Computer Vision and Image Understanding xxx (2009) xxx–xxx

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 xxxx

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

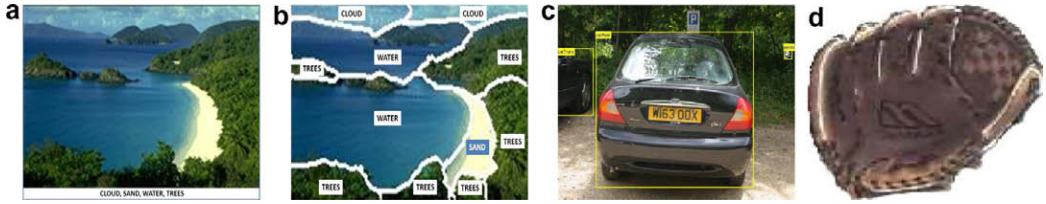

Fig. 1. Sample images for image-level AIA and region-level AIA.

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.

Please cite this article in press as: H.J. Escalante et al., The segmented and annotated IAPR TC-12 benchmark, Comput. Vis. Image Understand. (2009), doi:10.1016/j.cviu.2009.03.008

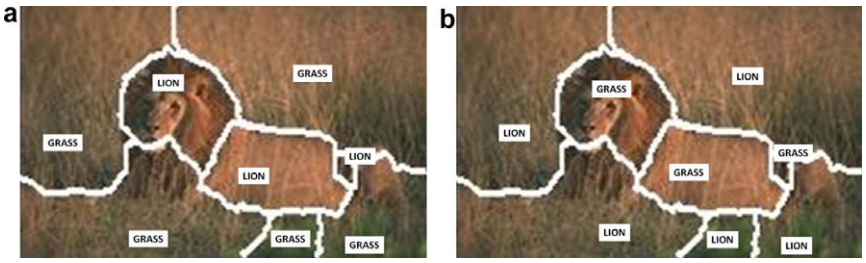

Fig. 2. Left: correct region-level annotation, right: incorrect region-level annotation.

[21] is another game that uses the image annotations generated by the ESP game. In this game, the goal is to provide the locations and geometric labels of objects. The resulting collection could be used to train algorithms for a variety of tasks, including region-level AIA. However, the number of annotators can be in the range of millions and, as a consequence, there is a lack of common annotation criteria to obtain consistent object localizations.

An important effort is also being carried out by Russell et al. [22], with the goal of creating benchmark collections for diverse computer vision applications. The LabelME project uses a web-based online tool for the segmentation and annotation of images. Segmentations are specified by drawing polygons around each object, while the annotation vocabulary is defined by the users.[2] The benefits of this collection are its size and that is publicly available; its limitation is its open vocabulary: regions can be assigned any word depending on the annotator background, and very different segmentations can be obtained for similar images as well.

Yao et al. [23] have also been putting tremendous effort into creating another large-scale benchmark collection. Currently, more than 630,000 images with very detailed manual segmentations have been considered in their project. The segmented objects are thereby decomposed and organized into a hierarchy similar to a syntactic tree in linguistics, and information about localization as well as 2D and 3D geometry is also available. The collection is divided into 13 subsets, according to the type of images and their applications. This collection has the potential to be a very useful resource for building visual dictionaries and as training data for learning algorithms. It could also be used to evaluate AIA methods; however, since the collection lacks ground-truth data to evaluate image retrieval (i.e. relevance judgments), it cannot be used to assess the impact of AIA methods on multimedia information retrieval.

There are several other collections available for the evaluation of object recognition algorithms [24], most notably the Caltech-101 [25] and Caltech-256 [16] benchmarks as well as the PASCAL VOC-2006 [15] and VOC-2007[3] collections. The type of images in such collections, however, are not suitable for the evaluation of AIA methods (see Section 1.2), yet even their use for the evaluation of object recognition methods has been questioned [24].

In fact, there are currently only a few collections that could, in principle, be used for the effective evaluation of region-level AIA methods. Most of these, however, are restricted to specific domains, such as cars [26], nature-roadways [27], animals [28], landscape vs. structured classification [29], and natural scene classification [30]. The size of the data sets and the restricted domains generate results that are not representative of those for general purpose AIA. Winn et al. [31], for example, segmented and annotated a small set of 240 images, considering nine labels only. More recent work by Shotton et al. [32] reports on the creation of a larger collection with 591 images and 21 labels, yet again the size of these data sets and the number of concepts are not adequate for evaluating AIA. Furthermore, Carbonetto et al. [8] provide three small data sets with a larger number of labels (from 22 to 56). To the best of our knowledge, these are the largest data sets publicly available that have been annotated at a region-level. Unfortunately, the data sets are still small and originate from the Corel collection.

A highly relevant collection to AIA is that provided by Barnard et al. [10]. It consists of 1041 images taken from a previous study for benchmarking segmentation algorithms [33]. The segmentations are thereby very detailed, and the annotations are specified using WordNet and according to well defined criteria. In principle, this methodology could be very useful for the segmentation and annotation of the IAPR TC-12 collection; however, the large size of the collection and the detailed segmentation and annotation processes introduced by Martin et al. [33] and Barnard et al. [10] make its application impractical.[4] A straightforward methodology was proposed for the evaluation of localization performance in region-level AIA, which facilitates the performance evaluation of methods that do not use the same vocabulary as that used in [10]. Therefore, although the data set is small and images come from the Corel collection, their main contribution comes from the evaluation methodology that allows for the assessment of AIA techniques using other collections.

The IAPR TC-12 benchmark was created with the goal of providing a realistic collection of images suitable for a wide number of evaluation purposes [12]. The collection comprises 20,000 images taken from locations around the world and comprising a varying cross-section of still natural images, including pictures of sports, actions, people, animals, cities, landscapes, and many other topics. Manual annotations in three languages are provided with each image. The IAPR TC-12 benchmark has been used to evaluate cross-language TBIR as well as CBIR and multimedia image retrieval methods [13,14]. It has also been used for object retrieval [34] and visual concept detection [35]. For the evaluation of visual concept detection, about 1,800 images were annotated with visual concepts; this was an early effort to use the collection for tasks related to AIA. However, only 17 concepts were considered for this task. Given the variety of images in the collection, this limited vocabulary is not sufficient for the automatic annotation of the entire collection, and annotations are currently only available at image-level. Previously, the IAPR TC-12 collection had also been used for the task of object retrieval, using the PASCAL VOC-2006 collection for training and the IAPR TC-12 as test set [34]. However, the number of objects was 10 and the accuracy of most of the methods was poor; the results showed that specialized collections are required for benchmarking different tasks.

