Segmenting and Annotating the *IAPR-TC12* Benchmark

H. Jair Escalante, Carlos Hernández-Gracidas, Jesús A. González, Aurelio López, Manuel Montes, Eduardo Morales, Enrique Sucar, Luis Villaseñor

Reporte Técnico No. CCC-08-005
1 de Noviembre de 2008

Luis Enrique Erro 1
Sta. Ma. Tonantzintla,
72840, Puebla, México.

Segmenting and Annotating the *IAPR-TC12* Benchmark

H. Jair Escalante Carlos A. Hernández Jesús A. González Aurelio López
Manuel Montes Eduardo Morales Enrique Sucar Luis Villaseñor

Computer Science Department
National Institute of Astrophysics, Optics and Electronics
Luis Enrique Erro # 1, Santa María Tonantzintla, Puebla, 72840, México
E-mail: {hugojair, carloshg, allopez, mmontesg, emorales, esucar, villasen}@inaoep.mx

Abstract

With the increasing storage of images worldwide, automatic image annotation has become a very active and relevant research area, however, it still lacks a benchmark specifically designed for this task, and in particular for region-level annotation. In this report we introduce the segmented and annotated IAPR-TC12 *benchmark, an extended resource for the evaluation of automatic image annotation (*AIA*) methods. We present a methodology for the manual segmentation and annotation of the images in this collection. The goal of this methodology is to obtain reliable ground truth data for benchmarking* AIA *and related tasks. For annotation, an ad-hoc vocabulary is defined and hierarchically organized. This hierarchy proved to be very useful for obtaining objective and structured annotations. Also, a soft measure for the evaluation of annotation performance is proposed, based on this hierarchy. Statistics on the segmentation and annotation processes give evidence of the reliability of the proposed approach. Visual attributes and spatial relations are also extracted from regions in segmented images. The latter feature will promote research on the use of (spatial) contextual information for* AIA *and image retrieval. The extended collection is publicly available and can be used to evaluate a variety of tasks besides image annotation; this resource can also serve to study the use of automatic annotations for multimedia image retrieval; the latter is a distinctive feature of the collection because, although there are several image annotation benchmarks, there is currently no collection that can be used to effectively evaluate the performance of annotation methods in the task they are designed for (i.e. image retrieval). We outline several applications and raise important questions that might be answered with the annotated collection; motivating research in the areas of image segmentation, annotation and retrieval as well as on machine learning.*

1 Introduction

The task of automatically assigning semantic labels to images is known as automatic image annotation (*AIA*). This research field has been identified as one of the *hot-topics* in the new age of image retrieval [2, 3, 4]. Besides being relatively new, there has been a significant progress in this task since the last decade [5, 6, 7, 8, 9, 10, 11, 12]. However, the lack of a benchmark collection specifically designed for this task, has caused most methods to be evaluated with small collections of unrealistic images [3, 5, 6, 7, 8, 9, 10, 11, 12]. For the same reason, most region-level methods have been evaluated for their image-level labeling ability. As a result, the correspondence performance (i. e. the capability to assign its correct label to a region) of such methods has not been reliably evaluated [7, 13]. The ultimate goal of *AIA* is to provide support for content-based image retrieval methods (*CBIR*).

Basically, the support consists of allowing image collections to be searched by using (*restricted*) natural language statements. This type of search is known as annotation-based image retrieval[1] (*ABIR*) and can be considered a special case of *CBIR* [4]. When *ABIR* and *CBIR* are compared, the former usually outperforms the latter. However, the image collections considered in such comparisons are well suited for *ABIR* (e. g. the Corel collection, see Section 3) and, therefore, the real advantage of *ABIR* over *CBIR* cannot be objectively evaluated.

In order to provide reliable ground-truth data for benchmarking *AIA* we propose the annotation of the *IAPR-TC12* collection, an established image retrieval benchmark [14]. This collection is composed of around 20,000 manually annotated images with free-text descriptions in three languages. In this report we justify the need of an *AIA* benchmark collection and describe *'why'* and *'how'* the annotation of the *IAPR-TC12* results in a suitable *AIA* benchmark collection. Furthermore, we outline interesting applications for the annotated *IAPR-TC12* collection. Once finished, the annotated collection could be used for evaluating *AIA* techniques as well as a source of training data for learning algorithms. Because this is an image retrieval benchmark, the annotated *IAPR-TC12* collection could also be used to objectively compare *ABIR* and *CBIR* techniques; these approaches may be compared to *TBIR* methods as well because the collection is already manually annotated. Further, the usefulness of combining information from different sources can be evaluated (i. e. free text + image + automatic annotations).

The *IAPR-TC12* collection already has several appealing features. Namely, the collection is an established benchmark for several tasks related to image retrieval, is large size, is composed of realistic images of diverse topics and has image-level annotations in three different languages [14]. We propose extending the benchmark by manually segmenting the entire collection and labeling each resulting region according to a carefully defined vocabulary. This extension will allow the evaluation of more multimedia tasks than those currently supported (e. g. region-level and image-level *AIA*, visual concept detection and object retrieval). Furthermore, we have identified several applications for the annotated collection as well as open questions that could be answered with this new resource.

The rest of this report is organized as follows. In the next Section we introduce preliminary information regarding *AIA* and we describe the usual methodology for the evaluation of *AIA* techniques. Next, in Section 3, we review existing collections highly related to the annotated *IAPR-TC12*. In Section 4, we describe the methodology we are following for the annotation of the *IAPR-TC12* collection, statistics on the segmentation and annotation processes are also described in that section. Next, in Section 5, we propose a new evaluation measure for *AIA*, experimental results with a number of classifiers are presented for illustrating the advantages of this measure. Then, in Section 6, we outline some applications and questions that might be answered with the annotated collection, showing the importance of this resource. Finally, in Section 7, we present conclusions derived from this work.

2 Preliminaries

In this section we introduce the *AIA* task and describe the main differences between this and the problem of object recognition, this due to the fact that region-level *AIA* and object recognition are frequently considered the same task. We also describe the usual methodology adopted for the evaluation of region-level *AIA* methods.

2.1 Automatic Image Annotation (*AIA*)

Textual descriptions in images are very useful because, when they are complete (i. e. the visual content of images as well as semantic information are available), standard information retrieval techniques have reported very good results for the task of image retrieval [15, 16, 17]. A complete description of images, however, can only be provided by humans and in some cases humans must have some expertise in the domain of the collection (e. g. a collection of medical images). Unfortunately, manually assigning textual information to images is both,

[1] Notice that *ABIR* is different from text-based image retrieval (*TBIR*), since the latter approach uses text that has been manually assigned to images.

expensive (in terms of human-hour costs) and subjective (due to the annotator criteria). Therefore, there has been a recent increment in the interest on automatically assigning textual information to images.

The task of automatically assigning words to images is known as *AIA*. There are two ways of facing this problem: at the image-level and at the region-level, see Figure 1. In the first case, keywords are assigned to the entire image as a whole, not specifying which words are related to which objects within the image. In the second approach, the assignment of annotations is at region-level within each image, providing a one-to-one correspondence between words and regions. The latter approach provides more information (e. g. spatial relations can be used) to the retrieval task and for this reason we consider it in this work. We should note that any region-level annotation is an image-level annotation, and thus the latter is a special case of the former.

In region-level *AIA* each image I is segmented into N_I regions, $r_{1,\ldots N_I}$. A region is normally represented by a vector of features. Given a fixed annotation vocabulary (i. e. a set of semantic labels) $W = \{w_1, \ldots, w_K\}$, the annotation of a region r_i is the label $w_i \in W$ that better describes r_i. Then, the *AIA* task consists of finding a mapping between r_i's and w_i's (i. e. $w_i = f(r_i)$).

The predominant approach to *AIA* is using semi-supervised latent variable models. Instances of this sort of models are random fields [10], hidden Markov models [11], correspondence latent Dirichlet allocation (*LDA*) [7, 9], probabilistic *LDA* [18], and cross-media relevance models [8], among many others (e.g. [3, 6, 7]). These methods are based on the formalism of graphical models and by introducing latent variables they attempt to model the region-label joint ($P(r_i, w_i)$) or conditional ($P(w_i|r_i)$) probabilities [7, 8, 10, 11]. The main advantage of these methods is that they only require *weakly labeled images* for training, that is, images with associated labels, without the need of the explicit correspondence between regions and labels, see Figure 1 left. The problem with these methods is that correspondence accuracy is low.

Supervised methods, on the other hand, consider the annotation problem as a classification task, with as many classes as labels are in the vocabulary. The goal is to find the best approximation to the map $w_i = f(r_i)$, given a set of N training region-label pairs $D = \{(r_1, w_1), \ldots, (r_N, w_N)\}$. The one-versus-all (*OVA*) formulation has been used in most of these works[2] [12, 21, 23, 24, 25]. Supervised methods have shown to outperform their semi-supervised counterparts. However, they require *strongly labeled images*, that is, images in which the correspondence between regions and labels is provided [12, 21, 23, 25]. The difference in accuracy is significant in favor of supervised methods and therefore it is worthwhile spending some time in building a training set of annotated regions. Alternatively, methods that can take advantage of unlabeled data can be used for obtaining these training samples.

[2]Multiple instance learning (*MIL*) methods are also popular in this task [26]. However, although they are supervised, *MIL* methods are trained on weakly labeled images, as a result the performance is comparable to that of latent variable models [13].

Figure 1. Sample images from three related tasks. From left to right: image-level annotation and region-level annotation (from the Corel subsets of Carbonetto et al. [10], the second image has been segmented with normalized cuts [61]), object recognition-detection (from the *PASCAL VOC-2007* data set [36]) and object recognition (from the Caltech256 data set [20]).

2.2 AIA and Object Recognition

Often region-level *AIA* is considered an object recognition task, this is true to some extent. In both, *AIA* and object recognition, the task is to assign the correct label to a region in a given image. In object recognition, however, training data consists of images where the object to recognize is centered and occupies more than 50% of the image (see Figure 1, rightmost image). Usually, no other object, from the set of objects to recognize, is present in the same image. In region-level *AIA*, training data consists of annotated regions from segmented images. However, the target object is not the main theme of the image. Furthermore, many other target objects are present in the same image (see Figure 1). In consequence, the difference in accuracy is significant. For illustration, accuracy in a benchmark object recognition collection is around 68% for 101 object categories [19], while accuracy of region-level *AIA* in an *AIA* collection is at most 45% for only 22 labels [10, 21, 22, 23].

Another difference lies on the type of objects to recognize. While in object recognition, the objects are very specific entities, like cars, gloves, bottles, soda-cans, and specific weapons, in region-level *AIA* the concepts are more general, for example: building, church, grass, trees and archaeological ruin. The differences between both tasks are due to the applications they are designed for. Object recognition systems are mostly related to surveillance, identification, and tracking, while *AIA* methods are specially designed for image retrieval. The *AIA* task is also related to other computer vision applications as well, including visual concept detection, object retrieval (which can be conceived as image-level *AIA*), and object detection. However, some specific aspects still make *AIA* different from these tasks.

2.3 Evaluation of Region-level AIA Methods

Image-level *AIA* is a special case of the region-level approach. In consequence, performance of region-level methods can be assessed with methodologies designed for image-level techniques. However, evaluating region-level methods by their ability of doing image-level *AIA* is not straightforward. This happens because by using image-level evaluation methodologies the correspondence performance of methods can not be assessed. Correspondence or localization performance is the capability of region-level methods of assigning the correct label to each region. This ability is not important for image-level *AIA* methods because their goal is to assign words to entire images.

One of the reference papers in *AIA* is due to Duygulu et al. [6]. In such paper the *AIA* problem is posed as one of machine translation (from visual-terms to words). The collection used in that work and the methodology proposed for evaluation have been adopted by most of the *AIA* methods [6, 7, 8, 9, 11]. However, while the data and methodology are well suited for assessing the image-level performance of methods, correspondence accuracy can not be objectively evaluated with them. In the rest of this section we focus on the evaluation methodology, the issue of the collection is discussed in Section 3.

The usual methodology for evaluating *AIA* methods is as follows. A set of segmented images with annotations at image-level is split into training and testing subsets. The training subset is used to learn model parameters. Then, the trained model is used for assigning labels to regions in the testing images. The labels assigned to these regions are merged obtaining an image-level annotation for each test image. Queries are formulated by using each of the labels from the annotation vocabulary. Such queries are then used for retrieving images from the testing set of images, annotated with the trained model. An image is said to be relevant to a query, if any of the labels assigned by the trained model to the image is contained in the ground truth image-level annotation of that image. The performance of *AIA* methods is then measured by counting, for each label, the number of relevant images to queries. Standard evaluation measures from information retrieval (e. g. recall and precision) are used for this task. In order to evaluate correspondence performance, Duygulu et al. analyze the correspondence results for 100 images [6] (500 were used in a later work [7]). However, this analysis can only give partial evidence of the true correspondence performance for the methods. Furthermore, one should note that, in most of the cases,

when *AIA* methods are evaluated the latter analysis is not carried out, and authors restrict themselves to evaluating image-level annotation performance [3, 6, 7, 8, 9, 11].

This approach has been adopted by most *AIA* methods regardless of the type of method (i. e. supervised or semi-supervised) and the goal of the method (i. e. region-level or image-level). However, *AIA* methods are evaluated by their ability for assigning labels to images as a whole, and therefore, their correspondence accuracy can not be determined. For illustration consider the annotations shown in Figure 2. As we can see both images are equally good if we just consider the image-level annotation (i. e. *'grass', 'lion'*). In consequence, both annotations would be similarly evaluated with the above described methodology. However, the annotation at the right is completely wrong if we look at correspondence performance. A better, yet simpler, methodology would be averaging the times a region is correctly labeled [13]. This simple measure would adequately evaluate the correspondence performance in both annotations (the annotated *IAPR-TC12* collection will allow this type of evaluation).

