# Towards a Region-Level Automatic Image Annotation Benchmark*

Hugo Jair Escalante, Manuel Montes and L. Enrique Sucar
Computer Science Department
National Astrophysics, Optics and Electronics Institute
Luis Enrique Erro # 1
Puebla, México,72840
e-mail: {hugojair, mmontesg, esucar}@ccc.inaoep.mx

**Abstract**

Automatic image annotation at region-level consists of the task of assigning labels to regions within segmented images. This is a very important task for the development and improvement of image retrieval methods. Many image annotation methods have been proposed so far, reporting relatively good results on this task. However the lack of benchmark collections have caused that region-level methods are often evaluated using image-level collections. On the other hand, a few region-level collections are available, although these are small and composed of unrealistic images; which difficult the evaluation of the expected annotation performance on real collections. In this paper we propose the creation of a benchmark collection for the evaluation of region-level automatic image annotation methods, in order to provide a valuable resource to the image annotation and retrieval communities. We describe a methodology for creating such a benchmark and present some work we have performed towards building it; work in progress and future directions are also discussed. The main goal of this paper is to obtain feedback from the benchmarking community as well as to establish collaborations and to contact sponsors in order to carry out this exhaustive, but very important, task.

## 1 Introduction

The task of assigning semantic labels (words) to images is known as image annotation. This is a very important step towards developing more effective image retrieval systems. For text-based image retrieval, annotations are indispensable features; while for content-based image

*We would like to thank Michael Grubinger by making available the *IAPR-TC12 Benchmark* [12], we acknowledge the copyrights of this collection due to him. We thank Allan Hanbury because of the facilities and support for the publication of this paper.

retrieval methods, annotations can provide them with semantic information for improving their performance. There are two ways of facing this problem, at image-level and at region-level. In the first case, labels are assigned to the entire image as an unit, not specifying which words are related to which objects within the image. In the second approach, which can be conceived as a high-level object recognition task, the assignment of labels is at region level; providing a one-to-one correspondence between words and regions. The last approach can provide more semantic information for the retrieval task, although it is more challenging than the former. Image annotation, however, is not an easy task; manual annotation is both infeasible for large collections and subjective due to annotator's criteria. In consequence, there is an increasing interest on developing automatic methods for image labeling.

In the last few years many interesting methods have been proposed for automatic image annotation (*AIA*) [7, 16, 23, 2, 8, 1, 4, 22, 6, 5, 19, 17, 21, 18]. Most of these approaches have reported relatively good results on the *AIA* task, however the evaluation criterion used in most of these works make their results not very reliable. Usually, evaluation of *AIA* methods is carried out using small collections of unrealistic images. Furthermore, the performance assessment of most of the region-level methods has been done using collections and evaluation protocols designed for image-level *AIA*, which can not provide a reliable estimation of the methods' accuracy. These sort of evaluations are done because of the lack of a reliable benchmark for the task of *AIA*. Until now, only a few region-level *AIA* collections are available. However collection's size, images' type and even copyrights make these collections not completely accessible and reliable for benchmarking. In this paper we propose the creation of a region-level *AIA* benchmark, which can also be used for evaluating image-level methods. Specifically, we propose the manual annotation at region-level of the *IAPR-TC12 benchmark*, a recently released collection of photographs manually annotated at image-level [12]. We describe some work already carried out for images' segmentation and feature extraction, as well as available tools for manual annotation. Furthermore, we describe work in progress and research issues that should be considered for the development of a reliable benchmark. The main goal of this paper is obtaining feedback from the benchmarking community that can help us for enhancing the reliability of the proposed benchmark; also we are looking for collaborations because this is a very exhaustive task.

The rest of this paper is organized as follows. In the next section we describe the motivation for building a benchmark on *AIA* at region-level. Then in Section 3 we describe the proposed methodology for the full annotation of the *IAPR-TC12 benchmark*. Next in Section 4 research advances are presented. Finally in Section 5 we summarize the proposal and discuss important issues that should be considered when creating this benchmark.

