented and annotated *IAPR TC-12* benchmark. This collection is a well-established image retrieval benchmark comprising 20,000 images manually annotated with free-text descriptions in three languages [12]. We extended this benchmark by manually segmenting and annotating the entire collection according to a carefully defined vocabulary. This extension allows the evaluation of further multimedia tasks in addition to those currently supported.

Since the IAPR TC-12 is already an image retrieval benchmark, the extended collection facilitates the analysis and evaluation of the impact of AIA methods in multimedia retrieval tasks and allows for the objective comparison of CBIR (content-based image retrieval), ABIR and TBIR techniques as well as the evaluation of the usefulness of combining information from diverse sources.

1.1. Automatic image annotation

Textual descriptions in images can prove to be very useful, especially when they are complete (i.e. the visual and semantic content of images is available in the description), with standard information retrieval techniques reporting very good results for image retrieval [13,14]. However, manually assigning textual information to images is both expensive and subjective; as a consequence, there has recently been an increasing interest in performing this task automatically.

☆ We thank editors and anonymous reviewers for their useful comments that helped us to improve this paper. This project was partially supported by CONACyT under project grant 61335.

* Corresponding author. Fax: +52 222 266 31 52.
E-mail addresses: hugojair@ccc.inaoep.mx (H.J. Escalante), michael.grubinger@gmx.at (M. Grubinger).

1077-3142/$ - see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.cviu.2009.03.008

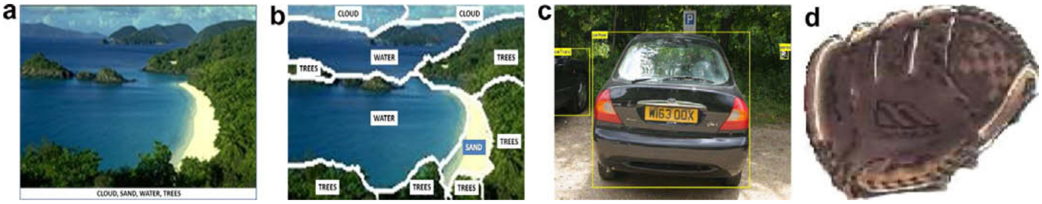

Fig. 1. Sample images for image-level AIA and region-level AIA.

There are two different approaches to AIA: at image-level and at region-level. Image-level AIA assigns labels to the image as a whole, not specifying which words are related to which objects within that image, while region-level AIA provides annotations at region-level within each image, or in other words, a one-to-one correspondence between words and regions. Hence, the latter approach offers more information (e.g. spatial relationships) that can be used to improve annotation and retrieval performance. Note that any region-level annotation is an image-level annotation. This work only considers region-level AIA.

Fig. 1 depicts sample images for both approaches, taken from three related tasks (from left to right): image-level annotation and region-level annotation (from the Corel subsets [8]), object detection (from the PASCAL VOC-2006 data set [15]) and object recognition (from the Caltech256 data set [16]).

The AIA challenge has been approached with semi-supervised and supervised machine learning techniques [3–11,17,18]. Supervised methods have thereby reported better results than their semi-supervised counterparts [9,17,18], but they also require a training set of region-label pairs, compared to semi-supervised methods that only need weakly annotated images. Hence, there exists a compromise between retrieval results and annotation effort, and both methods thereby offer complimentary benefits. An important feature of the extended IAPR TC-12 benchmark is that it supports both methods.

1.2. AIA and object recognition

Region-level AIA is often regarded as an object recognition task. Yet, this is true only to some extent and, therefore, object recognition benchmarks are not well-suited for AIA. In both, AIA and object recognition tasks, the common challenge is to assign the correct label to a region in an image. However, in object recognition collections the data consists of images whereby the object to recognize is often centered and occupies more than 50% of the image (see Fig. 1, rightmost image); usually, no other object from the set of objects to be recognized is present in the same image. In region-level AIA collections, in contrast, the data consists of annotated regions from segmented images, where the target object may not be the main theme of the image and many other target objects can be present in the same image (see Fig. 1).

Another difference lies in the type of objects to recognize. The objects in object recognition tasks are often very specific entities (such as cars, gloves or specific weapons), while the concepts in region-level AIA are more general (e.g. buildings, grass and trees). These differences are mainly due to the applications they are designed for: object recognition is mostly related with surveillance, identification, and tracking systems, whereas AIA methods are designed for image retrieval and related tasks.

1.3. Evaluation of region-level AIA methods

Duygulu et al. [4] adopted an evaluation methodology that has widely been used to assess the performance of both region-level and image-level AIA techniques, whereby AIA methods are used to label regions of images in a test set. For each test image, the assigned region-level annotations are merged to obtain an image-level annotation, which is then compared to the respective image-level ground truth annotation. To evaluate localization performance, the results for 100 images [4] (and 500 images respectively in subsequent work [5]) were analyzed. However, this analysis only gives partial evidence of the true localization performance as in most cases, when AIA methods are evaluated, this type of evaluation is not carried out [4–7]. Moreover, the performance of AIA methods is measured by using standard information retrieval measures such as precision and recall. While this choice can provide information of image-level performance, it cannot allow for the effective evaluation of localization performance. For example, consider the annotations shown in Fig. 2: according to the aforementioned methodology, both annotations have equal performance, however, the annotation on the right shows a very poor localization performance. A better and simpler methodology would be to average the number of correctly labeled regions [8,10]; this measure would adequately evaluate the localization performance of both annotations.

Yet, the image-level approach has been adopted to evaluate AIA methods regardless of their type (i.e. supervised or semi-supervised) or their goal (i.e. region-level or image-level) [4–7], due to the lack of benchmark collections with region-level annotations. In this paper, we therefore describe a segmented and annotated benchmark collection that can be used to evaluate AIA methods.

2. Related work

A widely used collection to evaluate AIA is the Corel data set [1,4–6,8,10,17]; it consists of around 800 CDs, each containing 100 images related to a common semantic concept. Each image is accompanied by a few keywords describing the semantic or visual content of the image. Although this collection is large enough for obtaining significant results, it exhibits several limitations that make it an unsuitable and unrealistic resource for the evaluation of image retrieval algorithms: (i) most of its images were taken in difficult poses and under controlled situations; (ii) it contains the same number of images related to each of the semantic concepts, which is rarely found in realistic collections; (iii) its images are annotated at image-level and therefore cannot be used for region-level AIA; (iv) it has been shown that subsets of this database can be tailored to show improvements [19]; (v) it is copyright protected, hence its images cannot be freely distributed among researchers, which makes the collection expensive; and (vi) it is no longer available.

In alternative approaches, computer games have been used to build resources for tasks related to computer vision. ESP [20], for example, is an online game that has been used for image-level annotation of real images. The annotation process ensures that only correct[1] labels are assigned to images. Unfortunately, the amount of data produced is considerably large, the images are annotated at image-level and the data is not readily available. Peekaboom

[1] The “correctness” is thereby measured by the agreement of annotators.