Despite being an established benchmark, the IAPR TC-12 collection is not well-suited for the evaluation of region-level AIA methods in its original form. Hence, an extension is needed in order to increase the tasks supported by the collection; such an extension is the main contribution of this work. Although there are several

[2] Any Internet user can thereby be a potential annotator in LabelME.

[3] http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html.

[4] Nevertheless, the general principles in their work form the base for the segmentation and annotation guidelines described in Section 3 hereinafter.

Please cite this article in press as: H.J. Escalante et al., The segmented and annotated IAPR TC-12 benchmark, Comput. Vis. Image Understand. (2009), doi:10.1016/j.cviu.2009.03.008

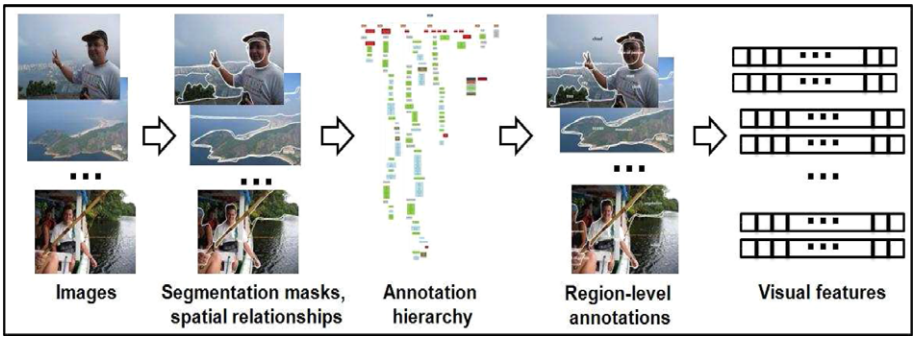

Fig. 3. Evaluation resources of the extended IAPR TC-12 benchmark.

collections available that can be used to evaluate the performance of AIA to some extent [8,10,20–23,28,31,32], there is no collection that can be used to study the impact of using AIA methods for image retrieval. The extended IAPR TC-12 collection will finally answer the call for a standardized benchmark for the evaluation of region-based AIA methods.

3. The annotation of the extended IAPR TC-12 benchmark

This section describes the methodology adopted for the extension of the IAPR TC-12 collection, which consists of the manual segmentation and annotation of the entire collection. This includes the definition of an annotation vocabulary, its hierarchical organization, guidelines for the segmentation and annotation processes, and statistics of the extended collection. We only focus on describing the main aspects of the project; further technical details can be found in [36]. Fig. 3 displays an overview of the evaluation resources provided by the extended IAPR TC-12 benchmark, which include segmentation masks, annotations, visual features and spatial relationships. These are now available for research purposes from both the ImageCLEF[5] and INAOE-TIA[6] websites.

3.1. Vocabulary

The vocabulary plays a key role in the annotation process because it must cover most of the concepts that one can find in the image collection. At the same time, the vocabulary should not be too large because AIA performance is closely related to the number of labels considered (see Section 4.1).

Several publications have reported on diverse annotation vocabularies in the past, with many of them being very specific for the type of images in particular collection considered. This work is based on a study carried out by Hanbury [37], in which a list of 494 labels was obtained by analyzing several AIA benchmark collections. We took this word list (H-list) as a reference and adapted it to the IAPR TC-12 collection. We also considered the list of nouns from the manual image annotations of that collection (A-list) and the list of nouns in the textual description of topics used for ImageCLEF 2006 and 2007 (T-list) [13,14].

First, we selected the labels appearing in at least two lists into a candidate list (C-list). Then, we manually filtered the C-list by (i) analyzing a large number of images from the IAPR TC-12 collection and discarding labels not present or rarely found in the images (e.g. *'binoculars'*,*'printer'*), and by (ii) considering the frequency of labels in the annotations of the IAPR TC-12 collection for filtering images: highly-frequent and useful words were kept (e.g. *'sky'*, *'mountain'*), while highly-frequent but useless and non-frequent words were not considered (e.g. *'background'*, *'marble'*). Finally, we incorporated some words in the H-list that had initially been dropped from the C-list (e.g. *'herd'*) as well as words identified by the authors (e.g. *'sky-red'*) that did not appear in any of the three lists into the final C-list. The latter procedure was iterated several times until the authors fully agreed on the final list.

Table 1 provides an overview of the annotation vocabulary obtained using the aforementioned methodology. Words included during the creation of the hierarchy are shown in **bold**. Abbreviations are as follows: ar-aerostatic, int-interior, per-person, ot-other, wo-wood, anim-animal, ins-instrument, sp-space, obj-object, veh-vehicle, fur-furniture, swim-swimming and refl-reflection.

3.2. Conceptual hierarchy

During the annotation process, the need of a hierarchical organization for the vocabulary arose. After we had carefully analyzed the images, the annotation vocabulary and the vocabulary of the existing manual annotations, we manually defined a hierarchy. We organized the annotation vocabulary mostly by using *is-a* relations between labels, but also included relations like *part-of* and *kind-of*. The hierarchy was thereby based on its usefulness for the annotation and representation of the images in the IAPR TC-12 collection, rather than considering the semantics of labels in general.

According to the proposed hierarchy, an object can be in one of six main branches: *'animal'*, *'landscape'*, *'man-made'*, *'human'*, *'food'*, or *'other'*. Fig. 4 shows the *'landscape'* branch of the hierarchy, using different colors to describe different levels; other branches and further details can be found in [36]. Some nodes in the same level (e.g. *'Vegetation'*) contain more descendant labels than others (e.g. *'Arctic'*), which is due to the type of images in the IAPR TC-12 collection. Some nodes are barely populated, like *'fruit'* - one could easily include all the names of fruits under this node, increasing the coverage on the variety of fruits; however, this would only lead to a large vocabulary that would scarcely be used for annotation, because the images in the collection do not contain a considerable diversity of fruits. The hierarchy reflects the principle adopted for defining the annotation vocabulary: *"keep it compact and ensure it covers most of the concepts present in the images of the IAPR TC-12 collection"*.

Barnard et al. [10] considered WordNet for the annotation of images; however, we did not base our hierarchy on WordNet as assigning a region to its nearest WordNet meaning would lead to ambiguities due to the subjective knowledge of annotators[7]; the use of WordNet would also make annotation slower.

[5] http://imageclef.org/photodata.
[6] http://ccc.inaoep.mx/~tia.

[7] E.g. a region of *'water'* may be labeled with *'river'* by an annotator, while the same region may be labeled with *'stream'*, *'watercourse'* or simply *'water'* by other annotators; even the same annotator may assign different WordNet concepts to similar regions in different images or at different times.