## 2 Motivation

*AIA* is a relatively new research area with the goal of providing visual-semantics into the image retrieval task. Many interesting approaches have been proposed [23, 2, 8, 1, 4, 22, 6, 5, 19, 17, 21, 18], based on statistical and probabilistic models [23, 2, 8, 1, 4, 5, 19], information retrieval models [22, 17, 21, 18, 15] and supervised learning [6, 9]. Most of these methods have been evaluated using different subsets of the Corel$^R$ image collection (annotated at image-level) and

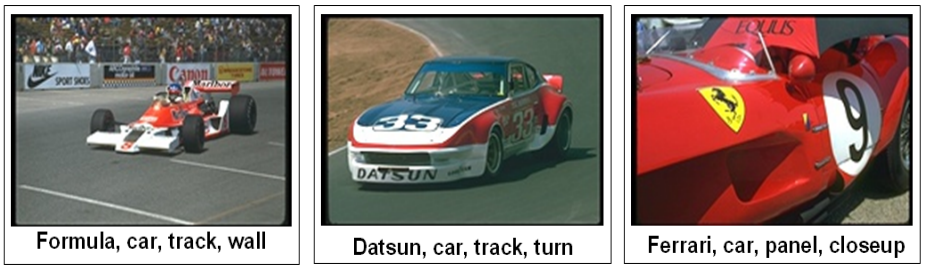


Figure 1: Sample images belonging to the concept *Auto Racing* are shown.

adopted a protocol proposed by Duygulu et al [8]. However, while this protocol may be some reliable for image-level annotation, region-level methods can not be reliably evaluated by using it.

## 2.1 Current image collections

Seminal papers on *AIA* are due to Mori and Duygulu et al, with the introduction of the co-occurrence and machine translation models, respectively [23, 8]. The image collection used by Duygulu et al in their reference work as well as the evaluation protocol proposed became an standard for the evaluation of *AIA* methods [16]. For their experiments, subsets of the Corel image collection were used [8]. The Corel collection 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 1 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 because of its commercial nature that make it not a reliable collection. 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. Third, Corel images are annotated at image level, limiting its applicability to image-level methods. Finally, given that the collection is commercial it is copyright protected and as a result images can not be distributed among researchers. Furthermore, the collection is no longer available hindering the evaluation for some methods.

Recently, a few efforts have been carried out towards developing a benchmark collection for the evaluation of *AIA* methods [14, 3, 5, 13]. Hanbury et al *re-annotated*, at image-level, a large subset of almost 60,000 images as containing animals or not, in most of these images the name of the animal is also provided [14]. For region-level annotation Hanbury et al provided a collection of 1289 manually segmented images of animals in which each segment was annotated. Barnard et al also presents a dataset of 1041 manually segmented images annotated at region-level too [3]. Images correspond to a broader concept domain than that of Hanbury et al; furthermore, Barnard et al considered WordNet, a semantic network, and a established methodology for the annotation process [3, 13]. Each region is therefore annotated according to a set of rules based on concepts and their synonyms as defined in WordNet. Carbonetto et al

provides smaller subsets of Corel images annotated at region-level [5]. Other benchmarks from the object recognition community are useless for evaluating *AIA* algorithms because images in such collections are also unrealistic and are limited to only a few objects[1]. Current segmentation algorithms are far away of providing accurate segmentations, and therefore performance evaluation on manually segmented images is not a reliable estimator of the performance of *AIA* methods on real scenarios. Furthermore the size of the above described collections is very small and all of the images are still unrealistic. In consequence a large collection of automatically segmented realistic images is needed.