The methodology described in this section has been adopted because of the lack of an available benchmark collection with annotations at region-level. In most of the cases researchers collect and manually annotate a small set of images. However, none of these sets have followed an objective methodology, and the resulting subsets are small, with a limited vocabulary and restricted to a type of images. The goal of this research is to provide a reliable benchmark collection that can allow the effective evaluation of correspondence performance for region-level *AIA* methods. Since image-level *AIA*, visual concept detection and object retrieval are closely related tasks, that can be considered special cases of region-level *AIA*, the resultant benchmark could be used for these tasks as well. Object recognition and detection methods could also be evaluated by using the collection, though we must emphasize that it is not specially designed for these tasks. Furthermore, it could be used for benchmarking *ABIR* and for studying the combination of diverse information extracted from images, free-text assigned to images and labels generated with *AIA* methods.

3 Related work

A widely used collection for evaluating *AIA* is the Corel collection. It was first used by Duygulu et al. and it consists of around 800 CD's, 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. For example, in Figure 3, sample images belonging to a common concept are shown. Besides the Corel collection is large enough for obtaining significant results. There are several problems with this collection that make it an unreliable benchmark. First, images are unrealistic because most of them were taken by professional photographers in difficult poses and under controlled situations. Second, it contains the same number of images related to each of the semantic concepts, as a result it is a balanced collection (rarely found in realistic collections). Third, Corel images are annotated at image level, limiting its applicability to image-level methods. Fourth, it has been shown that subsets of this database can easily be tailored to show improvements [27]. Finally, given that the collection is commercial

Figure 2. Sample image from the Corel subsets of Carbonetto et al. [10]. Left: correct region-level *AIA*, right: incorrect region-level *AIA*. Despite the right image has zero (the worst) correspondence performance, both annotations are equally correct at image-level.

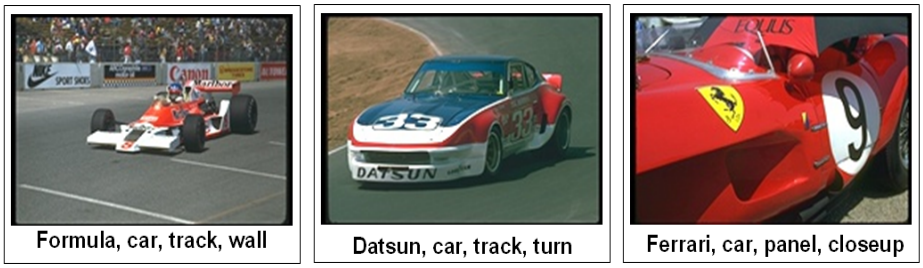

Figure 3. Sample images belonging to the concept *Auto Racing* from the Corel collection.

it is copyright protected; as a result obtaining the collection is expensive and images can not be distributed among researchers. Furthermore, the collection is no longer available, hindering the evaluation for new methods.

Recently, computer games have been used for automatically building resources for several computer vision tasks [28, 29, 30]. *ESP* is an online game that has been used for image-level annotation of real images [29]. The annotation process ensures that only correct (correctness is measured by the agreement of annotators) labels are assigned to images. The volumes of data that this game is able to produce are considerably large. However, images are annotated at image-level and the data is not publicly available [*personal communication*]. *Peekaboom* is another game that uses the annotated images generated with the *ESP* game [30]. In this game, the goal is to provide objects locations and geometry labels. The resulting collection could be used to train learning algorithms for a variety of tasks, including region-level *AIA*. However, since the number of annotators can be of the range of millions there are not objective criteria for neither the annotation process nor the object localization. As a result, the collection is not a reliable collection for benchmarking region-level *AIA*. Furthermore, as with the *ESP* game, data is not publicly available.

A very important effort is being carried out by Russell et al. with the *LabelME* project [31]. This large scale project is collecting large volumes of useful information for diverse computer vision applications. The goal is to obtain segmentation and annotation information by using an online annotation tool. Segmentations are provided by specifying polygons around each object, the annotation vocabulary is defined by the users. The advantages of this collection is that it is publicly available and it is composed of many annotated images. The problem with this collection for evaluating region-level *AIA* is that it has an open vocabulary, so that regions can be assigned any word depending on the annotator intuition, even very different labels may be assigned to the same image. Furthermore, it is no specially designed for region-level *AIA* methods.

Yao et al. are carrying out a valuable eort to create another large-scale benchmark collection; currently, more than 630,000 images have been considered in this project [32]. Manual segmentations are very detailed; segmented objects are decomposed and organized into a hierarchy similar to a syntactic tree in linguistics; information about localization, 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 will be a very useful resource for building visual dictionaries and as training data for learning algorithms. It can 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 effectively assess the impact of AIA methods on multimedia information retrieval.

There are several excellent object recognition collections for benchmarking [34]. Most notably the *Caltech-101* [19], *Caltech-256* [20], and the PASCAL *VOC-2006* [35], and *VOC-2007* [36] collections. The type of images in such collections, however, can not be used for evaluating *AIA* methods (see Section 2.2). Even their use for the evaluation of object recognition methods has been challenged [34]. In the *Caltech* data sets objects are centered and occupy more than 50% of the image, furthermore no other object is present in the images (see Figure 1) [19, 20, 34]. The PASCAL data sets are composed of more realistic images, however, the are only 10 objects in the *VOC-2006* data set and 20 in the *VOC-2007* collection. Moreover, these data sets have been developed

Collection	Size	Labels	Segmentation	Annotation	Caption	Type
Caltech-101[44]	9146	101	Automatic	Image-level	No	Objects
Caltech-256[44]	30608	256	Automatic	Image-level	No	Objects
Animals[45]	200	6	Automatic	Image-level	No	Animals
Scene[46]	1300	60	Automatic	Region-level	No	Nature
Caltech-101[46]	1680	28	Automatic	Region-level	No	Objects
MSRC2[47]	591	21	Automatic	Region-level	No	Objects
PASCAL[47]	5304	10	Automatic	Region-level	No	Objects
Caltech-4[48]	3188	4	None	Image-level	No	Objects
Caltech-101[48]	8677	101	None	Image-level	No	Objects
Events[49]	1040	300	Grid	Region-level	No	Events
LSCOM[50]	61901*	28	Grid	Region-level	No	TV programs
GoogleImages[51]	11182	18	None	Image-level	Yes	Animals, objects
MSRC1[52]	240	9	Automatic	Pixel-level	No	Objects
Caltech-4[53]	3188	4	Automatic	Image-level	No	Objects

Table 1. Analysis of several papers presented at *ICCV'07*, related to *AIA* or that involved image collections. The first column is the name of the image collection. Size is the number of images in the database. Labels is the number of words used for the experiments in the cited paper. Segmentation tells us if the image was manually/automatically segmented or not segmented. Annotation specifies the level of annotation (i. e. image-level, region-level or pixel-level). Caption, reflects if there is any extra meta-data associated to the image. The last column refers to the type of images in the collection. * This is the size of the entire collection, the number of considered images is not detailed in the respective paper [50].

for benchmarking object detection and localization [34] and, therefore, they are not well suited for evaluating region-level *AIA*.

There are only a few collections that can be used for (roughly) effectively evaluating region-level *AIA* methods. Most of these, however, are restricted to specific domains, including cars [37], nature-roadways [38], animals [39], landscape vs. structured classification [24], and natural scene classification [40]. The size of the data sets and the restricted domains make them not adequate for the evaluation of general purpose *AIA*. Winn et al. segmented and annotated a small set of 240 images, considering 9 labels only [41]. In a more recent work, a larger collection with 591 images and 23 labels was created by Shotton et al. [42]. However, the size of these data sets and the number of concepts are not adequate for evaluating *AIA*. Carbonetto et al. have provided three small data sets with a larger number of labels (from 22 to 56) [10]. To the best of our knowledge these are the largest data sets publicly available that have been annotated at a region-level. However, the data sets are still small and come from the Corel collection. Furthermore, the images were segmented with poor quality automatic segmentation methods (e. g. the normalized cuts algorithm [61], see Figures 1 and 2). A very relevant collection for *AIA* is that provided by Barnard et al. [13]. The collection consists of 1041 images taken from a previous study for benchmarking segmentation [43]. A straightforward methodology was proposed and followed for the annotation task. Annotations are specified using WordNet and well defined criteria. The data set is small and the images come from the Corel collection. However, the main contribution of that work is the evaluation methodology that can be used for assessing algorithms even using other collections. Thus, this evaluation methodology can be used with the segmented *IAPR-TC12* collection in conjunction with the one proposed in Section 5.

It is important not only to consider a review of other relevant image collections, but also to analyze the sets being used in current research. By looking at these image sets, it is easy to notice that although there are a number of image collections available, most of the researchers tend to use their own or to take subsets of them. The reason is that these image sets lack features which they consider relevant for their study purpose. Consequently, they find themselves in the necessity of developing or adjusting a collection for their research goal. These *ad hoc* collections have two main problems. The first one is the time consuming task of creating such set of segmented, annotated, and in general, processed images. The second, and probably the most important of them, is the incapacity of objectively comparing two similar works, since they do not use the exact same image collection. The first point tends to slow down an investigation, while the second is more related to the impact of a research, which is limited given the lack of a real standard to measure the degree of importance of the results obtained.

Figure 4. Sample images from the *IAPR-TC12* collection.

Table 1 summarizes some of the features of the image collections used in several papers[3] from the year 2007. In all of the cases, the reference given belongs to the paper where it was used, because in most of them, the creators of the image sets were not the authors of the papers. This table is intended not only to show what image collections are being used in current research, but also how they are actually being used. It can be seen that some names appear more than once along the table; as these collections are used in more than one paper.

From this table we can clearly appreciate that most of the works have used a collection of images about objects. Subsets from the *Caltech* collection are mostly used [19, 20], however, even for those papers using this collection there is not an agreement in the number of labels to consider. None of the collections have been manually segmented and the number of labels considered is small, with exception to *Caltech-256* [20] and the events collection [49]. Note, however, that these collections are designed for specific applications, namely object and event recognition. It should be noted that there is a single collection with meta-data information available [51]. However, meta-data in such collection consists of the HTML text associated with the images. This form of meta-data is not reliable, because HTML text is not controlled. The *IAPR-TC12* collection already offers reliable meta-data for each of the images in the collection. Regarding the size of the collection, only in a single reference have been used more than 20,000 images [44]. The *IAPR-TC12* benchmark was created with the goal of providing a realistic collection of images suitable for a wide number of evaluation purposes; providing images with associated written information [14]. The collection is composed of around $20,000$ images taken from locations around the world and comprising a varying cross-section of still natural images. Most of the images come from a travel company that organizes trips to South-America. The image collection includes pictures of sports, actions, people, animals, cities, landscapes, and many other topics. In Figure 4 sample images from the *IAPR-TC12* collection are shown. Manual annotations in three languages (English, German, and Spanish) are provided with each image. The annotation is at image-level and it is composed of an image identifier, a title, a free-text description of the semantic and visual content of the image, notes with additional information, and a specification of where and when the picture was taken. Further statistics about the *IAPR-TC12* collection can be obtained from [14]. At the moment, the image collection has been used for evaluating cross-language *TBIR* methods, *CBIR* methods, and methods that combine information from both text and images [16, 17]. It has also been used for object retrieval [54], and for measuring word association with application to region-level *AIA* [22]. Currently, it is being used for visual concept detection[4]. The *IAPR-TC12* collection has several positive properties that have established it as a benchmark. Namely, its applicability to several tasks related to image retrieval, it is a large size collection, a wide variety of topics is covered with images, it is composed of realistic images, and it has image-level annotations in three different languages. Because the ultimate goal of *AIA* is image retrieval, the *IAPR-TC12* collection is well suited for benchmarking *AIA* methods. The collection is already annotated at image-level, however, the annotation uses free-text and therefore it can not be used directly for evaluating image-level *AIA*. For the visual concept detection task at *ImageCLEF2008*, about $1,800$ images were annotated with visual concepts. This is an early effort for using the collection for tasks

[3] We considered those papers related to *AIA*, object recognition and image classification published in the proceedings of the 11^{th} IEEE International Conference on Computer Vision.

[4] *http://www.imageclef.org/2008/vcdt*

related to *AIA*. However, only 17 concepts were considered for this task. Given the variety of images this limited vocabulary can not be used for annotating the entire collection. Furthermore, we must emphasize that annotations are available at image-level only. Previously, the *IAPR-TC12* collection was used for the task of object retrieval, using the PASCAL *VOC-2006* collection for training and the *IAPR-TC12* as testing set [54]. However, the number of objects was *10* and accuracy of most of the methods was poor [54]. The results on this task show that specialized collections are required for benchmarking different tasks. As it has been stated by the creators of the *IAPR-TC12* collection [14], a benchmark is not supposed to be static, but evolving. Consequently, it is desirable to continue with the incorporation of additional features to existing benchmarks so they can be useful for evaluating particular tasks, and more important, useful for evaluating real-world tasks. In this work we justify the need of an *AIA* benchmark and propose a methodology for augmenting the *IAPR-TC12* collection. Our work consists of defining an appropriate vocabulary for annotation, developing an adequate concepts hierarchy for annotation and manually segmenting and annotating the entire collection of images.