Please cite this article in press as: H.J. Escalante et al., The segmented and annotated IAPR TC-12 benchmark, Comput. Vis. Image Understand. (2009), doi:10.1016/j.cviu.2009.03.008

Table 1
Annotation vocabulary of the extended IAPR TC-12 benchmark.

ar_balloon	**air_vehicles**	airplane	ancient_build	ape	animal	ant
antelope	apple	astronaut	arctic	baby	ball	balloon
beach	bear	beaver	bed	beetle	bench	bicycle
bird	boat	boat_rafting	bobcat	book	bottle	branch
bridge	building	bull	bus	bush	butterfly	cabin
cactus	camel	camera	can	**canine**	cannon	car
caribou	castle	cat	caterpillar	cello	chair	cheetah
child	child_boy	child_girl	chimney	church	church_int	city
clock	cloth	cloud	column	construction	**cons_ot**	coral
cougar	couple_per	cow	coyote	crab	crocodile	cup
curtain	deer	**desert**	desk	dish	diver	dog
dolphin	door	dragonfly	eagle	**edifice**	elephant	elk
fabric	face	feline	fence	field	fire	firework
fish	flag	flamingo	flock_birds	**floor**	floor_carpet	floor_court
floor_other	floor_wood	flower	flowerbed	food	forest	fountain
fowl	fox	fruit	**furniture**	furniture_ot	furniture_wo	**generic_obj**
giraffe	glacier	glass	goat	grapes	grass	ground
ground_veh	group_per	guitar	hand	handcraft	hat	hawk
head	hedgehog	helicopter	herd	highway	hill	horn
horse	house	**humans**	hut	ice	iguana	insect
island	jewelry	jaguar	kangaroo	kitchen_pot	koala	lake
lamp	**landscape**	leaf	leopard	lighthouse	lion	lizard
llama	lobster	log	lynx	mammal	**mammal_ot**	man
man_made	**man_made_ot**	mandril	marsupial	monkey	monument	motorcycle
mountain	mural_carving	**non_wo_fur**	mushroom	musical_inst	nest	**object**
ocean	ocean_anim	octopus	orange	**ot_entity**	owl	pagoda
painting	palm	panda	paper	parrot	penguin	person
per_rel_obj	piano	pigeon	plant	plant_pot	polar_bear	pot
primate	public_sign	pyramid	rabbit	rafter	railroad	reflection
reptile	rhinoceros	river	road	rock	rodent	roof
rooster	ruin_arch	sand	**sand_desert**	**sand_beach**	saxophone	school_fishes
scorpion	screen	seahorse	seal	semaphore	shadow	sheep
shell	ship	shore	sidewalk	**sky**	sky_blue	sky_light
sky_night	sky_ed	smoke	snake	snow	sp_shuttle	squirrel
stairs	starfish	statue	steam	strawberry	street	sun
surfboard	swim_pool	table	telephone	tiger	tire	tower
toy	train	trash	tree	**trees**	trombone	trumpet
trunk	turtle	umbrella	vegetable	vegetation	vehicle	**veh_tires**
viola	violin	volcano	wall	water	**water_refl**	**water_veh**
wooden_fur	waterfall	wave	whale	window	wolf	woman
wood	zebra					

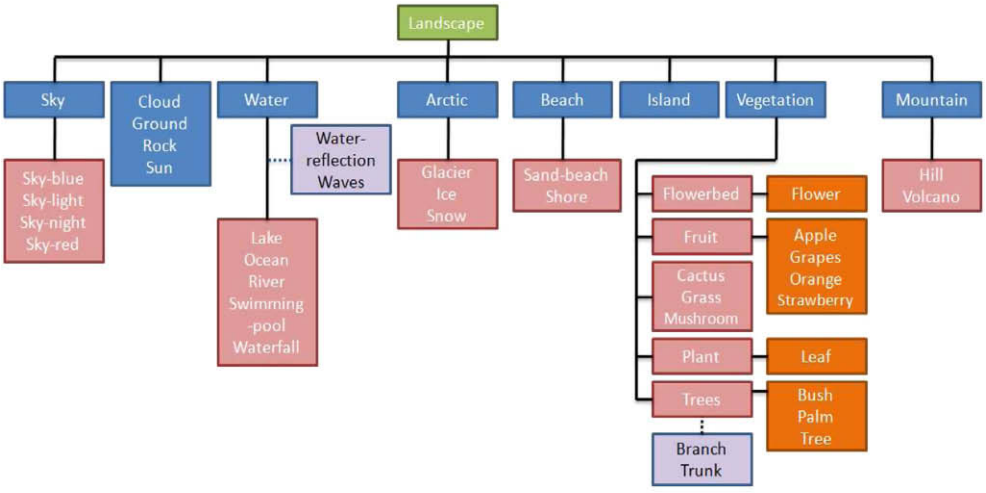

Fig. 4. Detailed view of the *'landscape-nature' branch*.

Hence, we considered more generic words which are more likely to be agreed on, thereby trading precision for agreement. A restricted hierarchy also allows a more concise and objective annotation of the collection; of course, some concepts may not be covered by the proposed hierarchy, although it proved to be very useful for the annotation of the IAPR TC-12 collection. The proposed hierarchy thereby resembles hierarchies proposed in related works [10,23,35].

The main purpose of this hierarchy is to facilitate and improve the subsequent annotation process, which can be achieved by

Please cite this article in press as: H.J. Escalante et al., The segmented and annotated IAPR TC-12 benchmark, Comput. Vis. Image Understand. (2009), doi:10.1016/j.cviu.2009.03.008