## 2.2 Evaluation of AIA methods

The evaluation protocol introduced by Duygulu et al has been used by most of *AIA* methods proposed [7, 16, 23, 8, 1, 4, 22, 19, 17, 21, 18]. It is designed to evaluate annotation performance by assessing image retrieval efficacy using automatically generated annotations. Intuitively, this protocol provides a measure of the labels overlap between the generated image-level annotations and the original image-level labels of the images. The protocol consist of splitting the image collection into training and test sets. The first set is used by the *AIA* methods for learning and training, then images in the test set are annotated using the trained method. Next, queries are defined by considering the set of labels in the test-set vocabulary. Queries consist of combinations of one, two, three and four keywords in such vocabulary. These queries are used for retrieving images, from the test collection, by considering the automatically generated labels. An image is said to be relevant to a query if the retrieved image includes the query in the original (manual) annotation of such image. Standard evaluation measures like precision and recall are used for evaluating the retrieval performance. As we can see annotation accuracy at region level can not be effectively evaluated under this protocol. Because we can never know if the annotation method is accurately assigning annotations to regions or, instead, if the method performed well by chance. This evaluation methodology can be useful for evaluating *AIA* methods at image-level; but even when it can be useful for giving an idea of the performance of region-level *AIA* methods, it can not be considered for reliable evaluations.

A more reliable methodology has been considered for the evaluation of annotation methods by Carbonetto et al [5]. They considered subsets of the Corel collection annotated at region-level. Under Carbonetto's methodology a measure of the percentage of correctly annotated regions is considered. This is a reliable estimate because we can know how many regions were correctly annotated. Instead of just measuring accuracy at image-level. The problem with this approach is that image collections annotated at region-level are required. Peter Carbonetto have made available small subsets of the Corel collection annotated at region-level. However these datasets have the same drawbacks of any Corel subset, namely the unrealistic nature of images.

[1] See for example the **ALOI** collection and the **Caltech** repository.

# 3 Towards and AIA benchmark

Our proposal for creating a benchmark on *AIA* consists of annotating at region-level a large collection of realistic images. Specifically we consider the *IAPR-TC12* benchmark [12] an actual standard for the evaluation of image retrieval methods. We selected this collection because it has already two desirable properties for an *AIA* benchmark, namely its size (large enough for obtaining significant results) and the images source (realistic photographs). The *IAPR-TC12* collection consist of 20,000 images annotated at image-level. Images consist of photographs taken by tourist around several places in the world. Annotations are available in English, German and Spanish. These were generated following a strict methodology and annotation rules [12]. The *IAPR-TC12* benchmark is copyright protected and it is available only for academic and research purposes.

The proposed methodology for creating the *AIA* benchmark consist of the following steps.

1. Segmentation
2. Feature extraction
3. Defining vocabulary
4. Defining annotation rules
5. Manual annotation of images
6. Publication of the benchmark

Since our objective is to provide a benchmark that can be used for evaluating both region-level and image-level *AIA* methods, we should obtain regions for each image in the collection. For this purpose we propose to use automatic segmentation methods because of the size of the collection, and because manual segmented collections can not give a reliable estimate of method's accuracy in real collections. We have considered two segmentation approaches, the first one is based on the normalized cuts algorithm [24], and the second one consists of splitting images into squared patches (grid segmentation). The first method is the most used algorithm in the *AIA* literature [8, 2, 1, 17, 21, 18, 19, 4], while with the second, better results have reported by *AIA* methods [5, 9].

Once the collection have been completely segmented visual attributes should be extracted from each region. For this step we propose extracting a large number of attributes from the regions, including color, texture, shape and orientation information. A large number of features will allow benchmark's users to perform feature selection in order to obtain the best set of features for their annotation methods.

A crucial step towards creating the benchmark consists of defining the annotation vocabulary for the collection, that is the set of labels to choose from for assigning them to regions. We may consider one of the annotation approaches identified by Hanbury: *free text*, *keywords* or *classification based on ontologies* [13]. *Free text* descriptions is one of the most easiest ways of annotation, because the user can annotate regions according to its own knowledge and vocabulary. However under this approach the same object can be labeled with different annotations and other inconsistencies may arose. The *keyword approach* is the most used in *AIA* collections, Corel for example uses it. Under this approach a predefined vocabulary of keywords (or even arbitrary keywords) are used for annotating images. This strategy is a good candidate for defining

Figure 2: Sample images from the *IAPR TC-12* collection. From left to right, original image, image segmented with normalized cuts and image segmented with the grid approach.

the benchmark vocabulary. The last approach consists of using *semantic networks* for assigning labels, this approach is similar to that of using keywords with the difference that annotations are arranged into a hierarchy of concepts, resulting in a semantically annotated collection.