The segmented and annotated collection can be used for benchmarking both region-level and image-level *AIA* methods. It can be used for the evaluation of closely related tasks, for example for visual concept detection and object retrieval [54]. It will also benefit the multimedia scientific community by allowing the study of combining free-text, labels and image features for different multimedia applications. The annotated collection will allow answering important questions that will benefit the *AIA* and image retrieval communities (e. g. we could answer whether *ABIR* outperforms *CBIR*, and if *AIA* is useful for image retrieval, see Section 6). Segmentations and annotations could be used to categorize the entire collection of images and for obtaining topics and relevance assessments for *AIA* and image retrieval. The collection can be used, to some extent, for the evaluation of segmentation algorithms. It will promote the study of the use of spatial relations for *AIA* and image retrieval. It will be a very useful resource for the machine learning community by providing a large data set for multi-class classification. In this respect, the collection can be used as a resource for training learning algorithms; the collection has many classes (allowing the study of classification with a large number of classes) and it is a very challenging domain (see Section 5.1); regions in images are visually and semantically related, allowing the study of structured prediction methods (one of the current research directions in machine learning [56, 57]); the annotation-hierarchy can be used for studying hierarchical classification and classification with a varying number of classes; finally, it will allow the study of the classification problem in highly imbalanced data sets (see Section 6). The study of the questions and problems mentioned in this section can give rise to a specialized track in the *ImageCLEF* forum or even to a machine learning challenge (see Section 6).

4 The Annotation of the IAPR-TC12 Benchmark

This section describes the methodology adopted for extending the IAPR-TC12 collection. The extension consists of manually segmenting and annotating the entire collection. This is one of the contributions of the paper, because it reflects to what extent the benchmark could be reliable for evaluating *AIA*. The key feature of the methodology is a hierarchical organization of the vocabulary that proved to be very helpful for annotation. This hierarchy is also used by an ad-hoc *AIA* evaluation measure proposed in this work (described in Section 5). Statistics on the segmentation and annotation of the *IAPR-TC 12* collection are also presented. This information gives evidence that the methodology we are following is well suited for the considered collection.

We have segmented and annotated half of the entire collection. The products derived from this work, namely: segmentation masks, annotations, visual features and spatial relations are publicly available for research purposes from the following website *http://ccc.inaoep.mx/˜tia/saiapr.*

4.1 Vocabulary

The vocabulary plays a key role in the annotation process because it has to cover most of the concepts that we may find in the collection of images. 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 5.1). Diverse annotation vocabularies have been considered in different works. Some of them are specific for the type of the images in the collection. In this work we consider a study carried out by Hanbury [55]. There, a list of around 494 labels is obtained by analyzing several *AIA* benchmark collections. We took this word list (*H*-list) as base and adapted it to the *IAPR-TC12* collection.

Using the *H*-list we created our ad-hoc vocabulary as follows. First, we extracted the nouns from the manual annotations of the *IAPR-TC12* collection (*A*-list). We did the same for the textual description of topics for *ImageCLEF 2006* and *2007* (*T*-list). Using these three lists (*H, A,* and *T*) we obtained a candidate list (*C*-list) of labels. The *C*-list was obtained by considering the words appearing in at least two lists. The *C*-list was then manually filtered by considering the following aspects: i) The type of images in the collection, we manually analyzed a large number of images, randomly chosen, and eliminated words that were not present in the images (e. g. *'binoculars','office','printer','sunflower'*). ii) We considered the frequency of occurrence of the words in the *IAPR-TC12* collection, highly frequent words were kept (e. g. *'sky', 'mountain', 'wall', 'table'*); while useless highly-frequent words were not considered (e. g. *'background', 'blue', 'black'*). Finally, words in the *H*-list that were initially dropped from the C-list (e. g. *'herd'*) and words identified by the authors (e. g. *'sky-red'*), that did not appear in any of the three lists, were incorporated into the final list. This last process was iterated several times until the authors totally agreed on the final list. The resulting list of words (*232*) is shown in Table 2.

4.2 Conceptual Hierarchy

When annotating the IAPR-TC12 benchmark the need of a hierarchical organization for the vocabulary arose, this is because a structured annotation was one of the main goals of this work. With this end in mind a hierarchical arrangement of the vocabulary was proposed. The hierarchy was manually defined by the authors after carefully analyzing the images, the annotation vocabulary and the vocabulary of manual annotations. The annotation vocabulary was organized mostly using is-a relations between labels; although, relations like part-of and sort-of were also included. One should note that the hierarchy was defined by thinking on its usefulness for annotation and the representation of images in the IAPR-TC12 collection rather than considering the semantics of labels. The purpose of this structure is to facilitate the annotation process by allowing the annotator to make the correct decision of which is the more adequate label for a given region. Reducing as much as possible, the ambiguities when annotating two not so visually similar regions with the same label, or on the other hand, to help distinguishing two visually similar regions, but different in the concept. The hierarchical organization of concepts is also helpful for the *soft* evaluation of *AIA* methods (see Section 5). In this respect, we propose an ad-hoc evaluation measure, based on the ontology, for the annotated *IAPR-TC12* collection. This hierarchy could also be useful for the organization and categorization of the *IAPR-TC12* benchmark. This would be very helpful for the creation of topics and relevance assessments for multimedia tasks using the *IAPR-TC12* collection.

In Section 4.1, we defined a vocabulary of 232 words. However, the final list of words in the ontology consisted of 275 words. The reason of this is that, when building the ontology, we found ourselves in the need of adding new words (most of them, intermediate words, and just in some cases, leaves), in order to have a better structure in the ontology tree, and also trying to make this ontology as coherent and formal as possible. The list of words added is shown in Table 3. Some other words were, on the other hand, discarded or changed in the same process; the list of such words, and the word or words used to replace them are in Table 4, including a brief explanation of why they were changed.

A general view of the proposed ontology may be seen in Figure 5. The levels in the ontology are shown in

airplane	animal	ant	antelope	apple	astronaut	arctic	baby
ball	balloon	bear	beaver	bed	beetle	bench	bicycle
bird	boat	bobcat	book	bottle	branch	bridge	building
bull	bus	bush	butterfly	cactus	camel	camera	can
canine	cannon	car	caribou	castle	cat	caterpillar	cello
chair	cheetah	child	child-boy	child-girl	chimney	church	church-interior
city	clock	cloth	cloud	column	construction	coral	cougar
couple-person	cow	coyote	crab	crocodile	cup	deer	desk
dish	diver	dog	dolphin	door	dragonfly	eagle	elephant
elk	fabric	face	feline	fence	field	fire	firework
fish	flag	flamingo	floor-carpet	floor-court-tennis	floor-other	floor-wood	flower
flowerbed	food	forest	fountain	fowl	fox	fruit	furniture-other
furniture-wood	giraffe	glacier	glass	goat	grapes	grass	ground
group-persons	guitar	hand	hat	hawk	head	hedgehog	helicopter
herd	highway	hill	horn	horse	house	hut	ice
iguana	insect	island	jaguar	lighthouse	lion	lizard	llama
lobster	log	lynx	mammal	man	monument	motorcycle	mountain
mushroom	musical-instrument	nest	ocean	ocean-animal	octopus	orange	owl
painting	palm	panda	paper	penguin	person	piano	pigeon
plant	polar-bear	pot	primate	public-sign	pyramid	rabbit	rafting
railroad	reflection	reptile	rhinoceros	river	road	rock	rodent
roof	rooster	ruin	sand	saxophone	scorpion	screen	seahorse
seal	semaphore	shadow	sheep	shell	ship	shore	sidewalk
sky-blue	sky-light	sky-night	sky-red (sun-set/dusk)	smoke	snake	snow	space-shuttle
squirrel	stairs	starfish	statue	steam	strawberry	street	sun
surfboard	swimming-pool	table	telephone	tiger	tire	tower	toy
train	trash	tree	trombone	trumpet	trunk	turtle	umbrella
vegetable	vegetation	vehicle	viola	violin	volcano	wall	water
waterfall	wave	whale	window	wolf	woman	wood	zebra

Table 2. Set of words selected with the methodology described in Section 4.1.

aerostatic-balloon	air-vehicles	ancient-building	ape	beach
boat-rafting	cabin	canine	construction-other	curtain
desert	edifice	flock-of-birds	floor	furniture
generic-objects	ground-vehicles	handcraft	humans	jewelry
kangaroo	kitchen-pot	koala	lake	lamp
landscape-nature	leaf	leopard	mammal-other	man-made
man-made-other	mandril	marsupial	monkey	mural-carving
non-wooden-furniture	object (not to be instantiated)	other-entity	pagoda	parrot
person-related-objects	plant-pot	rafter	ruin-archaeological	sand-beach
sand-desert	school-of-fish	sky	trees	vehicles-with-tires
water-reflection	water-vehicles	wooden-furniture		

Table 3. Set of words added to the vocabulary when the ontology was being built.

different colors in order to facilitate their visualization[5].

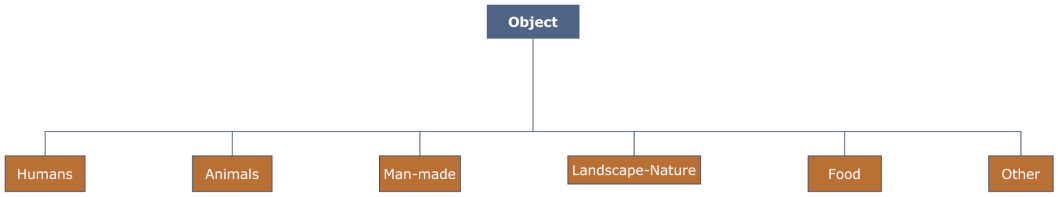

Figure 5. General view of the ontology created for annotating the IAPR-TC12 collection.

The root of the proposed ontology is the word *object* and from this root, four main categories are derived:

1. Animals. Every animal-related word in the ontology is contained in this category. Groups of animals are also instantiable, like *herd* (for mammals), *flock* (for birds), or *school* (for fish). These group labels are to be used when more than one animal belonging to the same group are found together. Animals are also subdivided into 5 subcategories, namely:

   (a) Mammal

   (b) Reptile

   (c) Bird

   (d) Ocean animal

   (e) Insect

   This branch is detailed in Figure 6.

2. Food. Any edible object is placed here. More detail on this branch may be found in Figure 7.

[5]The complete ontology is available for visualization at http://ccc.inaoep.mx/∼tia/ann_ont.htm

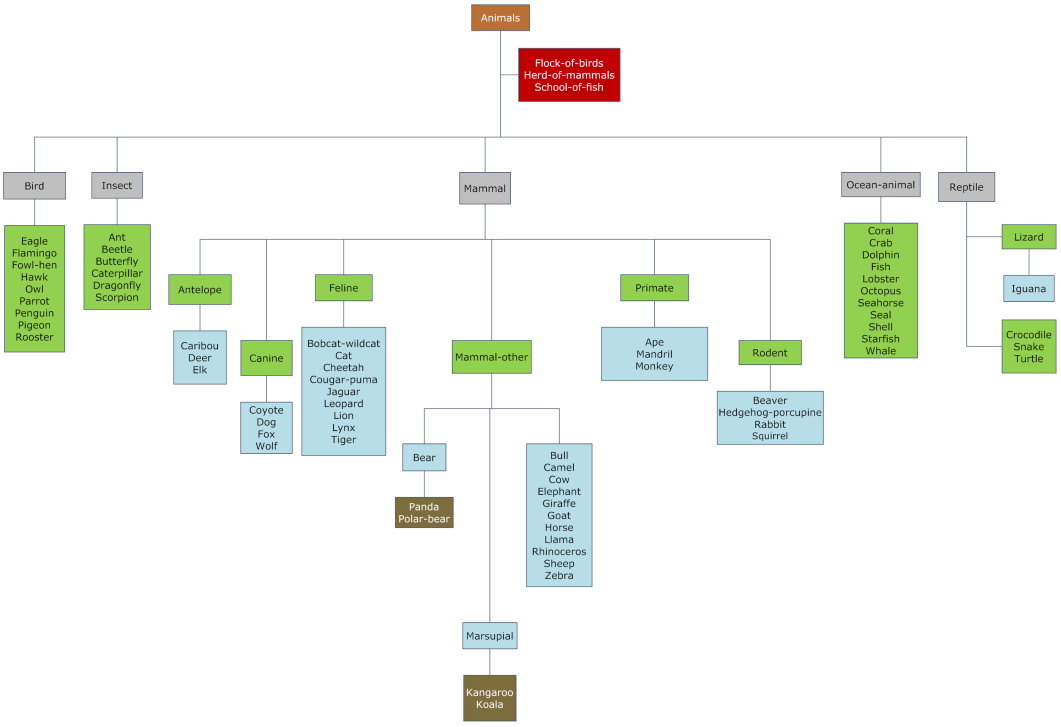

Figure 6. Detailed view of the branch *animal* in the ontology.

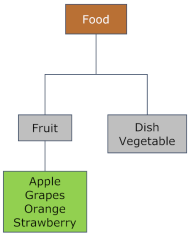

Figure 7. Detailed view of the branch *food* in the ontology.

Original word	substitute word(s)	Explanation
field	grass, vegetation	Field was considered too ambiguous
forest	trees	The only part of the forest that we considered could be instantiated was actually the trees
furniture-wood	wooden-furniture	More appropiate
pot	kitchen-pot, plant-pot	Divided into two words to avoid ambiguities
rafting	rafter	Rafting is more an action than an object (rafters, because of their indumentary were considered of interest)
reflection	water-reflection	The only kind of reflection considered of relevance was water reflection
ruin	ruin-archaeological	More explicit
sand	sand-beach, sand-desert	Also divided in order to avoid ambiguities
shadow	——	We considered shadows of no use for annotation or retrieval

Table 4. Set of words discarded or changed in the vocabulary when the ontology was being built.