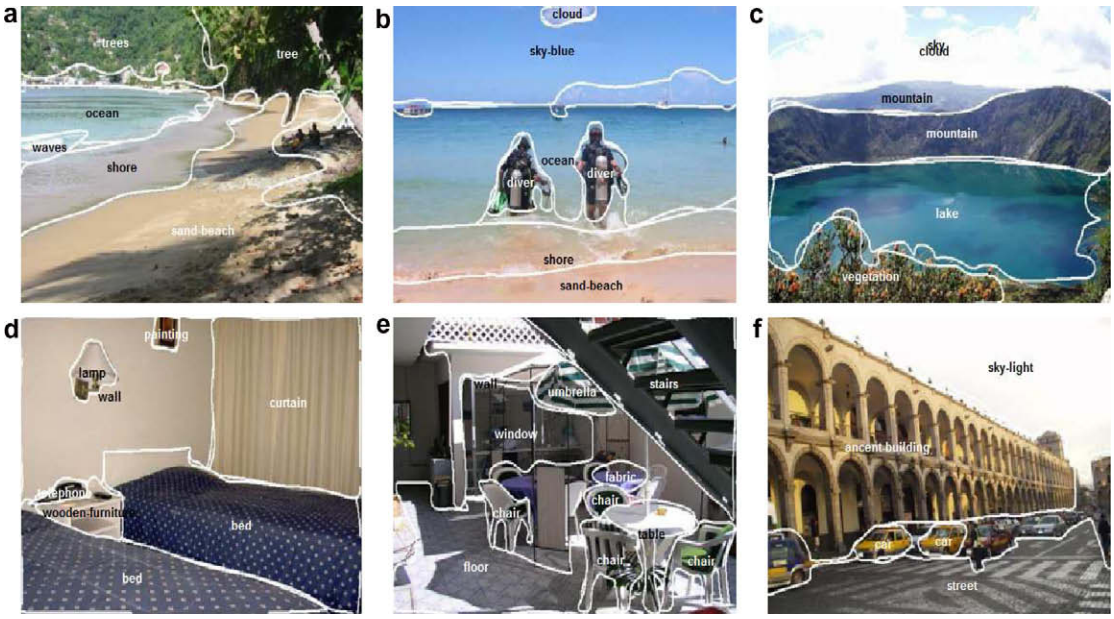

Fig. 5. Sample images from the extended IAPR TC-12 collection.

going through the hierarchy top-down each time a region needs to be labeled. This reduces ambiguities when annotating similar regions referring to different concepts (visual homonymy) and different regions about the same concept (visual synonymy). The hierarchical organization of concepts is also helpful for the *soft* evaluation of AIA methods (see Section 4).

3.3. Segmentation and annotation guidelines

Manually segmenting and annotating images are tasks prone to subjectivity, because different persons can segment and annotate the same image in completely different ways; yet even the same individual can provide different segmentations of the same image at different times. We therefore only considered four annotators to deal with this subjectivity. Moreover, in order to standardize the process of segmenting and annotating images, we defined a set of guidelines based on those created by Martin et al. [33] and Barnard et al. [10]. The goal was to make the segmentation and annotation processes as consistent as possible by reducing ambiguities and confusion among annotators. The main guidelines are summarized below.[8] During the segmentation and annotation process, annotators should:

- Avoid to segment regions that are too small with respect to the size of the image. What is considered too small depends on the object under consideration (e.g. *'bottle'* regions are usually smaller than *'mountain'* ones).
- Avoid to segment regions where the object is incomplete: at least one third of the object must be visible.
- Bear in mind that each region has to contain information from a single object.
- Provide a detailed segmentation when the shape of an object is highly relevant; otherwise, relax the segmentation in a way that it can be performed faster.
- Divide the object into more than one region if creating those smaller regions is easier than segmenting the original single one (for images in which shape is considered irrelevant).
- Avoid the segmentation of areas exhibiting excessive illumination, shadows or other conditions that make it difficult to segment these regions, and only segment what can be seen without difficulty.
- Go through the hierarchy top-down looking for the label that best describes the region and avoid to overuse general-purpose labels.
- Use the closest label to the term the region belongs to.
- Select the label in higher levels of the hierarchy whenever a suitable label is not found.
- Look for adequate labels in the 'other' branch if a correct label is not found under the category of interest.
- Segment object groups as a unit in case the respective group label exists in the vocabulary (e.g. *'persons'*, *'mammals'*, *'trees'*), and also segment its individual units as far as possible.

Although several tools are available for interactive segmentation [28,33], most of them are designed to provide rather detailed segmentations of images. This would certainly be desirable for the current work; however, the number of images and annotators involved make the use of such methods very time-consuming and therefore impractical. Hence, we developed a Matlab^R application, called ISATOOL (*Interactive Segmentation and Annotation Tool*), for the segmentation and annotation of the IAPR TC-12 collection. ISATOOL allows the interactive segmentation of objects by drawing points around the desired object, while splines are used to join the marked points, which also produces fairly accurate segmentation (compare Fig. 5) with much lower segmentation effort. Once a region is segmented, the user is then asked to provide a label for the region by using the hierarchy described in Section 3.2.

In addition, a set of simple visual features extracted from the regions and spatial relationships are provided with the extended collection. The following features were extracted from each region: area, boundary/area, width and height of the region, average and standard deviation in x and y, convexity, average, standard deviation and skewness in the RGB and CIE-Lab color spaces. Furthermore, spatial relationships are calculated and provided with the collection. Those include adjacent, disjoint, beside, X-aligned, above, below and Y-aligned [36]. We offer these spatial relations in order to promote the use of contextual information in the AIA

[8] We assume that the annotators know the full list of words in advance and that they are familiar with the segmentation and annotation tool.

Please cite this article in press as: H.J. Escalante et al., The segmented and annotated IAPR TC-12 benchmark, Comput. Vis. Image Understand. (2009), doi:10.1016/j.cviu.2009.03.008

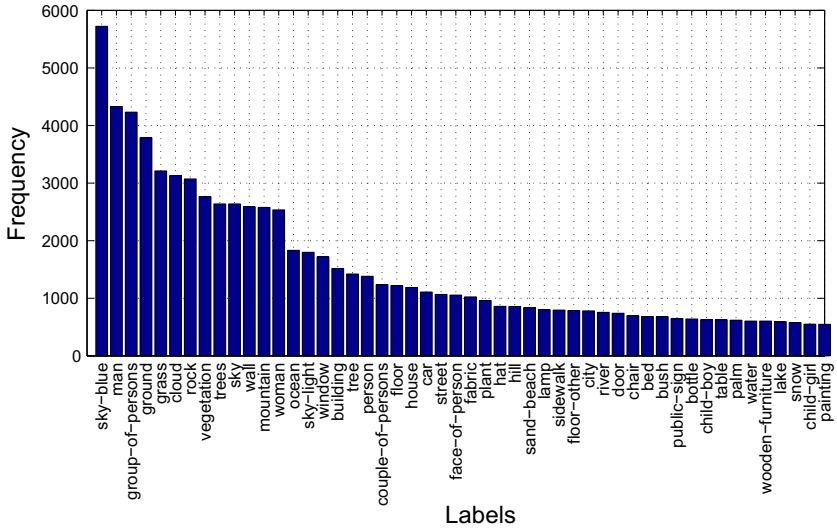

Fig. 6. Histogram of regions annotated with each label for the 50 most common labels.

task. All these features and spatial relationships have successfully been used in previous AIA work [8,17,18,38]; however, each user can extract their own set of features and spatial relationships since segmentation masks and images are publicly available.