The selection of a keyword vocabulary is ongoing work. Actually we are seriously thinking on using the list of keywords identified by Hanbury in a recent study [13]. This study comprises the analysis of the labels' vocabulary used in several *AIA* collections. Hanbury selected a set of keywords and arranged them according to an ontology [13]. The next step consists of defining a consistent methodology for the annotation process in order that several annotators could participate in the annotation task. This is another important step because of it depends the objective annotation of the collection and the consistency of the benchmark. We intend to establish several annotation rules, like those proposed by Barnard et al [3].

The main reason for creating a benchmark like this is to provide researchers with the segmented collection as a tool for the evaluation of their systems. Therefore the benchmark will be made publicly available for academic and research purposes, of course, under the agreement of the actual copyright owner of the collection. A future work extension could be the proposal of a track within the *ImageCLEF* forum for the evaluation of region-level image annotation systems.

# 4 Research status

At the moment we have segmented the *IAPR-TC12* collection into regions using both the normalized cuts algorithm and the grid approach. A total of 100,000 regions resulted from the segmentation with normalized cuts and around 480,000 regions are available under the grid segmentation approach (considering 24 patches per image, of course the size of the patches in the grid approach can be modified for obtaining smaller or bigger regions). Segmentation is not computational expensive, it may took almost 2 days for segmenting the entire collection. Sample segmented images from *IAPR-TC12* collection are shown in Figure 2. As we can see good image partitions can be obtained with the normalized cuts method, while other segmentations can be poor with this approach. The grid method always shows the same performance.

Simultaneously, when images were segmented visual attributes were also extracted from each of the regions. The set of attributes considered consisted of color, texture, shape and

orientation information. Therefore, each of the regions is described by a vector of attributes. The vectors' size varies because we used different tools for the distinct segmentation algorithms. Though in the near future we will extract the same patterns from both segmentations.

For the process of segmentation and feature extraction we have used two recently developed software tools. The first one is an interface developed at our institution that uses the normalized cuts algorithm for segmenting a given image collection [20]. This tool also includes methods for feature extraction and manual annotation of regions. Additionally the interface provides options for the re-segmentation of images and for joining adjacent regions annotated with the same label. A second interface for segmentation and manual annotation is that provided by Perter Carbonetto, this tool also includes options for segmentation with both normalized cuts and the grid approach. Further, this tool provides methods for soft-annotation of images, that is several labels can be assigned to each region.

Regarding the manual annotation step we have performed a first attempt for the annotation of the *IAPR-TC12* collection, however some issues were encountered. Four our participation in the photographic retrieval task at *ImageCLEF2007* we decided to make use of *AIA* methods for improving accuracy of retrieval [10]. For this purpose we created a training set of manually annotated regions. Randomly, a small subset of the collection segmented with normalized cuts was selected and manually annotated according to a defined vocabulary a keywords. For defining this vocabulary we looked at the textual part of the task's topics [11]. The keyword-annotations were defined according to a handmade ontology, built according to the nouns appearing in the topics. In Table 1 the label's vocabulary proposed for annotating images is shown, as well as the keywords related to each of the labels. As we can see some keywords like *building* and *person* comprise many concepts, though several other are not associated to any other keyword. This fact together with poor segmentation made difficult the process of annotation. Using this pseudo-ontology a small subset of images was annotated and used for training and *AIA* method in order to annotate the rest of the segmented collection. The automatically generated labels were used for expanding queries and documents in the ad hoc retrieval task at *ImageCLEF2007* [10].

Our experience at *ImageCLEF2007* give evidence that poor segmentation as well as the wrong definition of the annotation vocabulary can difficult the annotation process. The segmentation problem can be alleviated by using the grid segmentation approach. Because, even when segmentation is poor, it is consistent through any type of images. While segmentation algorithms like normalized cuts have irregular performance depending on the images. The vocabulary definition problem can de addressed by defining a consistent vocabulary of keywords, based on semantic knowledge (just as the keyword list proposed by Hanbury [13]). On the other hand, results obtained in the retrieval task give evidence that the use of annotation for image retrieval can help improving retrieval performance, although some issues should be addressed. Work in progress consists on the definition of the annotation vocabulary and the methodology to follow for the annotation process. We are requesting feedback and collaborations for both tasks[2].