3. Humans. Persons are put here. Groups and couples of persons are included as well. Given the large amount of humans in the dataset, body parts, like head, face and hand are considered. Particular cases taking into account age and gender are referred as subcategories of humans. Special cases are astronaut, diver, and rafter, given the special indumentary which might cause confusion if they are labeled together with the rest of the humans. Figure 8 details how this branch is composed.

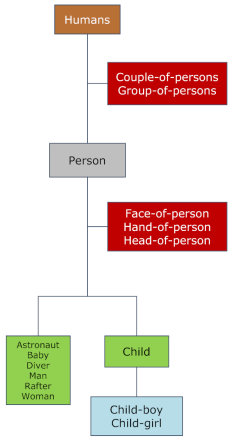

Figure 8. Detailed view of the branch *humans* in the ontology.

4. Man-made. Every object, for which its creation involved humans, is considered in this category. This is a general category whose subcategories actually determine the kind of object. City is another special case, where a group of elements together are taken as the representation, in this case, of a urban settlement. The subcategories of man-made are:

   (a) Construction. Any edification built by man.

   (b) Fabric. Fabric-made objects. Namely: cloth (in general), curtains, and flags.

(c) Floor. Surfaces built by men, including wooden floors, carpets, tennis-courts, or floors made of other material.

(d) Furniture. Either functional or decorative furniture.

(e) Handcraft. Objects which are hand-made, but from an artistic point of view.

(f) Musical-instrument. Different kinds of musical instruments.

(g) Vehicle. Transportations used for carrying humans or objects. These are divided, according to the way they move as:

  i. Air-vehicles

  ii. Ground-vehicles

  iii. Water-vehicles

(h) Man-made-other. Other man-made objects which are not found in the other classifications.

Figure 9 gives more detail regarding this branch.

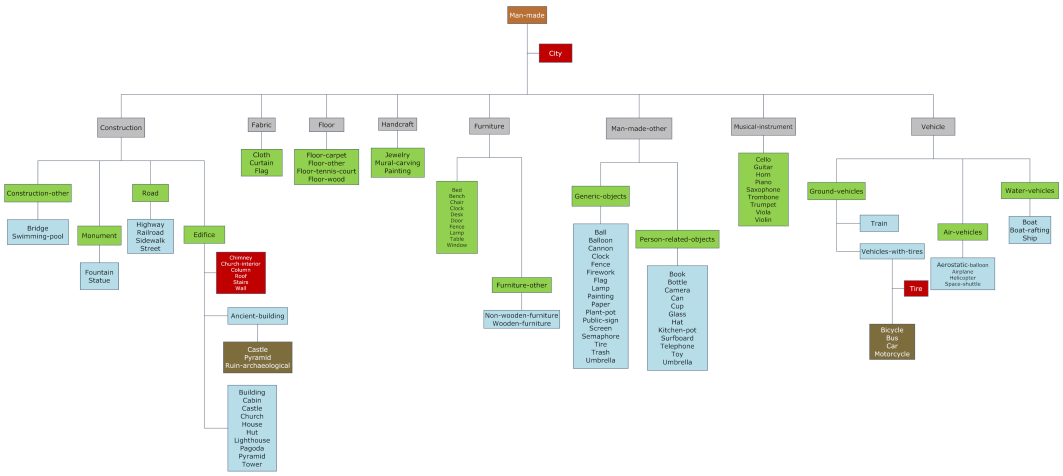

Figure 9. Detailed view of the branch *man-made* in the ontology.

5. Nature. Elements existing in nature (excluding animals and humans) can be found in this category. This also covers more general concepts such as landscapes. A more detailed view of this branch may be observed in Figure 10.

6. Other. Other labels which could not be classified in the other categories, but that were considered relevant at the moment of creating the vocabulary, were included in this category. Examples of these are fire, smoke, and steam. Figure 11 shows the contents of this branch.

Some labels in the ontology, corresponding to composed words, are named in a way that we considered would also allow for a faster search of the concept when annotating (consequently, some of these labels might not seem quite correct from a syntactic point of view). For example, to make reference to *blue sky*, *beach sand*, and *archaeological ruin* we use the labels *sky-blue*, *sand-beach* and *ruin-archaeological* respectively. In some other cases we make the label a composed word to emphasize the class we are refering to, and to avoid ambiguities; an example of this is *floor*, and its descending elements *floor-carpet*, *floor-wood*, *floor-tennis-court*, and *floor-other*, (note that the simple word *other* gives not enough information by itself). Finally, some labels are named using two names (in

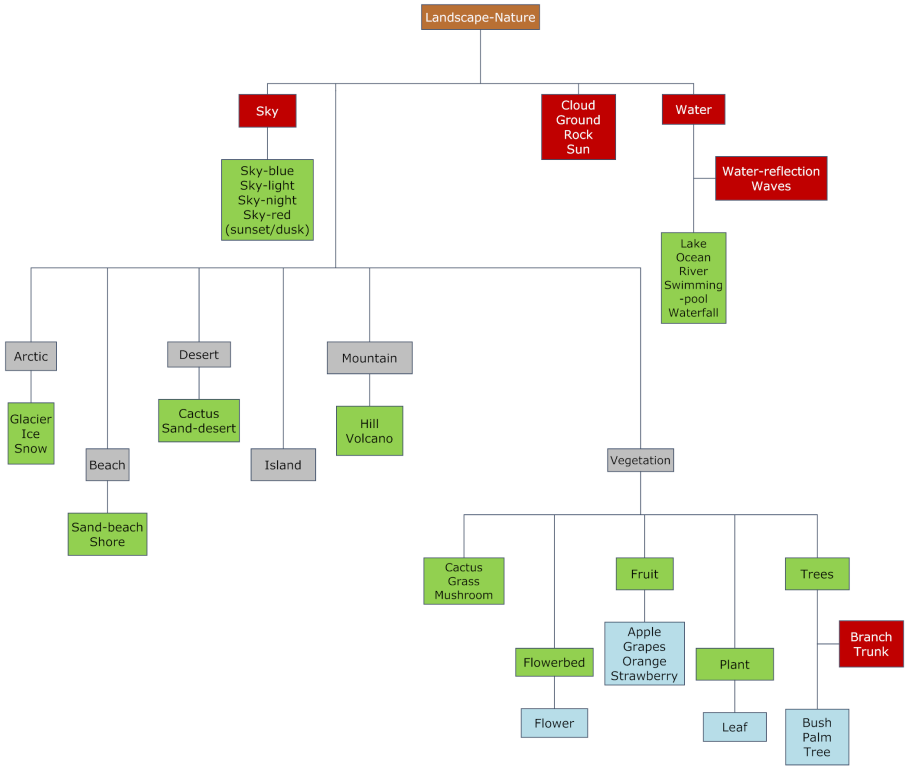

Figure 10. Detailed view of the branch *nature* in the ontology.

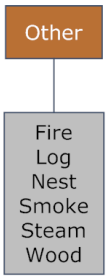

Figure 11. Contents of the branch *other* in the ontology.

a composed word fashion), such as *fowl-hen*, *bobcat-wildcat*, *cougar-puma*, when we considered there was more than one common name for them.

In a strict sense, the structure here introduced is not a real ontology, given the relaxations present in the tree, where some of the branches are considerably more detailed than others (some cases are not detailed at all); and where also some of the labels are not always strictly positioned in the structure according to *is-a* relations. Although the ontology was built mainly based on *is-a* relations, other relations were also included when they were considered relevant; these relations are *part-of* and the concept *group-of*.

The relaxations in the structure, refer to the level of detail, the kind of relations considered, and also, the repetition of some terms in two or more branches of the ontology, in cases where a label can actually be thought of in more than one way. It is important to remark that there is a reason for these relaxations, since this pseudo-ontology is more annotation-oriented than semantic-oriented.

There are similar works based on hierarchical models for vocabulary and annotations, such as the one presented by Yao et al. [32], with what they call a tree list of terms; or the visual concept detection task at *ImageCLEF 2008*, where the labels must be assigned according to a 17-class hierarchy. It is important to mention that these two works we refer to, did not exist by the time our own project started, so our ontology was not based on them. Nevertheless, these works also show a certain tendency to structure labels in a hierarchical, and more structured way than a simple list of apparently unrelated words. This is probably the most important similarity between these works and our proposed ontology, since these independent ideas produced similar structured annotation hierarchies.

The first case shows an extremely detailed structure, where we may see that a big number of its elements have an equivalent term in our ontology (even classes like *man-made*, *mammal*, and *vehicle*, appear in a similar way). However, being it intended to be a multipurpose database (scene and activity classification, aerial images, popular and general objects, and text, to name a few of its subsets), several labels in this structure will not be suitable at all for certain images. This considerably reduces the actual number of possible labels for images of the type of the ones found in the *IAPR-TC12*, reducing the level of detail as well. As an example, while in their classification all of the mammals are put at the same level, in our ontology *mammal* is a four-level branch. In the the second case, most of the elements they defined have some kind of equivalent term in our ontology. However, there is also an important difference in the level of detail, given the fact that their hierarchy consists of three levels at most (though almost all of their labels are in the second level). If the structure is too general, this imposes certain restrictions and forces in some cases to group not so similar regions under the same label, which may affect the annotation ability.

4.3 Segmentation and Annotation Guidelines

Segmenting and annotating an image are tasks which can be performed by completely automatic methods, by human-assisted methods, or even manually. When humans participate in these tasks, they become mostly subjective, since two different persons may segment the same image in completely different ways. This means that a region that seems important to a person, might be irrelevant to another, and the same happens for the shape and size of the region, which will vary from person to person. Even the same individual will probably have different segmentations if they segment the same image more than once at different times.

Because the process of segmenting and annotating images is so hardly standardized, we considered as one of our first priorities to perform this task by a reduced group of persons (in our particular case, the size of the designated group was of 4 persons), so the segmentations and annotations were as consistent as possible. For the same reason, a set of segmentation and annotation guidelines were defined to reduce ambiguities and confusion (it is assumed that the annotator knows the full list of words in advance and that they are familiarized with the segmentation and annotation tool).

The set of guidelines for segmentation is:

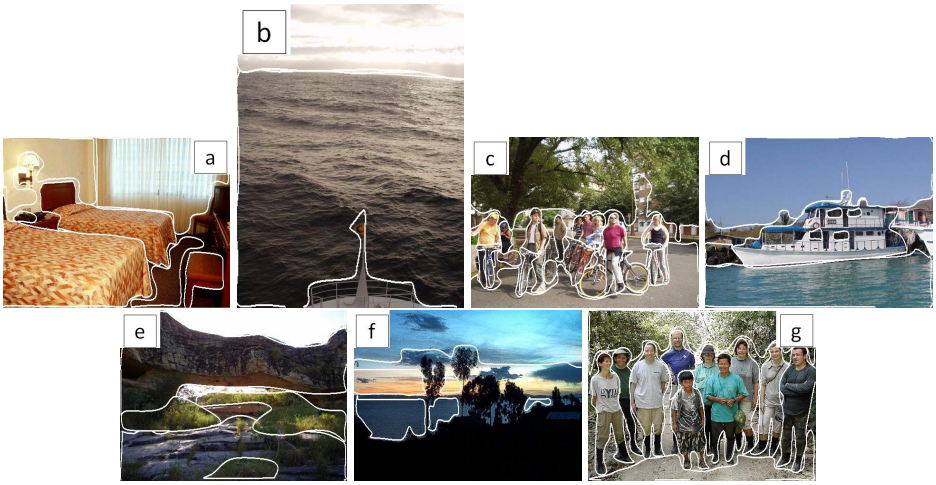

Figure 12. Some images from the IAPR-TC12 collection, used in this case to exemplify the segmentation rules defined. Images are shown manually segmented in order to facilitate distinguishing the regions that are mentioned.

1. Regarding the size of the region to segment, the annotator must avoid segmenting too small regions (with respect to the image size). The annotator must keep in mind that this size is mostly determined by the type of this region. Some objects are expected to look big with respect to the image, while others, even in perfect conditions will most of the times look relatively small with respect to the total image size. This makes this size condition mostly subjective, but nevertheless, necessary, in order to obtain only useful annotated regions.

2. With respect to the object itself, the annotator must avoid segmenting regions where the object is incomplete. This means that at least a third of the object must be visible in the image for the corresponding region to be considered useful.

3. About the contents of the segmented regions, the annotator must consider that this region should contain information from just one object. This means that a segment should not contain parts from more than one entity.

4. Regarding the shape of the object to be segmented (when such shape is relevant), the annotator is advised to try to keep it; otherwise, if the annotator considers the shape of the object to be of no importance, then they may relax the segmentation in a way that it can be performed faster and easier.

5. Likewise, also regarding the shape of the object, when it is considered to be irrelevant, the annotator may divide the object in more than one region if they believe that creating those smaller regions is easier than segmenting the original single one.

6. With respect to the image quality, the annotator must take into account that several images were captured under bad conditions, and in such cases lots of shadows or excessive illumination make it difficult to segment their regions. The recommendation is to avoid such areas and just segment what can be seen without much difficulty by the annotator.