3.4. Statistics of the collection

The 20,000 images in the collection have been segmented and annotated according to the aforementioned methodology; we thereby segmented a total of 99,555 regions, which were annotated using 255 labels, representing about 93% of the total vocabulary. Only 19 labels of *'animals'* and *'musical-instruments'* were not used. On average, 4.97 regions have been segmented per image; these regions occupy an averaged 15.61% of the images. Like in many AIA collections, there are some labels that have been used in a considerable number of regions, while others have barely been used at all. Fig. 6 plots the number of regions annotated with the 50 most frequent labels. The 10 most common labels are: *'sky-blue' (5722), 'man', (4330), 'group-persons' (4232), 'ground' (3785), 'grass' (3211), 'cloud' (3128), 'rock' (3071), 'vegetation' (2763), 'trees' (2638), and 'sky' (2637)*. The most common labels thereby correspond to the type of images present in the collection (i.e., pictures of people on vacation trips [12]). There are 26, 122 and 147 labels that have been used in more than 1000, 100 and 50 regions, respectively.

A total of 200 leaves in the hierarchy have been used for annotation; Table 2 shows the distribution of annotations for the nodes in the first level of the hierarchy. **Frequency** thereby indicates the number of regions annotated with labels in or below each node, and **Norm. Freq.** shows the frequency amortized by the number of descendants of each node. There are more than 45,000 regions annotated with labels below the *'Landscape'* node; it has 45 descendants, of which 33 are leaves. Labels below the *'Man-made'* node are also very popular; more than 34,000 regions have been annotated with labels from below that node, with 110 nodes as descendants, of which 88 are leaves. *'Humans'* is a node with many regions as well; however, its number of descendants is small when compared to other nodes at the same level.

The normalized frequency (third row in Table 2) shows the average number of labels assigned to each descendant in the considered nodes. The branch *'Humans'* is the one with the most annotations per descendant; *'Landscape'* and *'Man-made'* come next. This, again, reflects the type of images in the collection: *'Humans'* appear in many images, since most of the images were taken by/for tourists. Furthermore, a large number of pictures were taken in rural South-American places; therefore, there are also many images that contain labels from the *'landscape-nature'* branch.

The normalized frequency of spatial relationships is described in Table 3; **No. Ex.** provides the number of regions that present each relationship, while **Norm. Freq.** shows its normalized frequency. The most frequent relations are *beside* and *disjoint*, with 25.65% and 23.21%, respectively. Note that *beside* is a generalization of the *left* and *right* relations, and this is reflected in its frequency. *X*-alignment and *Y*-alignment are low frequency relations, with 7.68% and 6.82%, respectively. Finally, the propor-

Table 2
Distribution of annotations for labels in and below the nodes in the first level of the hierarchy.

Label	*'Animal'*	*'Humans'*	*'Food'*	*'Man-made'*	*'Landscape'*	*'Other'*
Frequency	1991	16,950	705	34,024	45,308	622
Norm. Freq.	28.44	1210.71	117.5	309.35	1006.84	103.67
Descendants	70	14	6	110	45	6
Leaves	56	12	5	88	33	6

Table 3
Frequency of spatial relationships among regions in the extended IAPR TC-12 collection.

ID	Adjacent	Disjoint	Beside	*X*-alig	Above	Below	*Y*-alig
No. Ex.	176,466	404,716	447,212	133,970	231,169	231,169	118,844
Norm. Freq.	10.12%	23.21%	25.65%	7.68%	13.26%	13.26%	6.82%

Please cite this article in press as: H.J. Escalante et al., The segmented and annotated IAPR TC-12 benchmark, Comput. Vis. Image Understand. (2009), doi:10.1016/j.cviu.2009.03.008

tions obtained by *above* and *below* reflect their symmetry property, both with 13.26%

4. An evaluation measure for the benchmark

The extended IAPR TC-12 collection would certainly allow the evaluation of region-level AIA methods by using common classification-performance measures such as the area under the ROC curve, misclassifications rate or squared root error, as these measures can effectively assess the localization performance of annotation methods. However, they could prove to be too rigorous for the current state-of-the-art technology in region-level AIA. Consider, for instance, the case in which the correct label for a given region is *'water'* and the model under study classifies such a region as *'river'*. In this situation, a classification-performance measure would consider the assignment as totally incorrect, despite this prediction being partially correct.

In order to give *partial credit* to these annotations, we introduce a novel evaluation measure for region-level AIA in the context of the annotated IAPR TC-12 collection. This additional[9] measure $e_{hierarchy}(t,p)$ is based on the annotation hierarchy described in Section 3.2 and is defined as

$$e_{hierarchy}(t,p) = \left[\mathbf{1}_{in-path(t,p)} \times \frac{|f_{depth}(t) - f_{depth}(p)|}{\max(f_{depth}(t), f_{depth}(p))}\right] \quad (1)$$

where $\mathbf{1}_{in\text{-}path(t,p)}$ is an indicator function that takes the value of 1 when both the predicted label p and the true label t are in the same path of the annotation hierarchy, and $f_{depth}(x)$ is the depth of label x within the hierarchy. Intuitively, $e_{hierarchy}(t,p)$ assigns an error value to a label predicted by a model, proportional to its normalized distance (within the hierarchy) with respect to the ground-truth label. A predicted annotation will be evaluated as partially good if and only if it appears in the same branch as the correct label. Unless the true and predicted labels are different, $e_{hierarchy} = 0$ by definition.

In addition, $e_{hierarchy}$ can be modified such that only more specific (general) predictions are considered as partially correct; this can be achieved by altering $\mathbf{1}_{in\text{-}path(t,p)}$ such that it only takes the value of 1 when the predicted label p is more specific (general) than the true label t. Furthermore, a threshold can be set on $e_{hierarchy}$ such that only certain types of errors are considered as partially correct.

4.1. Benchmarking AIA

This section presents experiments and results on region-level AIA using subsets of the extended IAPR TC-12 collection. We illustrate how this collection can be used to evaluate region-level AIA and compare the novel soft evaluation measure to a hard one. We approach the challenge of AIA as one of multi-class classification with as many classes as labels in our vocabulary, whereby we consider state-of-the-art classifiers over subsets of the already annotated regions. Table 4 describes the classifiers with their respective (default) parameters that we chose for our experiments; these classifiers are included in the CLOP machine learning toolbox[10]. We considered different subsets of data according to the frequency of annotated regions per class. Table 5 describes the selected data subsets and the distribution of the classes and examples. For each data subset, the available data were randomly split into disjoint training (70%) and test (30%) sets. In each experiment, k-classifiers were trained under the *one-versus-all* (OVA) formulation (k being the number of classes). OVA is a widely used approach for supervised AIA [9,17,18,29,38]. Under this schema, a binary classifier is trained for each label; when the kth-classifier is trained, only examples from class k are regarded as positive and all other examples as negative.