[2]We will appreciate any comment, suggestion and help offer, if interested please contact the first author of this paper via e-mail.

| **Keyword** | **Associated keywords** |
|---|---|
| animal | fish, bird, reptile, kangaroo, . . ., seals, sea lions (any animal) |
| boat | - |
| building | accommodation, tourist-accommodation, hotel, hostel, cities, stadium, school, bridge, grandstand, ruin, wall |
| church | mosque, cathedral |
| clouds | fog |
| flag | - |
| furniture | bed, tv, room |
| grass | football, ground, sports-field |
| mountain | landscape, volcanoes, sights |
| other | - |
| person | group of persons, people, footballers, players, families, god daughter, tennis player, god children, god son, guide, woman, girls |
| plate | meat dish, dish |
| prize | medals, trophies, cups |
| road | straight road, highway, square, street |
| sand | salt pan, beach, desert, salt heap, salt pile |
| sky | - |
| snow | winter |
| statue | monument |
| sun | sunset |
| swimming-pool | - |
| tower | telescope, lighthouse |
| trees | - |
| vehicle | motorcycle, car, buses, bicycles, forklifts, trains |
| water | sea, lake, waterfall, river |

Table 1: Annotation vocabulary defined for the creation of a training set of annotated regions for *INAOE-TIA's* participation at ImageCLEF2007 [10].

# 5 Conclusions

We have proposed the creation of a benchmark for the evaluation of region-level image annotation methods. Our proposal comprises the segmentation, feature extraction and annotation at region-level of images in the *IAPR-TC12* collection. The main goal of this paper is to obtain feedback from the benchmarking community in order to create a robust standard collection. We have already segmented the collection using normalized cuts and a grid approach, and visual features have been extracted from each of the resulting regions. Furthermore, software tools available for manual annotation have been briefly described. Currently we are in the process of defining the vocabulary of keywords allowed for annotation. The next stage of the project will consist of annotating the generated regions, a slow and time-consuming process that should be done some day and start as soon as possible.

# References


[1] K. Barnard, P. Duygulu, N. de Freitas, D. Forsyth, D. Blei, and M. I. Jordan. Matching words and pictures. *Journal of Machine Learning Research*, 3:1107–1135, 2003. 2, 4, 5

[2] K. Barnard and D. Forsyth. Learning the semantics of words and pictures. In *International Conference on Computer Vision*, volume 2, pages 408–415. IEEE, IEEE, 2001. 2, 5

[3] Kobus Barnard, Quanfu Fan, Ranjini Swaminathan, Anthony Hoogs, Roderic Collins, Pascale Rondot, and John Kaufhold. Evaluation of localized semantics: Data, methodology and experiments. *International Journal of Computer Vision (to appear)*, 2007. 3, 6

[4] D. Blei and M. Jordan. Modeling annotated data. In *Proceedings of the 26th annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 127 – 134. ACM press, 2003. 2, 4, 5

[5] P. Carbonetto, N. de Freitas, and K. Barnard. A statistical model for general context object recognition. In *Proceedings of the 8th European Conference on Computer Vision*, pages 350–362, 2005. 2, 3, 4, 5

[6] G. Carneiro, A. B. Chan, P. J. Moreno, and N. Vasconcelos. Supervised learning of semantic classes for image annotation and retrieval. *IEEE Trans. on PAMI*, 29(3):394–410, 2007. 2

[7] R. Datta, J. Li, and J. Z. Wang. Content-based image retrieval - approaches and trends of the new age. In *Proceedings ACM International Workshop on Multimedia Information Retrieval*, Singapore, 2005. ACM Multimedia. 2, 4