7. About the presence of groups in images, the annotator must take into account if the label related to that group is present in the ontology; if that is the case, they must segment such group as a unit (with its corresponding

group label), and as far as possible, also segment its units individually (with their corresponding individual label as well). The purpose of group labels is mainly helping in cases where a cluttered scene makes difficult the segmentation of each and every individual in a group.

As an example of rule 1, Figure 12 (a), shows objects like *telephones* and *lamps*, which are often small with respect to the image, in contrast with objects like *curtains* and *beds*, which are most of the times of a considerable size also with respect to the image. Figure 12 (b) shows an example of rule 2; here, it is possible to appreciate that although there is a part of what seems to be a *ship* or a *boat*, it must not be segmented. The main problem with this kind of objects is the difficulty to be correctly identified, and since a human would certainly have trouble identifying it, we must expect its features would not be of any help either. For rule 3, in Figure 12 (c), we may see that if we segmented the *group of persons* along with the *bicycles*, we would get wrong feature measurements (because more than one kind of objects would be represented in a single region), and this must be avoided since this tends to cause confusion when regions are automatically classified. An example of rule 4 may be found in Figure 12 (d), where elements such as the *sky* or the *ocean* may be segmented in a relaxed way, but elements like the *boat* must be segmented more carefully, since its shape may help defining a pattern to better represent this class of objects. An example of rule 5 is in Figure 12 (e), where *vegetation* is segmented using more than one single region. In the case of rule 6, for example, Figure 12 (f) shows an example of how excessive shadows make it difficult to distinguish relevant objects in the image. Figure 12 (g) exemplifies rule 8; here the *group of persons* is segmented as a single object, and only the most distinguishable *persons* in the image are individually segmented.

Guidelines for annotation:

1. In reference to the most adequate label to use, the annotator should look for the exact label they believe the region belongs to by searching the ontology top-down, ie., going from the most general labels to the most particular ones. Whenever a suitable label is not found they must resort to the label in the upper level to assign it to the region of interest. This means that a word must be used only when the annotator is completely sure of its meaning and that it really fits the region of interest, otherwise, a more general label should be used. However, the annotator is advised to try to not overuse general-purpose labels.

2. With respect to the branch *other*, in several cases, when a correct label is not found under the category of interest, there is a chance of finding an adequate one by navigating the *other* subbranch (though in some cases it does not exist, and this means that the annotator must resort to an upper level).

An example of rule 1 is shown in Figure 13 (a); in this case, a *parrot* appears in the image, so we must find the corresponding label in the ontology, in case it did not exist, we should use the closest label in meaning (which in this case is *bird*). In the case of rule 2, for example, for labeling the region with the horse in Figure 13 (b) if we look for the label *horse* in the branch *mammal*, we will find that we can actually locate such label under *mammal-other*.

4.4 Segmentation and Annotation Process

For the segmentation and annotation of the *IAPR-TC12* collection we developed an interactive software tool in Matlab^{TM} (called *ISATOOL*, for *Interactive Segmentation and Annotation Tool*). In *ISATOOL* the segmentation-annotation process is as follows. First an image is loaded and the user starts the segmentation process by marking points in the image. Such points should surround objects present in the image. Instead of using straight lines for joining the marked points (just like it is done with the *LabelMe* project [31]) we considered splines in this work. This is because most objects have an irregular shape and using straight lines would require the user to mark more points. The number of images will make impractical such an approach. As we can see, in Figure 14, accurate segmentations can be obtained with this approach.

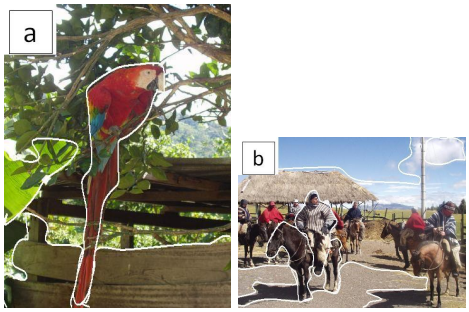

Figure 13. Some images from the IAPR-TC12 collection, used in this case to exemplify the annotation rules defined. Images are also shown manually segmented in order to facilitate distinguishing the regions that are mentioned.

The user is asked to segment the objects in the image (see Section 4.3 for a description of which objects to segment). Once an object is segmented the interface asks the user if they are satisfied with the segmentation of the object. If they are not satisfied, the user can restart the segmentation process until they are satisfied. Otherwise, the user is required to specify the label from the vocabulary that better describes the object.

For annotation, the user pushes a button and the hierarchy described in Section 4.2 is displayed. The user must navigate through the hierarchy and eventually select the word that better describes the identity of the object, see Section 4.3. This feature of *ISATOOL* obligates the annotator to select the label that better describes the object according to our hierarchy. Opposed to selecting the label that better describes the object according to the annotators knowledge [31, 29, 30, 10]; or selecting a word that is related to the object and appears first in a list of words, even when it is not the best label. The user is again asked to confirm if they are satisfied with the annotation of the current region. The segmentation-annotation process is repeated for each object of considerable size within the image.

Once all of the objects have been segmented in the current image the user pushes a button for extracting information from the current image and loading the next image to segment. Such information consists of the segmentation masks and visual features extracted from each region. We store two different segmentation masks per image: the individual and the global segmentation masks. In individual masks we store a single segmentation mask for each of the segmented objects. In the global segmentation mask we generate a single segmentation mask for each image. This is done by combining the individual segmentation masks of the objects in the same image. Regions are collocated in this global mask by considering their area. This way, large regions are put first (on the back) and smaller regions are put later (on the front). Information of spatial relations between regions is also extracted at this time (see Section 4.5).

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 both color spaces *RGB* and *CIE-Lab*, for a total of 27 features. We selected these features because these were considered in previous *AIA* works [10, 21, 22, 23]. However, each user can extract their own set of features since we will provide segmentation masks, and the images of the collection can be obtained as well [14]. As future work we will consider a significantly larger number of features and will perform feature selection for region-level *AIA* and *ABIR*.

4.5 Spatial Relations

Among the potentially useful features we can extract from images we find spatial relations. These features have proved to be very useful for region-level *AIA* [10, 21, 22, 67, 68]. Spatial relations are helpful to know the relative

Figure 14. Sample images from the *IAPR-TC12* collection segmented and annotated as described in Section 4.

position of an object with respect to other objects used as reference in the same scene. Based on the research shown in [21] we extract as additional image features, seven different spatial relations. They are divided into three groups which are:

- Topological relations. Theoretically there are eight possible topological relations between two surfaces in a 2D space, and these are: *disjoint*, *contains*, *inside*, *touching*, *covers*, *covered by*, *overlapped*, and *equal*. In practice, however, only the first four of these relations can be verified between regions in an image. We simplify these relations to two, and then for every pair of regions, we consider they are either *adjacent* or *disjoint*.

- X-relations. X-relations determine how a region is positioned with respect to another, regarding the X axis of the image. One region is *beside* or *X-aligned* to any other region.

- Y-relations. Same as X-relations, but regarding the Y axis of the image. One region is *above*, *below* or *Y-aligned* to another region.

Adjacent is a simplification of the possible topological relations where there is some level of contact between the two images (*contains*, *inside*, and *touching*), while X an Y relations are a subclassification of order relations. These order relations are computed with respect to the center of mass of each region. X-alignment is determined considering a vertical stripe with a proportional width with respect to the image width (and centered at the center of mass of the object of interest); if the center of mass of the region of reference falls into this stripe, it is considered to be X-aligned. Similar procedure is followed to determine Y-alignment.

The spatial relations considered here, are computed in a binary fashion, i.e., for every pair of regions (R_i, R_j) in an image, three spatial relations are computed (one per group) between R_i and R_j and three more between R_j and R_i. Spatial relations are directly computed from the segmentation masks.

We strongly believe that the computed relations are of interest for future research in the use of spatial information among regions in images. It has been proved that the use of context can be helpful in the object detection-recognition [63] and *AIA* tasks [10, 21, 22, 67, 68]; other works using spatial relations for annotation or retrieval are [65, 66]. For object recognition there are several collections that offer spatial information [30, 31, 32]. However, the only collection that provides spatial information for region-level *AIA* is that of Carbonetto et al. [10]. Despite it has been very useful [10, 21, 22], that collection is a subset of the Corel collection and has only used three spatial relations[6] (up, below and next-to). Further, this collection has other limitations described in Section 3. When we complete the segmentation and annotation of the *IAPR-TC12* collection we will provide seven spatial relations for $20,000$ images. This information will support research in the use of spatial relations for region-level *AIA*, and *ABIR*.

4.6 Statistics of the Collection

At the moment, more than 10,000 images out of the 20,000 have been segmented and annotated; the entire collection will be available at the end of 2008. Currently, 47,153 regions compose the segmented collection, for which 239 labels have been used; which represents 86% of the total of the vocabulary. In average 4.71 regions have been segmented per image. The area of each region occupies 15.69% of the image in average. As it is usual in AIA collections, there are some labels that have been used for a considerable number of regions while others have been very sparsely used. Figure 15 plots the number of regions annotated for the 50 more frequent labels. The ten more common labels are *'sky-blue'* (3260), *'man'*, (2349), *'cloud'* (1709), *'group-persons'* (1695), *'ground'* (1648), *'grass'* (1527), *'woman'* (1513), *'rock'* (1447), *'vegetation'* (1418), and *'trees'* (1320). We can clearly

[6]These spatial relations have also been extracted from images in order to complement the benchmark; they will be distributed with the collection. For this task and for that of feature extraction we used the code provided by Peter Carbonetto http://www.ubc.edu.ca/∼pcarbo/

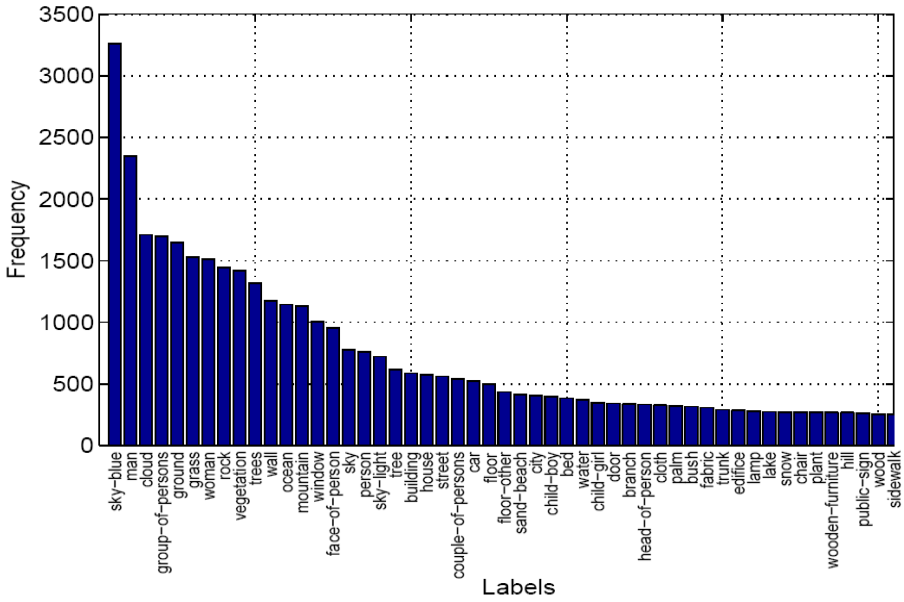

Figure 15. Histogram that shows the the number of regions annotated with each label for the top-50 more common labels.

appreciate that the most common labels agree with the type of images present in the collection (*i.e. pictures of people in vacation trips* [14]). There are 14, 85 and 119 labels that have been used in more than 1,000, 100 and 50 regions respectively.

A total of 218 leaves in the hierarchy have been used for annotation; Table 5 shows the distribution of annotations for the nodes in the first level of the hierarchy. There are more than 21,000 regions annotated with labels below the *'Landscape'* node; it has 46 descendants of which 33 are leaves. More than 15,000 regions have been annotated with labels from the *'Man-made'* node; which is also a large number of regions, however, note that the number of descendants for *'Man-made'* is of 113 nodes, from which 85 are leaves. *'Humans'* is a node with many regions as well, however, its number of descendants is small when compared to the other nodes at the same level. The normalized frequency (third row in Table 5) shows the average number of labels assigned to each descendant in the considered nodes.We can see that the branch *'Humans'* is the one with more annotations per descendant, *'Landscape'* and *'Man-made'* come next. This fact, again, reflects the type of images in the collection: in most of the images appear *'Humans'* since most of the images were taken by/to tourists; most of the pictures were taken in South-American natural places,therefore, there are many images that contain labels from the *'landscape-nature'* branch. Regarding spatial relations, the normalized frequency of spatial relations extracted is described in Table 6. We can see that the most frequent relations are beside and disjoint, with 25.93% and 23.62%, respectively. Note that beside is a generalization of left and right relations, and this is reflected in its frequency. X-alignment and Y-alignment are low frequent relations, with 7.4% and 7.16%, respectively. Finally, the proportions obtained by above and below reflect their symmetry property, both with 13.09%

5 An Evaluation Measure for the Benchmark

As introduced in Section 2.3, most region-level *AIA* methods have been evaluated by their ability of assigning labels at image-level; standard measures from information retrieval have been used for measuring the image-level annotation performance of these region-level *AIA* methods. We have seen (Section 2.3), however, that this evaluation methodology can not provide a reliable estimate of correspondence accuracy. This approach has been adopted by most researchers because of the lack of a suitable collection for benchmarking region-level *AIA*.

