Semantic Cohesion for Image Annotation and Retrieval

Hugo Jair Escalante Balderas,
Manuel Montes, Enrique Sucar

Reporte Técnico No. CCC-10-004
22 de abril de 2010

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

Abstract of PhD Thesis
Semantic Cohesion for Image Annotation and Retrieval

Graduated: Hugo Jair Escalante Balderas

Supervisors: Manuel Montes and Enrique Sucar

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

Abstract

We present methods for image annotation and retrieval that are based on the semantic cohesion among terms. On the one hand, we propose a region labeling technique that assigns an image the labeling that maximizes an estimate of the semantic cohesion among candidate labels associated to regions in segmented images. On the other hand, we propose document representation techniques that are based on the semantic cohesion among multimodal terms that compose to images. Additionally, we extended a benchmark collection of the evaluation of the proposed techniques.

1 Introduction

Nowadays images are the main source of information available after text; this fact is due to the availability of inexpensive image registration (e.g., photographic cameras and cell phones) and data storage devices (large volume hard drives), which have give rise to the existence of millions of digital images stored in many databases around the world. However, stored information may be useless if we cannot access the specific data we are interested on. Thus, the development of effective methods for the organization and exploration of image collections is a crucial task.

Image retrieval has been an active research area since more than two decades ago [48, 27, 8, 36, 35, 43]. However, despite that substantial advances have been achieved so far, most of the reported work focuses on methods that consider a single modality (i.e., either image or text), limiting the effectiveness and applicability of such methods. On the one hand, text-based methods are unable to retrieve images that are visually similar to a query image. On the other hand, image-based techniques cannot retrieve relevant images to queries that involve non-visual information (e.g., about places, events or dates). Further, visual methods present additional complications; for example, the need of specifying query images, providing relevance feedback and, more importantly, the ambiguity on determining the underlying user information need from a sample image.

Because of the above limitations, in the last few years there has been an increasing interest from the scientific community in the study and development of retrieval techniques that incorporate both visual and textual information [36, 6, 7, 2]. Most of researchers that adopt the latter approach attempt to exploit the complementariness and diversity of information from different modalities available in multimodal images (i.e., images that are composed by terms of at least two modalities). Despite that such approach seems logical and intuitive, it is not easy to develop methods that can yield satisfactory retrieval results. Hence, current techniques fail at exploiting the availability of multimodal information for effectively representing the content of images. Furthermore, in many databases, images are not accompanied with textual information, which further complicates the application and development of multimodal retrieval methods.

The image retrieval problem is therefore more complex in collections where images are not annotated as in a preliminary step images must be associated with keywords. Since manually annotating images is a time consuming and labor expensive task, automatic image annotation (AIA) methods are considered instead [30, 8]. Thus, AIA methods are very important for allowing the application of multimodal image retrieval techniques in un-annotated image collections. Despite current AIA methods are still limited in several aspects, recent results have gave evidence that the use of labels, as generated by such methods, is helpful for improving the retrieval performance in both annotated and un-annotated image collections [12, 13].

Both tasks, image annotation and image retrieval, are closely interrelated and hence they can be studied jointly. Accordingly, in this thesis we face the problems of image annotation and retrieval with the goal of improving the performance of current techniques and overcoming some of their limitations. More specifically, we focused on the region-level AIA task with the goal of giving support to multimodal image retrieval methods that attempt to exploit the redundancy and complementariness of information as provided by labels and text. Implicitly, our study includes the evaluation on the benefits of using labels for image retrieval in realistic scenarios. The rest of this document summarizes our research and outlines the main findings of our work.

2 Motivation

Despite AIA methods have been studied for a while there are several aspects that have received little attention so far. For example, most AIA methods have been applied and evaluated mainly in image collections that have been considered "easy" by the specialized community [40, 28]. Also, labels, as generated by AIA methods, have been used for image retrieval in very restricted retrieval settings: for example, searching for images by using the labels assigned to images [32, 1, 4, 33, 5, 25, 3]. Hence, the actual usefulness of AIA methods for image retrieval has not been properly evaluated so far.

Besides, the use of AIA methods has been restricted to databases where the images have not any associated text (i.e., un-annotated collections). Even when this was the main purpose for developing AIA methods in the first place [39, 10, 3], a hypothesis of this work is that labels can also be helpful for annotated image collections [12, 13]. Such hypothesis is based on the fact that text and AIA labels can provide complimentary information (see Figure 1). On the one hand, textual descriptions assigned by users describe the image content at a very high level of semantics; for example, making reference to places, events and dates (see a and c in Figure 1). On the other hand, the labels as used in AIA describe the content of images at a low level of semantics, making reference to visual objects present in the image, for example: clouds, sand, trees or sky (see b and d in Figure 1). Therefore, it is clear that both modalities provide information that can be complimentary[1] and redundant[2] at the same time. Thus, it is reasonable to assume that we can exploit this multimodal information to develop effective multimodal image retrieval methods.