Table 4
Classifiers used in the experiments.

Classifier	Description	Parameters
Zarbi	A simple linear classifier	—
Naive	Naive Bayes classifier	—
Klogistic	Kernel logistic regression	—
Neural	Feedforward Neural Network	Units = 10, Shrinkage = 0.1, Epochs = 50
SVC	Support Vector Classifier	Kernel = *Poly*, Degree = 1, Shrinkage = 0.001
Kridge	Kernel ridge regression	Kernel = *Poly*, Degree = 1, Shrinkage = 0.001
RF	Random Forest	Depth = 1, Shrinkage = 0.3, Units = 100

Although OVA is very simple, it has proved to be competitive in comparison to more sophisticated and complex approaches. One of the main challenges in OVA classification is choosing a way to combine the outputs of binary classifiers such that errors with respect to unseen data are minimized [39]. Since the goal of the experiments is only illustrative, we adopted the simplest strategy for merging the outputs of the individual classifiers: when more than one classifier was triggered, we preferred the prediction of the classifier with the higher confidence about its respective class.

For evaluation purposes, we considered the percentage of correct classifications (e_{hard}), a widely used measure for assessing the performance of classifiers, and compared it to $e_{hierarchy}$ as described in Eq. (1). We thereby say that a predicted annotation is correct whenever $e_{hierarchy} < 1$. Fig. 7 shows the average e_{hard} (left) and $e_{hierarchy}$ (right) for the classifiers and data sets described in Tables 4 and 5, and compares them to two baselines: baseline-1 is the performance one would get by randomly selecting labels, and baseline-2 is the performance one would get by assigning always the majority class.

The graphs indicate that both measures decrease similarly as the number of labels increases, because e_{hard} is a special case of $e_{hierarchy}$. For the data sets A–D, both measures obtain the same result, as there are no hierarchical relations among the five most frequent classes. However, there are small differences for the data sets E–H, for which information from the hierarchy can be used by $e_{hierarchy}$. The average differences between e_{hard} and $e_{hierarchy}$ are of 7.79%, 6.8%, 6.08% and 5.91% for the data sets E, F, G and H, respectively. This reflects that the soft measure is, indeed, less rigid for the evaluation of the classifiers, although the difference is not very large. Thus, $e_{hierarchy}$ becomes increasingly useful to differentiate between the performance of annotation methods when the number of classes is high, as the existing measures then often fail to discriminate between methods.

Table 6 shows sample labels assigned to test regions that were classified as incorrect by e_{hard} but as correct by $e_{hierarchical}$. In this experiment, all the labels used for annotation were considered (i.e., 255 classes); **Correct** shows the correct label for the region and **predicted** stands for the label predicted by the model.

All the labels considered as partially good are indeed closely related to the correct label, which is highly useful for the evaluation of ABIR methods. The value of $e_{hierarchy}$ reflects how precisely the AIA method is able to predict the correct class, which makes $e_{hierarchy}$ a very well-suited measure for the evaluation of region-based AIA methods using the extended IAPR TC-12 collection.

5. Applications for the extended IAPR TC-12 Benchmark

The original IAPR TC-12 benchmark has already been used for the performance evaluation of methods in several information

[9] The new measure should be seen as an alternative evaluation of performance in addition to the existing classification-performance measure and does not intend to replace it.
[10] http://clopinet.com/CLOP/.

Please cite this article in press as: H.J. Escalante et al., The segmented and annotated IAPR TC-12 benchmark, Comput. Vis. Image Understand. (2009), doi:10.1016/j.cviu.2009.03.008

Table 5
Distribution of classes and examples for the subsets used in the experiments.

ID	A	B	C	D	E	F	G	H
No. classes	2	3	4	5	10	25	50	100
No. examples	10,052	14,284	18,069	21,280	35,517	59,759	77,631	93,167

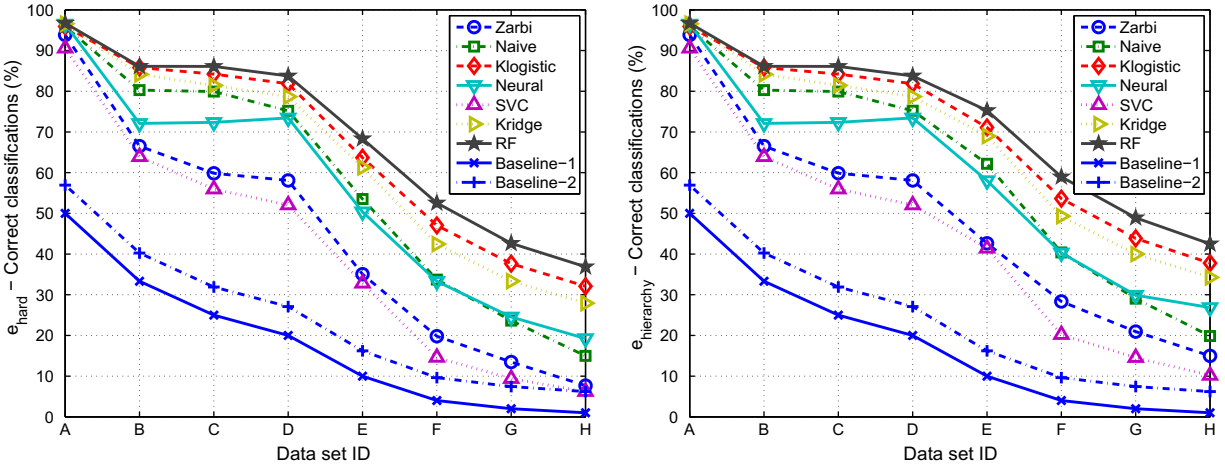

Fig. 7. Percentage of correct annotations for the considered classifiers and data sets.

Table 6
Analysis of labels evaluated as non-erroneous with $e_{hierarchy}$ and as errors using hard measure.