[8] P. Duygulu, K. Barnard, N. de Freitas, and D. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *Seventh European Conference on Computer Vision*, volume IV of *LNCS*, pages 97–112. Springer, 2002. 2, 3, 4, 5

[9] H. J. Escalante, M. Montes, and L. E. Sucar. Word co-occurrence and mrf's for improving automatic image annotation. In *In Proceedings of the 18th British Machine Vision Conference (BMVC 2007) To appear*, Warwick, UK, September, 2007. 2, 5

[10] H. Jair Escalante, C. A. Hernandez, A. Lopez H. Marin-Castro, E. Morales, L. E. Sucar, M. Montes, and L. Villasenor. Inaoe-tia participation at imageclef2007. In *Working Notes of the CLEF (to appear)*, Budapest, Hungary, 2007. CLEF. 7, 8

[11] M. Grubinger, P. Clough, A. Hanbury, and H. Müller. Overview of the ImageCLEF 2007 photographic retrieval task. In *Working Notes of the 2007 CLEF Workshop*, Budapest, Hungary, September 2007. 7

[12] M. Grubinger, P. Clough, H. Müller, and T. Deselaers. The iapr tc-12 benchmark: A new evaluation resource for visual information systems. 2005. 1, 2, 5

[13] Allan Hanbury. Review of image annotation for the evaluation of computer vision algorithms. Technical Report 102, PRIP, Vienna University of Technology, 2006. 3, 5, 6, 7

[14] Allan Hanbury and Alireza Tavakoli Targhi. A dataset of annotated animals. In *Proceedings of the Second MUSCLE / ImageCLEF Workshop on Image and Video Retrieval Evaluation*, Czech Repulic, 2006. 3

[15] J. S. Hare, P. H. Lewis, P. G.B. Enser, and C. J. Sandom. A linear-algebraic technique with an application in semantic image retrieval. In H. Sundaram, editor, *ACM International Conference on Image and Video Retrieval, CIVR*, volume 4071 of *LNCS*, pages 31–40. ACM, Springer-Verlag, 2006. 2

[16] J. S. Hare, P. H. Lewis, P. G.B. Enser, and C. J. Sandom. Mind the gap: Another look at the problem of the semantic gap in image retrieval. In Hanjalic A. Chang, E. Y. and Eds. Sebe, N., editors, *Proceedings of Multimedia Content Analysis, Management and Retrieval*, volume 6073, San Jose, California, USA, 2006. SPIE. 2, 3, 4

[17] J. Jeon, V. Lavrenko, and R. Manmatha. Automatic image annotation and retrieval using cross-media relevance models. In *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 119–126, New York, NY, USA, 2003. ACM Press. 2, 4, 5

[18] J. Jiwoon and R. Manmatha. Using maximum entropy for automatic image annotation. In P. Enser, editor, *Procc. international conference on image and video retrieval (CIVR 2004)*, volume 3115 of *LNCS*, pages 24–32, Dublin IR., 2004. Springer. 2, 4, 5

[19] V. Lavrenko, R. Manmatha, and J.Jeon. A model for learning the semantics of pictures. In Sebastian Thrun, Lawrence Saul, and Bernhard Scholkopf, editors, *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA, 2004. 2, 4, 5

[20] H. Marin-Castro, L. E. Sucar, and E. F. Morales. Automatic image annotation using a semi-supervised ensemble of classifiers. In *In Proceedings of the 12th Iberoamerican Congress on Pattern Recognition (CIARP 2007), To appear*, 2007. 7

[21] D. Metzler and R. Manmatha. An inference network approach to image retrieval. In *Proceedings of International Conference Image and Video Retrieval*, volume 3115 of *LNCS*, pages 42–50. Springer, 2004. 2, 4, 5

[22] F. Monay and D. Gatica-Perez. Modeling semantic aspects for cross-media image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, to appear. 2, 4

[23] Y. Mori, H. Takahashi, and R. Oka. Image-to-word transformation based on dividing and vector quantizing images with words. In *First International Workshop on Multimedia Intelligent Storage and Retrieval Management (in conjunction with ACM Multimedia Conference 1999)*, 1999. 2, 3, 4

[24] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000. 5