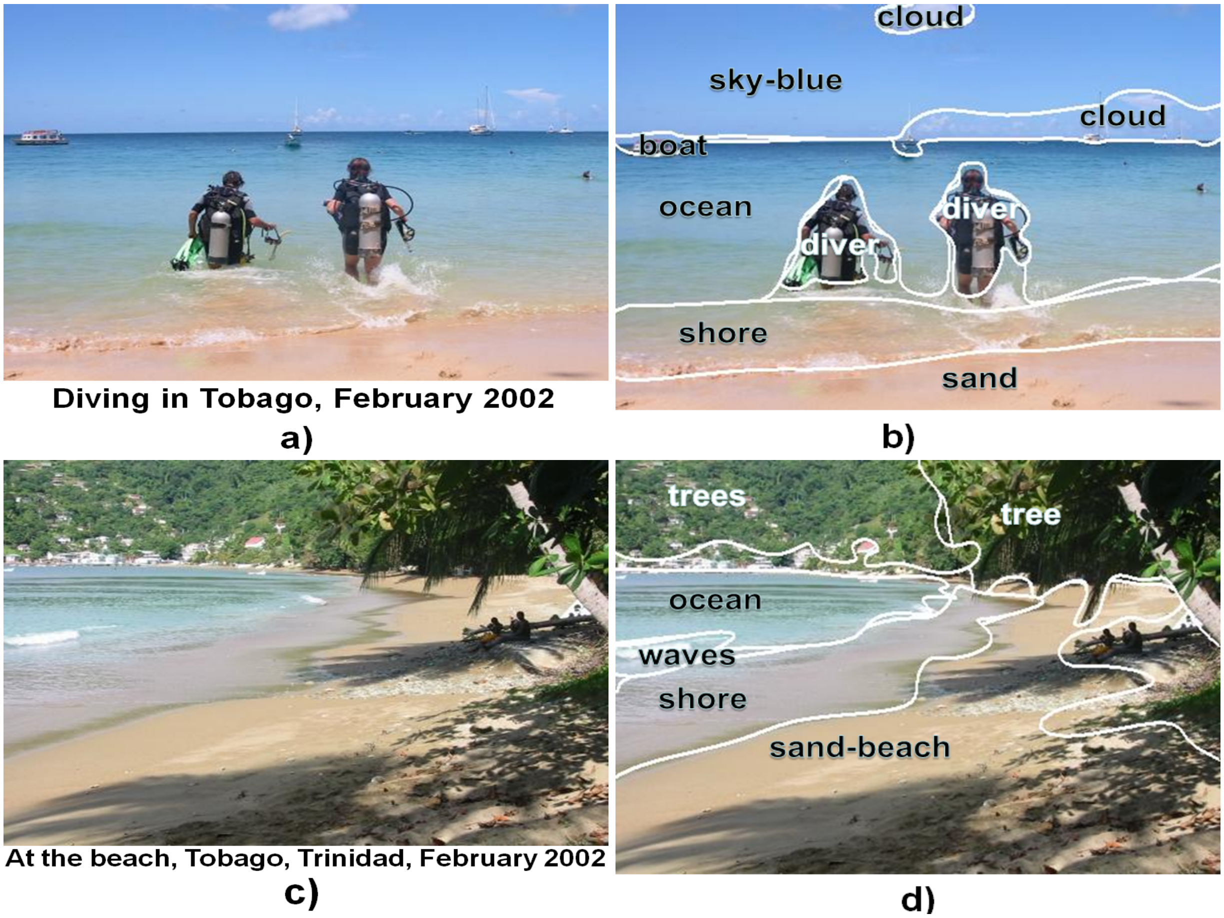

Figure 1. Illustration of the complementariness between text (a y c) and labels from AIA (b y d). Images taken from the SAIAPR TC12 collection [15].

One should note that the above complementariness and redundancy of information can be exploited in collections where the images have been manually annotated by users. Whereas there exist many annotated image collections (e.g., medical image collections[3], images available in the Web[4], image collections from newspapers[5] and magazines[6] as well as some personal image collections[7]) most of the existing images have

[1] For example, in Figure 1 the labels in b complement the annotation in a, combining both sources of information we can know that the image contains two divers at the shore of a beach in Tobago, where the sky is blue, there is a boat in the background and the picture was taken in 2002.

[2] For example, in Figure 1 the labels in d are redundant to the annotation in c, the combined information refers to the shore of a beach at Tobago with vegetation in the background.

[3] http://www.irma-project.org/

[4] http://images.google.com/

[5] http://www.belga.be/

[6] http://www.nationalgeographic.com/photography/

[7] http://www.flickr.com/

not been manually annotated; for those image collections AIA labels offer important benefits as the main goal of developing AIA methods is to support users of un-annotated image collections [39, 10, 8, 32, 4, 33, 3].