Correct	Predicted	$e_{hierarchy}$	Correct	Predicted	$e_{hierarchy}$
tree	trees	0.25	child-boy	child	0.25
person	child	0.33	sky	sky-light	0.33
food	dish	0.5	vegetation	bush	0.5

retrieval research areas such as CBIR, multimedia information retrieval, cross-language TBIR, object retrieval and visual concept detection. The additional evaluation resources provided by this work make the fully segmented and annotated IAPR TC-12 benchmark now also attractive to evaluation tasks that have not been covered by the test collection before. Some likely evaluation scenarios and potential future applications include the following:

- *Benchmarking.* The extended benchmark is well-suited to evaluate the tasks of region-level AIA, image-level AIA, visual concept detection, and object retrieval. It should also be considered to assess the performance of object detection and object recognition techniques (e.g. face and skin detection). Furthermore, the benchmark facilitates the evaluation of segmentation performance and motivates further research on the use of the spatial context for AIA, ABIR and CBIR.
- *Multimedia information retrieval.* The extended collection can be used to shed some light on the clarification of assumptions commonly made in multimedia image retrieval. For example, the assumption that the use of information from AIA methods can be useful to improve the retrieval performance of CBIR or TBIR methods; despite being intuitively sound, it has neither been proved theoretically nor empirically. Similarly, the collection could contribute to the research of the extent that automatic segmentation affects the retrieval and annotation performance, among several other interesting aspects [36]. Furthermore, the spatial relationships can be used to allow complex queries on the collection, whereby the interest could lie in finding objects in specific positions with respect to other objects, motivating research on the use of spatial relations for multimedia information retrieval.
- *Machine learning applications.* The annotated collection could even be used to bridge the gap between the machine learning and multimedia information retrieval communities because it allows the analysis and evaluation of multi-class classification for a large number of labels[11], hierarchical classification for AIA, spatial data mining, multi-modal clustering, classification for imbalanced data sets, and the application of structured prediction methods to the problems of AIA, ABIR, and CBIR.

6. Conclusion

The IAPR TC-12 benchmark is an established image retrieval test collection that has several attractive features: it contains a large-size image collection comprising diverse and realistic images, it offers textual annotations in three languages, and it provides query topics, relevance assessments and a set of measures for the evaluation of image retrieval performance. Benchmarks, however, are not static by nature but have to evolve and develop according to the needs of the tasks they are designed for and the emergence of new evaluation needs.

Hence, in this paper, we introduced the segmented and annotated IAPR TC-12 collection, an extension to the existing benchmark which increases the number of evaluation tasks that can be accommodated for and which also significantly augments the number of potential future applications for the test collection. In particular, we described the methodology adopted for the systematic extension of the IAPR TC-12 collection, including the definition of an ad-hoc annotation vocabulary, its hierarchical organization and well defined criteria for the objective segmentation and annotation of images. We also proposed an additional evaluation measure based on the hierarchy, with the goal of assessing localization performance of region-level AIA methods.

[11] Note that the hierarchy of labels can be used to study the performance of classifiers with an increasing number of labels.

Please cite this article in press as: H.J. Escalante et al., The segmented and annotated IAPR TC-12 benchmark, Comput. Vis. Image Understand. (2009), doi:10.1016/j.cviu.2009.03.008

Statistics of the extended collection give evidence that the adopted methodology is reliable and well-suited for the extended IAPR TC-12 benchmark. In addition, initial results using the proposed evaluation measure indicate that it can effectively evaluate the performance of region-level AIA methods. Another contribution of this work is the identification of potential future applications and evaluation scenarios for the extended IAPR TC-12 benchmark, which is likely to motivate further research that will significantly contribute to advance the state-of-the-art technology in research areas such as segmentation, AIA, CBIR, TBIR, ABIR and machine learning.