Label	*'Animal'*	*'Humans'*	*'Food'*	*'Man-made'*	*'Landscape'*	*'Other'*
Frequency	1,002	9,214	337	15,822	21,802	300
Norm. Freq.	16.43	614.27	48.14	138.79	463.87	50
Descendants	60	14	6	113	46	5
Leaves	47	12	5	85	33	5

Table 5. Distribution of annotations for labels in and below the nodes in the first level of the hierarchy described in Section 4.2. **Frequency** shows the number of regions annotated with labels in or below each node. **Norm. Freq** shows the frequency amortized by the number of descendants of each node.

ID	**Adjacent**	**Disjoint**	**Beside**	**X-alig**	**Above**	**Below**	**Y-alig**
No. Ex	77,306	187,904	206,308	58,902	104,124	104,124	56,962
Norm. Freq.	9.72%	23.62%	25.93%	7.4%	13.09%	13.09%	7.16%

Table 6. Frequency of spatial relations among regions in the extended IAPR-TC12 collection.

Evidently, with the annotated *IAPR-TC12* collection correspondence accuracy of region-level *AIA* methods can be evaluated in a reliable way by using common classification-performance measures, widely used for evaluating machine learning methods (e. g. area under the *ROC* curve, balanced error rate, percentage of misclassifications, squared root error, etcetera). These measures can effectively assess the correspondence performance of annotation methods. However, they can be too *hard* for the current state of the art in region-level *AIA*. Consider, for example, the case in which the correct label for a given region is *'trees'* (plural) and the model under study classifies such a region as *'tree'* (singular), see Figure 14 top leftmost image. In this situation a classification-performance measure would consider the assignment as totally incorrect, despite this prediction is partially correct.

In order to give *partial credit* to those annotations, we propose a new[7] evaluation measure for region-level *AIA* in the annotated *IAPR-TC12* collection. The proposed measure is quite simple, yet it can provide reliable *soft* evaluations of correspondence performance. It is based on the annotation hierarchy introduced in Section 4.2, and it is described in Equation (1).

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

Where $\mathbf{1}_{in-path(t,p)}$ is an indicator function that takes the unit value when both, the predicted label p and the true one t, are in the same path of the annotation ontology. $f_{depth}(x)$ is the depth of label x within the hierarchy. Intuitively, $e_{ontology}(t,p)$ assigns an error value to a label, predicted by a model, proportional to its distance (within the hierarchy) with respect to the ground truth label. The distance should be properly normalized in order to ensure that labels with different depths in the hierarchy are equally evaluated. A predicted annotation will be evaluated as partially good if and only if it appears in the same branch, in the hierarchy, than the correct label. Note that this measure only applies when the true and predicted labels are different, otherwise its value is zero.

For illustration, consider the path of the labels for the above example, i. e. *'tree'* is the predicted label and *'trees'* is the correct one. The respective paths in our ontology are: *object→landscape-nature→vegetation→trees→tree* and *object→landscape-nature→vegetation→trees* (see Figure 10). Clearly, the predicted label, *'tree'*, is a child (i. e. a more specific label) of the correct one, *'trees'*. For this case we would assign a prediction error of $e_{ontology}(trees, tree) = 0.25$ by using Equation (1). If we use a standard classification-performance measure (e. g. $e_{hard} = \mathbf{1}_{t \neq p}$) the error would be of 1. As we can see this measure can provide a reliable estimate of correspondence performance for region-level *AIA* methods. This will be evident in the next Section, where we compare standard classification algorithms by using the evaluation measure proposed in this section.

[7]This measure is not supposed to replace a classification-performance measure for evaluating region-level *AIA*. Instead, it is supposed to provide an alternative or/and complementary evaluation performance, so region-level methods could be evaluated by a combination of a classical classification-performance measure and the one proposed here.

Classifier	Description	Parameters
Zarbi	A simple linear classifier	-
Naive	Naive Bayes classifier	-
Neural	Feedforward Neural Network	Units=10, Shrinkage=0.1, Epochs=50, Balanced=1
SVC	Support Vector Classifier	Coef0=0, Kernel=*Polynomial*, Degree=1, Shrinkage=0.001
Kridge	Kernel ridge regression	Coef0=0, Kernel=*Polynomial*, Degree=1, Shrinkage=0.001
RF	Random Forest	Depth=1, Shrinkage=0.3, Units=100

Table 7. Classifiers from the CLOP toolbox considered in the experiments. The parameters of the classifiers are shown as well, see [62] for more details.

ID	A	B	C	D	E	F	G	H
No. Classes	2	3	4	5	10	25	50	100
No. Examples	3168	4408	5602	6784	11186	18381	24071	28955

Table 8. Subsets considered in the experiments with distribution of classes and the total number of examples considered in each subset.

One should note that this measure is well suited for evaluating region-level *AIA* methods that have as ultimate goal supporting image retrieval. This is because methods that most of the times do not assign the correct label, but assign a word that is highly related to the correct one, will be evaluated high. The labels assigned by this sort of methods will be still very useful because for image retrieval (and information retrieval en general) words related to the correct annotation of the image can be considered an expansion of the annotation. For example, if the correct labels for an image are *'vegetation', 'bush', 'sky', 'vegetation', 'sky'* and an *AIA* method assigns the following labels, *'leaf', 'vegetation', 'sky-blue', 'leaf', 'sky-blue'* (see first column in Table 9). Then, it is very likely that this image will be equally relevant/irrelevant to a query, independently of annotation used (e. g. both annotations describe and image showing vegetation and sky).

5.1 Benchmarking AIA

In this section we show annotation results on region-level *AIA* using a subset of the annotated collection. The goal of these experiments is to illustrate how the annotated collection can be used for region-level *AIA* (and multi-class classification) and to compare the soft evaluation measure to a hard one. We face the problem of *AIA* as one of multi-class classification with as many classes as labels are in our vocabulary. We considered state of the art classifiers over a subset of the total number of annotated regions. The classifiers we considered with their respective parameters are described in Table 7. These classifiers are included in the *CLOP*[8] machine learning toolbox [62].

We considered different subsets according to the frequency of annotated regions per class. A description of the data subsets we considered and the distribution of classes and examples are shown in Table 8. For each data subset the available data were split into, disjoint, training (70%) and testing (30%) sets randomly. In each experiment we trained $k-$classifiers under the *one-versus-all* (*OVA*) formulation (with k as the number of classes). Under this schema one binary classifier is trained for each label. The $k^{th}-$classifier is trained taken as positive those examples from class k and the rest as negative. Despite *OVA* being very simple, it has proved to be competitive to more sophisticated and complex approaches [58, 59]. Furthermore, *OVA* is the widely used approach for supervised *AIA* [12, 21, 22, 23, 24]. One of the main challenges in *OVA* classification is that of choosing a way to combine the outputs of binary classifiers so that error on unseen data is minimized [58, 59]. Since the goal of the experiments is only illustration we adopted the simplest strategy for merging outputs of the individual classifiers. When more than one classifier is triggered we preferred the prediction of the classifier with lower cross-validation error [58]. When no classifier is triggered for a test instance, we assign it the majority class.

For the evaluation of the above classifiers we considered the widely used measure for assessing the performance of classifiers: the percentage of correct classifications (e_{hard}). We compare this *hard* evaluation measure to that described in Equation

[8] http://clopinet.com/CLOP/

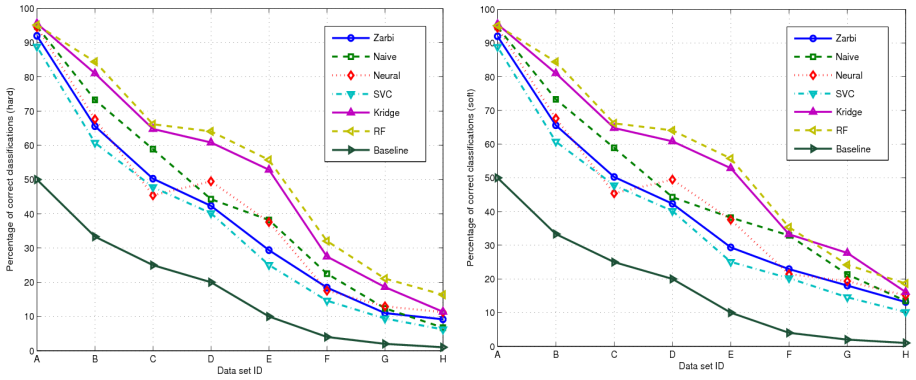

Figure 16. Percentage of correct annotations for the classifiers and data sets described in Tables 7 and 8. Left, results obtained with the hard evaluation measure; right, results obtained with the evaluation measure proposed in this paper. Baseline (the lower line in each plot) is the accuracy we would get by randomly (uniformly) selecting labels.

Ground-truth	Predicted	Error	Ground-truth	Predicted	Error	Ground-truth	Predicted	Error
vegetation	leaf	0.5	tree	vegetation	0.5	plant	vegetation	0.33
bush	vegetation	0.5	trees	vegetation	0.33	tree	vegetation	0.5
sky	sky-blue	0.33	trees	vegetation	0.33	tree	vegetation	0.5
vegetation	leaf	0.5	sky-light	sky	0.33	palm	vegetation	0.5
sky	sky-blue	0.33	sky	sky-blue	0.33	trees	vegetation	0.33
tree	vegetation	0.5	palm	vegetation	0.5	sky	sky-blue	0.33
sky	sky-blue	0.33	vegetation	leaf	0.5	water	waterfall	0.33
vegetation	leaf	0.5	bush	vegetation	0.5	branch	vegetation	0.5
sky	sky-blue	0.33	vegetation	leaf	0.5	trees	vegetation	0.33
road	highway	0.25	tree	vegetation	0.5	vegetation	leaf	0.5

Table 9. Analysis of some labels evaluated as non-erroneous with the $e_{ontology}$ measure and as wrong with the hard evaluation measure. Ground-truth shows the correct label for the region, predicted stands for the label predicted by the model and error represents the $e_{ontology}$ value for the label predicted.

(1) (we say a predicted annotation is correct if it obtains $e_{ontology} < 1$). In Figure 16 we plot the average e_{hard} (left) and $e_{ontology}$ (right) for each class and for each classifier we considered. We can clearly appreciate that both measures decrease similarly as the number of labels considered increases. The plot illustrates the fact that the *hard* evaluation measure is a special case of our *ontology-based* measure. Both plots obtain very similar accuracy values for the data sets *A-E*, (i.e. 5 classes at most). For the data sets *F-H* there is a small difference in the considered evaluation measures. This reflects how the soft measure is indeed less rigid for the evaluation of the classifiers.

In Table 9 we analyze the results of a random run for 100 classes (it is expected that the performance measure works better for a larger number of labels). We consider a subset (randomly selected) of the labels assigned to test regions that were classified as wrong with the hard measure and as correct by using the measure we propose. In the first column we show the ground truth label and in the second one we show the label predicted by the algorithm, the error value of each of these labels is also shown. We can clearly appreciate that the ontology-based measure effectively assess the performance of this classifier. All of the labels considered as partially good are indeed closely related to the correct label. It is clear that if a method obtains good performance under this measure it will be very useful for *ABIR*.

6 Applications of the annotated IAPR-TC12 Benchmark

The *IAPR-TC12* collection is an established image retrieval benchmark that has been used for the evaluation of a wide variety of *CBIR* and *TBIR* methods [16, 17]. Further, it has been used for the evaluation (and for the study) of methods that combine visual and textual information [16, 17]. The work proposed in this paper will complement the collection by

allowing the effective evaluation of several other tasks related to image retrieval. The number of applications for the *IAPR-TC12* collection will be increased as well with our work.

So far we have introduced the annotated *IAPR-TC12* collection and presented statistics about the segmentation-annotation process. In this section we extend the ideas introduced in previous sections, outlining possible applications and important questions that could be answered, to some extend, by using the fully segmented and annotated *IAPR-TC12* benchmark. Applications in the following areas, for the annotated collection, have been identified.

- Image annotation
- Image retrieval
- Image segmentation
- Machine learning

6.1 Image Annotation

The main subject of this research is on the evaluation of region-level *AIA* methods. However, image-level *AIA* methods could be evaluated directly with the annotated collection as well. Because we can always create an image-level annotation from the region-level annotations for regions within the same image. Image-level *AIA* can not be evaluated with the current annotations of the images in the *IAPR-TC12* collection. This is because these annotations were provided from the annotators as free-text. Furthermore, the size of the vocabulary from the *IAPR-TC12* collection is in the order of $10,000$ words for any language when stop words are removed and stemming is applied.

Currently, the *ImageCLEF2008* is running a visual concept detection track. This task can be considered as an special case of image-level *AIA*. The participants are provided with $1,800$ weakly labeled images to train their model, 17 concepts (represented by labels) are considered. Then, participants are required to identify these concepts in images within the *IAPR-TC12*. However, the organizers do not have *complete* relevance assessments, in the sense that they do not know which of the $20,000$ images in the *IAPR-TC12* collection actually contain each concept. It is still not clear how participants will be evaluated in this track. One way would be to obtain relevance assessments in the way it is done with other *ImageCLEF* tasks (i. e. by looking at the submissions of the participants). Another option could be using the free-text annotations of images. However, in any case the organizers do not certainly known in which of the $20,000$ images each concept is actually present. The annotated *IAPR-TC12* collection could be used to obtain *complete* relevance assessments for the concepts proposed in this track, and for many other concepts (those considered in our vocabulary). The same applies for the object retrieval task proposed in *ImageCLEF2007* [54], with the difference that now images from the same collection could be used for training and testing [54].

Spatial relations have been already used to improve region-level *AIA* or to provide better annotations based on these relations [10, 21, 22, 67, 68]. The spatial relations we provide with the annotated *IAPR-TC12* collection could be used for similar experiments intending to use these high-level image features. These spatial relations could also be used to enrich image annotations, adding them as meta-data to the original annotations for complex queries.

In the proposed vocabulary for annotation we have considered several concepts that have been widely studied in the field of object detection and recognition (e. g. *'bottle', 'car', 'face', 'hand'*). With the annotated *IAPR-TC12* collection we will have information of the presence/absence and localization for each concept as well as information of spatial relations. In consequence this collection could be used for the evaluation of methods for object detection and recognition. Techniques for face and skin detection can be assessed as well.

6.2 Image Retrieval

The ultimate goal of both image-level and region-level *AIA* is to allow image collections to be searched by keywords. It is assumed that the use of keywords would be better than query-by-example in a simple *CBIR* system [3, 4]. Some researchers have recently assumed that the use of *AIA* methods in combination with *CBIR* and *TBIR* techniques give better results than using *CBIR* or *TBIR* methods alone [60]. These assumptions, however, have not been proved neither theoretically nor empirically.

The annotated *IAPR-TC12* collection can be used for validating these assumptions. Particularly, we could give empirical evidence on whether the performance of *ABIR* is superior to that of *CBIR*, and how well *ABIR* does with respect to *TBIR* methods. We could also study the benefits of combining labels obtained from *AIA* methods, manual annotations, and visual

information from images for the task of image retrieval. Furthermore, we could provide an empirical bound on the image retrieval (*ABIR*) performance that can be expected by using the manual annotations for the entire *IAPR-TC12* collection.

In Section 4.4 we described the features we extracted from regions. It is possible to extract other features and perform a feature selection process for *AIA* and *ABIR*. Feature selection is the task of choosing a subset from the total of available features so that the predictive performance of methods is not affected (see [64] for a comprehensive introduction). Very often the selection of the adequate subset of features is more important than the classification-retrieval task itself. While this issue has been already approached by Desselears et al. for *CBIR* [69], with the annotated collection we could perform an extensive comparison of features for both tasks: *AIA* and *ABIR*. This is an important issue because almost in every work a particular set of features is considered. In these works the improvements in performance of *AIA* methods could be due to the set of features used instead of the method itself.

Spatial relations could also be used in the retrieval process, in order to give a form of comparing two images, other than low-level image features. Works like [70, 65] explore this possibility by creating retrieval frameworks supporting spatial information. The spatial information we provide, could also be used to allow complex queries on the image set, where the interest could be on finding objects in specific positions with respect to others. Meta-data regarding spatial relations in the image, like the one we suggest in Section 6.1, could be used for these more complex queries, based not only on the original data, but also on these spatial relations.

6.2.1 An ImageCLEF track

The applications and questions mentioned above can motivate the proposal of a region-level *AIA* track at *ImageCLEF*. Also, it is possible to propose an *ABIR* track and allow the use of *AIA* methods for the current ad-hoc photographic retrieval task. Actually, this is the tendency shown by the *ImageCLEF2008* organizers with the introduction of the visual concept detection task (image-level annotation) for giving support to participants in the retrieval task. The proposal of a multi-modal (labels, text, and images) image retrieval track is another feasible application for the annotated collection.

By means of these tracks we could study diverse interesting topics. For example, research advances in region-level *AIA* and multi-class classification (see Section 6.4); what combination of methods (e. g. *CBIR, ABIR* and *TBIR*) performs better for image retrieval; the benefits of combining information from multiple modalities for the task of image retrieval; the benefits of considering spatial relations into the annotation and retrieval process, among many others.

6.3 Image Segmentation

Most segmentation benchmarks offer many different segmentations for the same image [43], so that the overlap between the regions obtained with segmentation algorithms and those from ground truth data are compared. Because the *IAPR-TC12* collection will be manually segmented, segmentation performance can be evaluated, to some extend, with this collection. The proposed methodology states that most of the objects of considerable size, within our vocabulary, should be segmented (see Section 4.3); also, several images have been over-segmented (see Section 4.4). In consequence, the segmented *IAPR-TC12* collection can give us an estimate of segmentation performance for automatic methods with respect to the segmentation and annotation rules we defined (see Section 4.3).

Just as the annotated *IAPR-TC12* collection can be used for providing a bound in the retrieval performance of *ABIR*, the same collection can be used to provide a bound in the annotation accuracy of region-level *AIA* methods if we could have a *perfect* segmentation. That is, *which annotation accuracy can be obtained with manual segmentations?*. This is an interesting question because we often make the assumption that low accuracy of region-level *AIA* methods is due to the poor segmentation quality of images. Furthermore, we could study the issue of what would happen when you train with manually segmented regions and test with images that have been segmented with automatic algorithms (e. g. normalized cuts or grid patches [61, 10]).

6.4 Machine Learning Applications

The annotated collection could be used to bridge the gap between the machine learning and information retrieval communities. In the rest of this section we outline some important learning applications of the annotated collection.

Multi-class classification is the generalization of the binary classification problem to the case in which an observation may belong to more than two classes [58, 59]. This problem is of great interest in the machine learning community, because many interesting real-world problems involve multi-class classification. For example, text categorization, pixel-wise image

classification, land covering classification, star-galaxy classification, genotype categorization, and digit recognition are all multi-class classification problems. The machine learning community could use the annotated collection for studying the classification problem for a large number of labels. One should note that the number of classes could be reduced if we consider as classes the nodes at a particular level of the hierarchy of concepts. To the best of our knowledge, no research on the effect on the performance of classifiers with an increasing number of classes has been performed. The concepts hierarchy can be used to such end by increasing the number of labels according its levels. Also, little research on hierarchical classification has been performed for *AIA*, the hierarchy could also be useful for studying this issue.

Currently, structured prediction is one of the *hot-topics* on machine learning [57, 56]. The problem consists of making predictions of composed objects and sequence of objects that have strong dependency relationships among them [57, 56]. Structured domains include, machine translation (words in sentences are related to each other in any language), document markup (Web pages are related by hyperlinks), hand written word recognition (digital letters are related to each other in words), among many others. Region-level *AIA* can be considered a semi-structured domain because regions in the same image are usually closely related (visually and semantically) to each other. Several methods for semi-structured prediction have reported successful results in this task [10, 21, 22]. The application of structured prediction approaches could be useful as well. Since we provide manual annotations for a large number of regions in images together with spatial relations information, the collection could be used for exploring the application of structured prediction methods to the problem of *AIA*. The collection can also be used for doing spatial data mining.

As we can see in Section 4.6 the number of training samples per class is highly imbalanced. This is a feature present in most realistic data domains and, therefore, it is an important machine learning problem. The annotated *IAPR-TC12* collection can be used for benchmarking this important problem directly.

The machine learning applications identified in this section can be studied in a machine learning challenge or competition. Machine learning competitions promote the collaboration among researchers and can advance significantly the state of the art in some tasks. Up to this day, there have been many successful machine learning challenges, ranging from feature and variable selection[9], performance prediction[10], model selection[11], agnostic-learning vs prior-knowledge [12], causality discovery[13], reinforcement learning[14], time series prediction[15] and classification[16].

7 Conclusions

The *IAPR-TC12* image collection is an established image retrieval benchmark that has several attractive features. Namely, it is a large size collection, it is composed of diverse and realistic images, it offers textual annotations in three languages and it provides relevance assessments for evaluating image retrieval performance. A benchmark, however, is not supposed to be static, but to evolve according to the needs of the task it is designed for and the emergence of new related-issues. In this paper we introduced the segmented and annotated *IAPR-TC12* collection, an extension to the benchmark that will increase the number of tasks that can be *benchmarked* with the collection. This extension will also augment significantly the number of applications for the *IAPR-TC12* collection.

We described a methodology for the manual segmentation and annotation of the images in the *IAPR-TC12* collection. The methodology includes the definition of an ad-hoc annotation vocabulary and well defined criteria for objectively segmenting and annotating images. A hierarchical organization of the vocabulary is proposed for the structured and objective annotation of images. This pseudo-ontology is an added value to the collection because it can be used for the categorization of images and for the identification of relevance assessments for image retrieval and related tasks. For machine learning, the hierarchy can promote research on multi-class classification (varying the number of classes), hierarchical classification and structured prediction as well. Visual attributes and spatial relations are extracted from regions in segmented images. The latter feature is another contribution to the collection since it will allow the study of the benefits of using spatial information for the tasks of

[9] http://www.nipsfsc.ecs.soton.ac.uk/

[10] http://www.modelselsect.inf.ethz.ch/

[11] http://clopinet.com/isabelle/Projects/NIPS2006/

[12] http://www.agnostic.inf.ethz.ch/

[13] http://www.agnostic.inf.ethz.ch/

[14] http://rl-competition.org/

[15] http://www.neural-forecasting-competition.com/

[16] http://home.comcast.net/∼nn-classification/

AIA and image retrieval. Statistics on the annotation-segmentation process[17] give evidence that the methodology we follow is reliable and well suited for the *IAPR-TC12* collection.

A new evaluation measure for *AIA*, based on the hierarchy, is proposed. This measure is an alternative to the *hard* evaluation measures that consider a predicted label as correct if and only if it is the same as the ground truth label. While *hard* measures effectively assess the performance of *AIA*, they can be too rigid for the current state of the art in *AIA*. Further, some predicted labels deserve partial credit because of its relationship (visual and/or semantic) with the ground truth labels (e. g. *trees* and *tree*). Initial experimental results give evidence that the measure effectively assess *AIA* performance, and that labels evaluated as correct, with the *soft* measure, can be very useful for image retrieval.

Another important contribution of this work is the identification of applications for, and important questions that can be answered with, the annotated *IAPR-TC12* collection, that will help to advance in different research areas. The following list summarizes specific applications we have identified for the annotated collection:

1. Benchmarking:
   (a) Region-level and image-level *AIA*,
   (b) Visual-concept detection and object retrieval,
   (c) Object detection-recognition, face and skin detection.
   (d) *ABIR*,
   (e) Segmentation.
2. Studying the combination of the following information for image retrieval:
   (a) Manual annotations,
   (b) Automatically generated labels,
   (c) Visual features from images and/or regions,
   (d) Spatial relations.
3. Studying the advantages of using spatial relations for:
   (a) *AIA*,
   (b) *ABIR*,
   (c) *CBIR*.
4. Comparing and combining different approaches to image retrieval:
   (a) *CBIR*,
   (b) *TBIR*,
   (c) *ABIR*.
5. Using the hierarchy of concepts for:
   (a) Organization and categorization of the collection,
   (b) Topic creation,
   (c) Relevance assessments creation.
6. Solving machine learning problems, such as:
   (a) Training learning algorithms,
   (b) Multi-class classification with varying number of classes,
   (c) Hierarchical classification,

[17] We have segmented and annotated, approximately, one quarter of the entire collection. We expect to have segmented and annotated the full collection by the end of this year.

(d) Structured prediction,

(e) Classification on imbalanced data sets,

(f) Spatial data mining,

(g) Feature selection.

It is clear that the annotated collection will be useful not only for benchmarking *AIA* and *ABIR*. It will also offer benefits and promote research on other multimedia and machine learning sub-areas. The above research directions can motivate the proposal of an *ImageCLEF* track or a machine learning challenge. The collection can also be used for reinforcing current *ImageCLEF* tracks that make use of the un-annotated *IAPR-TC12* collection.

Statistics on the segmentation-annotation process and initial annotation results show that the methodology we follow is straightforward and give evidence that the resulting annotated collection will be a reliable benchmark for *AIA* and related tasks. Moreover, the contributions mentioned in this section show that the annotated *IAPR-TC12* benchmark will be a very important resource for the multimedia and machine learning areas, motivating research in several directions. Once finished, the collection will be made publicly available[18] for academic and research purposes.

References

[1] H. J. Escalante, M. Grubinger, M. Montes and L. E. Sucar, "Towards a Region-Level Automatic Image Annotation Benchmark," *Proc. of the Third MUSCLE / ImageCLEF Workshop on Image and Video Retrieval Evaluation*, Budapest, Hungary, 2007.

[2] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Image Retrieval: Ideas, Influences, and Trends of the New Age," *ACM Computing Surveys*, 2008. (to appear).

[3] J. S. Hare, and P. H. Lewis, and P. G.B. Enser, and C. J. Sandom, "Mind the Gap: Another Look at the Problem of the Semantic Gap in Image Retrieval", *Proc. of Multimedia Content Analysis, Management and Retrieval*, Vol. 6073, SPIE, San Jose, California, USA, 2006.

[4] M. Inoue, "On the Need for Annotation-based Image Retrieval", *IRiX '04: Proc. of the Workshop on Information Retrieval in Context*, SIGIR, Sgeffield, UK, 2004.

[5] Y. Mori, and H. Takahashi, and R. Oka, "Image-to-word transformation based on dividing and vector quantizing images with words", *MISRM'99 First Intl. Workshop on Multimedia Intelligent Storage and Retrieval Management*, 1999.

[6] P. Duygulu and K. Barnard and J. F. G. de Freitas and D. A. Forsyth, "Object Recognition as Machine Translation: Learning a Lexicon for a Fixed Image Vocabulary", *ECCV '02: Proc. of the 7th European Conf. on Comp. Vis.-Part IV*, 97–112, LNCS 2353, Springer-Verlag, London, UK, 2002.

[7] K. Barnard, and P. Duygulu, and N. de Freitas, and D. Forsyth, and D. Blei, and M. I. Jordan, "Matching Words and Pictures", *J. Mach. Learn. Res.*, Vol. 3, 1107–1135, 2003.