References

[1] R. Datta, D. Joshi, J. Li, J.Z. Wang, Image retrieval: ideas, influences, and trends of the new age, ACM Computing Surveys 40 (2) (2008) 1–60.
[2] M. Inoue, On the need for annotation-based image retrieval, in: IRiX'04: Proceedings of the ACM-SIGIR Workshop on Information Retrieval in Context, Sheffield, UK, 2004, pp. 44–46.
[3] Y. Mori, H. Takahashi, R. Oka, Image-to-word transformation based on dividing and vector quantizing images with words0, MISRM'99 First International Workshop on Multimedia Intelligent Storage and Retrieval Management, Orlando, Florida, USA, 1999.
[4] P. Duygulu, K. Barnard, N. de Freitas, D.A. Forsyth. Object recognition as machine translation: learning a lexicon for a fixed image vocabulary, in: ECCV'02: Proceedings of the 7th European Conference on Computer Vision, LNCS vol. 2353, Springer, London, UK, 2002, pp. 97–112.
[5] K. Barnard, P. Duygulu, N. de Freitas, D.A. Forsyth, D. Blei, M.I. Jordan, Matching words and pictures, Journal of Machine Learning Research 3 (2003) 1107–1135.
[6] J. Jeon, V. Lavrenko, R. Manmatha, Automatic image annotation and retrieval using cross-media relevance models, in: Proceedings of the 26th International ACM-SIGIR Conference on Research and Development on Information Retrieval, Toronto, Canada, 2003, pp. 119–126.
[7] D. Blei. Probabilistic Models of Text and Images, PhD Thesis, University of California, Berkeley, California, USA, 2004.
[8] P. Carbonetto, N. de Freitas, K. Barnard, A statistical model for general context object recognition, in: ECCV'04: Proceedings of the 8th European Conference on Computer Vision, LNCS vol. 3021, Springer, Prague, Czech Republic, 2004, pp. 350–362.
[9] G. Carneiro, A.B. Chan, P.J. Moreno, N. Vasconcelos, Supervised learning of semantic classes for image annotation and retrieval, IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (3) (2007) 394–410.
[10] K. Barnard, Q. Fan, R. Swaminathan, A. Hoogs, R. Collins, P. Rondot, J. Kaufhold, Evaluation of localized semantics: data methodology and experiments, International Journal of Computer Vision 77 (1–3) (2008) 199–217.
[11] H.J. Escalante, J.A. González, C.A. Hernández, A. López, M. Montes, E. Morales, L.E. Sucar, L. Villaseñor, Annotation-Based Expansion and Late Fusion of Mixed Methods for Multimedia Image Retrieval, Evaluating Systems for Multilingual and Multimodal Information Access – 9th Workshop of the CLEF, LNCS, Springer, Aarhus, Denmark, 2009, in press.
[12] M. Grubinger, Analysis and Evaluation of Visual Information Systems Performance, PhD Thesis, School of Computer Science and Mathematics, Faculty of Health, Engineering and Science, Victoria University, Melbourne, Australia, 2007.
[13] P. Clough, M. Grubinger, T. Deselaers, A. Hanbury, H. Müller, Overview of the ImageCLEF 2006 photographic retrieval and object annotation tasks, Evaluation of Multilingual and Multimodal Information Retrieval – 7th Workshop of the CLEF, LNCS vol. 4730, Springer, Alicante, Spain, 2006 (printed in 2007), pp. 579–594.
[14] M. Grubinger, P. Clough, A. Hanbury, H. Müller, Overview of the ImageCLEF 2007 photographic retrieval task, Advances in Multilingual and Multimodal Information Retrieval – 8th Workshop of CLEF, LNCS vol. 5152, Springer, Budapest, Hungary, 2007 (printed in 2008), pp. 433–444.
[15] M. Everingham, A. Zisserman, C.K.I. Williams, L. Van Gool, The PASCAL Visual Object Classes Challenge 2006 (VOC2006) Results, Tech. rep., University of Oxford, Oxford, UK, 2006.
[16] G. Griffin, G. Holub, P. Perona, The Caltech-256, Tech. rep., California Institute of Technology, Pasadena, California, USA, 2007.
[17] H.J. Escalante, M. Montes, L.E. Sucar, Word co-occurrence and Markov random fields for improving automatic image annotation, in: Proceedings of the 18th British Machine Vision Conference, vol. 2, Warwick, UK, 2007, pp. 600–609.
[18] H.J. Escalante, M. Montes, L.E. Sucar, Multi-Class PSMS for Automatic Image Annotation, Applications on Swarm Intelligence, Springer, 2008.
[19] H. Müller, S. Maillet-Marchand, T. Pun, The truth about corel - evaluation in image retrieval, in: CIVR'02: Proceedings of the International Conference on Image and Video Retrieval, LNCS vol. 2383, Springer, London, UK, 2002, pp. 38–49.
[20] L. von Ahn, L. Dabbish, Labeling images with a computer game, in: CHI'04: Proceedings of the ACM-SIGCHI Conference on Human Factors in Computing Systems, Vienna, Austria, 2004, pp. 319–326.
[21] L. von Ahn, R. Liu, M. Blum, Peekaboom: a game for locating objects in images, in: CHI'06: Proceedings of the ACM-SIGCHI Conference on Human Factors in Computing Systems, Montréal, Québec, Canada, 2006, pp. 55–64.
[22] B. Russell, A. Torralba, K.P. Murphy, W. Freeman, LabelMe: a database and web-based tool for image annotation, International Journal of Computer Vision 77 (1–3) (2008) 157–173.
[23] B. Yao, X. Yang, S. Zhu, Introduction to a large-scale general purpose ground truth database: methodology, annotation tool and benchmarks, in: EMMCVPR'07: Proceedings of Energy Minimization Methods in Computer Vision and Pattern Recognition, Hubei, China, 2007, pp. 169–183.
[24] J. Ponce, T.L. Berg, M. Everingham, D.A. Forsyth, M. Hebert, S. Lazebnik, M. Marszalek, C. Schmid, B.C. Russell, A. Torralba, C.K.I. Williams, J. Zhang, A. Zisserman, Dataset issues in object recognition, Toward Category-Level Object Recognition LNCS, vol. 41, Springer, 2006, pp. 29–48 (Chapter 2).
[25] L. Fei-Fei, R. Fergus, P. Perona, Learning Generative Visual Models from Few Training Examples: An Incremental Bayesian Approach Tested on 101 Object Categories, CVPR'04: Workshop on Generative-Model Based Vision, Washington, DC, USA, 2004, p. 178.
[26] S. Agarwal, A. Awan, D. Roth, Learning to detect objects in images via a sparse part-based representation, IEEE Transactions on Pattern Analysis and Machine Intelligence 26 (11) (2004) 1475–1490.
[27] C.K.I. Williams, F. Vivarelli, Using Bayesian neural networks to classify segmented images, in: Proceedings of IEEE International Conference on Artificial Neural Networks, Cambridge, UK, 1997, pp. 268–273.
[28] A. Hanbury, A. Tavakoli-Targhi, A dataset of annotated animals, in: Proceedings of the 2nd MUSCLE/ ImageCLEF Workshop on Image and Video Retrieval Evaluation, Alicante, Spain, 2006, pp. 19–27.
[29] A. Vailaya, A. Jain, H. Zhang, On Image classification: city versus landscape, Pattern Recognition 31 (1998) 1921–1936.
[30] J. Vogel, B. Schiele, semantic modeling of natural scenes for content-based image retrieval, International Journal of Computer Vision 72 (2) (2007) 133–157.
[31] J. Winn, A. Criminisi, T. Minka, Object Categorization by Learned Universal Visual Dictionary, in: ICCV'05: Proceedings of IEEE International Conference on Computer Vision, Beijing, China, 2005, pp. 1800–1807.
[32] J. Shotton, J. Winn, C. Rother, A. Criminisi, TextonBoost: joint appearance, shape and context modeling for multi-class object recognition and segmentation, in: ECCV'06: Proceedings of the 9th European Conference on Computer Vision, Graz, Austria, 2006, pp. 1–15.
[33] D. Martin, C. Fowlkes, D. Tal, J. Malik, A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics, in: ICCV'01: Proceedings of IEEE International Conference on Computer Vision, Vancouver, BC, Canada, 2001, pp. 416–421.
[34] T. Deselaers, A. Hanbury, V. Viitaniemi, et al., Oerview of the ImageCLEF 2007 Object Retrieval Task, Advances in Multilingual and Multimodal Information Retrieval – 8th Workshop of the CLEF, LNCS Vol. 5152, Springer, Budapest, Hungary, 2007 (printed in 2008), pp. 445–471.
[35] T. Deselaers, A. Hanbury, The Visual Concept Detection Task in ImageCLEF 2008, Evaluating Systems for Multilingual and Multimodal Information Access – 9th Workshop of the CLEF, LNCS, Springer, Aarhus, Denmark, 2009, in press.
[36] H.J. Escalante, C. Hernandez, J. Gonzalez, A. Lopez, M. Montes, E. Morales, E. Sucar, L. Villaseñor, Segmenting and Annotating the IAPR TC-12 Benchmark, Tech. rep., Department of Computational Sciences, INAOE, CCC-08-05, Puebla, México, 2008.
[37] A. Hanbury, Review of Image Annotation for the Evaluation of Computer Vision Algorithms, Tech. rep., PRIP, Vienna University of Technology, 102, Vienna, Austria, 2006.
[38] C. Hernandez, L.E. Sucar, Markov Random Fields and Spatial Information to Improve Automatic Image Annotation, in: Proceedings of the 2007 Pacific-Rim Symposium on Image and Video Technology, LNCS vol. 4872, Springer, Santiago, Chile, 2007, pp. 879–892.
[39] C. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.

Please cite this article in press as: H.J. Escalante et al., The segmented and annotated IAPR TC-12 benchmark, Comput. Vis. Image Understand. (2009), doi:10.1016/j.cviu.2009.03.008