[8] J. Jeon and V. Lavrenko and R. Manmatha, "Automatic image annotation and retrieval using cross-media relevance models," *Proc. of the 26th Intl. ACM-SIGIR Conf. on Research and Development on Informaion Retrieval*, 119–126, Toronto, Canada, 2003.

[9] D. Blei, "Probabilistic Models of Text and Images," *Ph. D. thesis Universtity of California, Berkley*, 2004.

[10] P. Carbonetto, and N. de Freitas, and K. Barnard, "A Statistical Model for General Context Object Recognition," *ECCV '04: Proc. of the 8th Europ. Conf. on Comp. Vis.*, 350–362, LNCS 3021, Springer-Verlag, Canada, 2004.

[11] G. Iyengar and P. Duygulu and S. Feng and P. Ircing and S. P. Khudanpur and D. Klakow and M. R. Krause and R. Manmatha and H. J. Nock and D. Petkova and B. Pytlik and P. Virga, "Joint visual-text modeling for automatic retrieval of multimedia documents," *Proc. of the 13th ACM Intl. Conf. on Multimedia*, 21–30, Singapore, 2005.

[12] G. Carneiro and A. B. Chan and P. J. Moreno and N. Vasconcelos, "Supervised Learning of Semantic Classes for Image Annotation and Retrieval," *IEEE Trans. on PAMI*, Vol. 29-3, 394–410 2007.

[18] Previous authorization of the creators of the *IAPR-TC12* collection.

[13] K. Barnard, Q. Fan, R. Swaminathan, A. Hoogs, R. Collins, P. Rondot, J. Kaufhold, "Evaluation of Localized Semantics: Data, Methodology, and Experiments," *Int. J. Comput. Vis.*, Vol. 77: 199–217, 2008.

[14] M. Grubinger, P. Clough, and H. Müller and T. Deselaers, "The IAPR TC-12 Benchmark: A New Evaluation Resource for Visual Information Systems," *Proc. of the Intl. Workshop OntoImage'2006 Language Resources for CBIR*, 2005.

[15] A. Goodrum, "Image Information Retrieval: An Overview of Current Research," *Journal of Informing Science*, Vol 3-2, 2000.

[16] P. Clough, and M. Grubinger, and T. Deselaers, and A. Hanbury, and H. Müller, "Overview of the ImageCLEF 2006 photographic retrieval and object annotation tasks," *Working Notes of the CLEF Workshop*, 2006.

[17] P. Clough, and M. Grubinger, and T. Deselaers, and A. Hanbury, and H. Müller, "Overview of the ImageCLEF 2007 photographic retrieval task," *Working Notes of the CLEF Workshop*, 2007.

[18] F. Monay and D. Gatica-Perez, "Modeling Semantic Aspects for Cross-Media Image Indexing," *IEEE Trans. on PAMI*, Vol 29-10, 1802–1817, 2007.

[19] L. Fei-Fei, R. Fergus and 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 Vis.*, 2004.

[20] G. Griffin, G. Holub, P. Perona, "The Caltech-256", *Caltech Technical Report*, 2007.

[21] C. Hernández-Gracidas and L. E. Sucar, *Markov Random Fields and Spatial Information to Improve Automatic Image Annotaion.* Proc. of the the 2007 Pacific-Rim Symposium on Image and Video Technology, 879–892, Springer-Verlag 2007.

[22] H. J. Escalante, and M. Montes and L. E. Sucar, "Word Co-occurrence and Markov Random Fields for Improving Automatic Image Annotation," *Proc. of the 18th British Machine Vis. Conf.*, Vol. 2, 600-609, Warwick, UK, 2007.

[23] H. J. Escalante, and M. Montes and L. E. Sucar, "Multi-Class PSMS for Automatic Image Annotation," *Submitted to Applications on Swarm Intelligence*, 2008.

[24] A. Vailaya, and A. Jain, and H. Zhang, "On Image Classification: City versus Landscape," *Pattern Recognition*, Vol. 31, 1921–1936, 1998.

[25] B. Bradshaw, "Semantic based image retrieval: a probabilistic approach," *Proc. of the 8th ACM Intl. Conf. on Multimedia*, California, USA, 167–176 2000.

[26] O. Maron, T. Lozano-Perez, "A framework for multipleinstance learning," *NIPS '98: Proc. of the Neural Information Processing Systems Conf*, MIT Press, 1998.

[27] H. Müller, S. Maillet-Marchand, and T. Pun "The Truth about Corel - Evaluation in Image Retrieval," *CIVR 02: Proc. of the Int. Conf. on Image and Video Retrieval*, Londond, UK, LNCS 2383, Springer-Verlag, 38–49, 2002.

[28] L. von Ahn, "Games with a Purpose," *IEEE Comp. Magazine*, June 2006.

[29] L. von Ahn and L. Dabbish, "Labeling Images with a Comp. Game," *CHI 04: Proc. of the ACM Conf. on Comp.-Human Interaction*, 319–326, Vienna, Austria, 2004.

[30] L. von Ahn, R. Liu and M. Blum, "Peekaboom: A Game for Locating Objects in Images," *CHI 06: Proc. of the ACM Conf. on Comp.-Human Interaction*, 55–64, Montréarl, Québec, Canada, 2006.

[31] B. Russell, A. Torralba, K. P. Murphy, and W. Freeman, "LabelMe: a Database and Web-Based Tool for Image Annotation," *MIT AI LAB MEMO AIM-2005-025*, 2007.

[32] B. Yao, X. Yang, and S. Zhu, "Introduction to a Large-Scale General Purpose Ground Truth Database: Methodology, Annotation Tool and Benchmarks," *EMMCVPR '07: Proc. of Energy Minimization Methods in Comp. Vis. and Pattern Recognition*, 169–183, Hubei, China, 2007.

[33] S. C. Zhu and D. Mumford, "A Stochastic Grammar of Images," *Foundations and Trends in Comp. Graphics and Vis.*, Vol 2, No. 4, 259–362, 2006.

[34] 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, and A. Zisserman "Dataset Issues in Object Recognition," *Chapter 2 in Toward Category-Level Object Recognition*, LNCS 4170, Springer-Verlag, 29-48, 2006.

[35] M. Everingham, A. Zisserman, C. K. I. Williams, L. Van Gool, "The PASCAL Visual Object Classes Challenge 2006 (VOC2006) Results," *http://www.pascal-network.org/challenges/VOC/voc2006/results.pdf*, 2006.

[36] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results," *http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html*, 2006.

[37] S. Agarwal, Awan, and D. Roth, "The UIUC Image Database for Car Detection," *Available from http://l2r.cs.uiuc.edu/simcogcomp/Data/Car/*, 2002.

[38] C. K. I. Williams and F. Vivarelli, "Using Bayesian neural networks to classify segmented images," *Proc. of IEEE Intl. Conf. on Artificial Neural Networks*, 268–273, 1997.

[39] A. Hanbury, A. Tavakoli-Targhi, "A Dataset of Annotated Animals," *Proc. of the Second MUSCLE / ImageCLEF Workshop on Image and Video Retrieval Evaluation*, Czech Repulic 2006.

[40] J. Vogel and B. Schiele, "Semantic Scene Modeling and Retrieval for Content-Based Image Retrieval," *Intl. Int. J. Comput. Vis.*, Vol. 72:2, pp. 133-157, April 2007.

[41] J. Winn, A. Criminisi and T. Minka, "Object Categorization by Learned Universal Visual Dictionary," *ICCV '05 Proc. IEEE Intl. Conf. on Comp. Vis.*, 1800–1807, Beijing, China, 2005

[42] J. Shotton, J. Winn, C. Rother, and A. Criminisi, "TextonBoost: Joint Appearance, Shape and Context Modeling for Multi-Class Object Recognition and Segmentation," *ECCV '06: Proc. European Conf. on Comp. Vis.*, 1–15, Graz, Austria, 2006

[43] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," *ICCV '01 Proc. IEEE Intl Conf. on Comp. Vis.*, Vol. II, pp. 416-421, 2001.

[44] A. Bosch, A. Zisserman, and X. Munoz, "Image Classification using Random Forest and Ferns," *Proc. of the 11th IEEE Intl. Conf. on Comp. Vis.*, Brazil, 2007

[45] N. Ahuja, and S. Todorovic, "Learning the Taxonomy and Models of Categories Present in Arbitrary Images," *ICCV '07: Proc. of the 11th IEEE Intl. Conf. on Comp. Vis.*, Brazil, 2007

[46] L. Cao, and L. Fei-Fei, "Spatially Coherent Latent Topic Model for Concurrent Segmentation and Classification of Objects and Scenes," *ICCV '07: Proc. of the 11th IEEE Intl. Conf. on Comp. Vis.*, Brazil, 2007

[47] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie, "Objects in Context," *ICCV '07: Proc. of the 11th IEEE Intl. Conf. on Comp. Vis.*, Brazil, 2007

[48] A. Kapoor, K. Grauman, R. Urtasun, and T. Darrell, "Active Learning with Gaussian Processes for Object Categorization," *ICCV '07: Proc. of the 11th IEEE Intl. Conf. on Comp. Vis.*, Brazil, 2007

[49] L. Li, and L. Fei-Fei, "What, where and who? Classifying events by scene and object recognition," *ICCV '07: Proc. of the 11th IEEE Intl. Conf. on Comp. Vis.*, Brazil, 2007

[50] J. Liu, and M. Shah, "Scene Modeling Using Co-Clustering," *ICCV '07: Proc. of the 11th IEEE Intl. Conf. on Comp. Vis.*, Brazil, 2007

[51] F, Schroff, A. Criminisi, and A. Zisserman, "Harvesting Image Databases from the Web," *ICCV '07: Proc. of the 11th IEEE Intl. Conf. on Comp. Vis.*, Brazil, 2007

[52] J. van de Weijer, C. Schmid, and J. Verbeek, "Using High-Level Visual Information for Color Constancy," *ICCV '07: Proc. of the 11th IEEE Intl. Conf. on Comp. Vis.*, Brazil, 2007

[53] H. Deng, W. Zhang, E. Mortensen, T. Dietterich, and L. Shapiro, "Principal Curvature-Based Region Detector for Object Recognition," *CVPR '07: Proc. of the IEEE Conf. on Comp. Vis. and Patt. Recog.*, 1–8, 2007

[54] T. Deselaers, A. Hanbury, V. Viitaniemi, et al., "Overview of the ImageCLEF 2007 Object Retrieval Task," *Working Notes of the CLEF Workshop*, 2007.

[55] A. Hanbury, "Review of Image Annotation for the Evaluation of Comp. Vis. Algorithms," *Technical Report PRIP, Vienna University of Technology*, 102, 2006.

[56] L. Getoor and B. Taskar editors, *Introduction to Statistical Relational Learning.* MIT, Press, 2007.

[57] G. Bakir, T. Hofmann, B. Schölkopf, A. Smola, B. Taskar and S. V. N. Vishwanathan editors, *Predicting Structured Data.* MIT, Press, 2007.

[58] C. Bishop, *Pattern Recognition and Machine Learning.* Springer, 2006.

[59] R. Rifkin, and A. Klautau, "In Defense of One-Vs-All Classification," *J. Mach. Learn. Res.*, 5(Jan):101–141, 2004.

[60] H. Jair Escalante, and C. A. Hernandez, and H. Marin-Castro, A. Lopez, and E. Morales and L. E. Sucar, and M. Montes and L. Villasenor, "Towards Annotation Based Query and Document Expansion," *PostProc. of the CLEF Workshop*, LNCS, Springer-Verlag, 2008.

[61] J. Shi and J. Malik, "Normalized Cuts and Image Segmentation," *IEEE Trans. on PAMI*, Vol. 22-8, 888–905 2000.

[62] I. Guyon, A. Saffari, H. Jair Escalante, G. Bakir, and G. Cawley, "CLOP: a Matlab Learning Object Package," *Demonstration session, Neural Information and Processing Systems Conf. (NIPS)*, Vancouver B.C., Canada, 2007.

[63] A. Torralba, "Contextual Priming for Object Detection," *Intl. Int. J. Comput. Vis.*, Vol 53-2, 169–191, 2003.

[64] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, Vol 3, 1157–1182, 2003.

[65] Lee, S. and Hwang, E.: "Spatial Similarity and Annotation-based Image Retrieval System," Proc. of the IEEE Fourth Intl. Symposium on Multimedia Software Engineering, 33–36, 2002

[66] Wang, Z, Feng, D., Chi, Z. and Xia, T.: "Annotating Image Regions Using Spatial Context," Proc. of the IEEE Eighth Intl. Symposium on Multimedia, 55–61, 2006

[67] Yuan, J., Li, J. and Zhang, B.: "Exploiting Spatial Context Constraints for Automatic Image Region Annotation," Proc. of the 15th Intl. Conf. on Multimedia, 595–604, 2007

[68] Yu, D., and Ip, H.: "Automatic Semantic Annotation of Images using Spatial Hidden Markov Model," Proc. of the IEEE Intl. Conf. on Multimedia and Expo, 305–308, 2006

[69] T. Deselaers, D. Keysers, H. Ney, "Features for Image Retrieval: An Experimental Comparison," *Information Retrieval*, 03/2008, 77–107, Springer, 2008.

[70] Lim, S. and Lu, G.: "Spatial Statistics for Content Based Image Retrieval," Proc. of the Intl. Conf. on Information Technology: Coding and Computing, 155–159, 2003