Labels are helpful for un-annotated image collections because without them the retrieval task in those collections requires of a considerable amount of user participation; for example, for providing sample images to be compared with those stored in the database [48, 43] (in content-based image retrieval systems) or for browsing through predefined categories and for providing relevance feedback [44, 36] (in interactive retrieval systems). Despite that the interaction with the user may improve the retrieval performance, it should be avoided whenever possible as it makes the retrieval task anti-natural (providing an image similar to those we want to retrieve) or tedious (for some users it may be bothering to browse through the collection or to provide relevance feedback in order to find the required images). In consequence, AIA methods are critical to simplify and to improve the accessibility to images in un-annotated image collections: by providing images with labels, the users may formulate queries by using keywords avoiding the need to provide sample images; also, with the availability of labels, interactive relevance systems can be improved and simplified as well. Further, information fusion techniques can be used to combine visual information and labels with the goal of improving the retrieval performance [14].

3 On the semantic cohesion

We propose solutions to the image annotation and image retrieval problems based on a modeling process that accounts for the semantic cohesion among terms[8] [21, 22]. In particular, we face the region-level AIA problem and focused on multimodal image retrieval from documents that are composed of both text (assigned by users) and labels (generated by AIA methods). Before describing our proposed solutions we define what we will understand by semantic cohesion throughout this document[9]:

- **Semantic cohesion:** *Semantic cohesion is the degree of relationship between terms within a document according to their meaning in a certain context.*

The semantic cohesion reflects the degree of affinity of the terms in a document according to their meaning or their use in the context given by other terms that occur in the same document. Intuitively, the more the semantic cohesion among the terms the higher the probability that such terms are used together in similar contexts. For example, in the case that the terms are words (i.e., textual modality), the terms “snow” and “polar bear” have more semantic cohesion than the terms “snow” and “lion”; in the case that we have terms from two different modalities (e.g., text and labels), the words “accommodation”, “hostel” and the labels “swimming pool”,“hotel” have more semantic cohesion than the words “accommodation”, “hostel” and the labels “church”, “swimming pool”.

In this work we adopted a formulation based on term occurrence counts such that we estimate the semantic cohesion among terms through occurrence and co-occurrence statistics. Such statistics provide useful information about terms usage that can approximate the actual association between terms. This formulation also offers practical advantages, as the calculus of the considered statistics is a rather simple and efficient task. There are several other options to estimate the semantic cohesion (e.g., by using lexical resources like

[8] By term we refer to the building blocks by which documents are constituted; for example, terms in textual documents can be words, n-grams or phrases.

[9] We would like to emphasize that it is not our intention a formal study on the semantics of terms or documents, nor on the extraction or on the use of semantic knowledge from a strict point of view, thus by semantics we will refer to any information about the meaning or the use of terms.

WordNet [38] or other sources of knowledge), although, adopting other formulations may be more complicated or more computationally expensive; nevertheless, we will explore other sources of information for estimating the semantic cohesion in future work.

4 Goals

The general goal of this research is as follows:

"The development of image annotation and image retrieval methods that can exploit the semantic cohesion among multimodal terms for improving the effectiveness of current techniques".

Accordingly, we have the following specific goals:

- To develop effective AIA methods that can take advantage of the semantic cohesion among labels for improving the labeling performance of current techniques and that can give support to the task of multimodal image retrieval.
- To develop image retrieval methods that can exploit the semantic cohesion among labels and text to improve the performance of unimodal and standard multimodal techniques.
- To evaluate the effectiveness of AIA methods and the impact of the use of AIA labels into the multimodal image retrieval task.

5 Research overview and main results

In this work we propose methods for the annotation and retrieval of images that are based on the idea of the semantic cohesion among multimodal terms. The rest of this section summarizes our work, for further details we refer the reader to the thesis document [11] or to the following derived research papers [15, 18, 22, 21]. Before presenting the developed methods, we describe the extension we proposed to a benchmark collection for allowing the evaluation of region-level AIA methods and of image retrieval methods that consider AIA labels.

5.1 The SAIAPR TC12 collection

Because of the lack of a suitable database to evaluate the methods we propose, part of our work included the development of a benchmark image collection. Specifically, we proposed the extension of the IAPR TC12 collection[10], an already benchmark data set for the evaluation of image retrieval methods [28, 29]. The extension consisted on the manual segmentation and annotation of each image in the IAPR TC12 collection, according to predefined rules and by using a hierarchical organization of the vocabulary that we defined. The proposed hierarchy is composed of six branches: "Animal", "Humans", "Food", "Man-made", "Landscape" and "other". Figure 2 shows the "Landscape" branch and Table 1 shows statistics of the used labels per branch.

Summarizing, a total of $20,000$ images have been manually segmented and the resultant $99,535$ regions were manually labeled by using a vocabulary of 255 labels. The most used labels were (with quantities

[10] http://imageclef.org/photodata

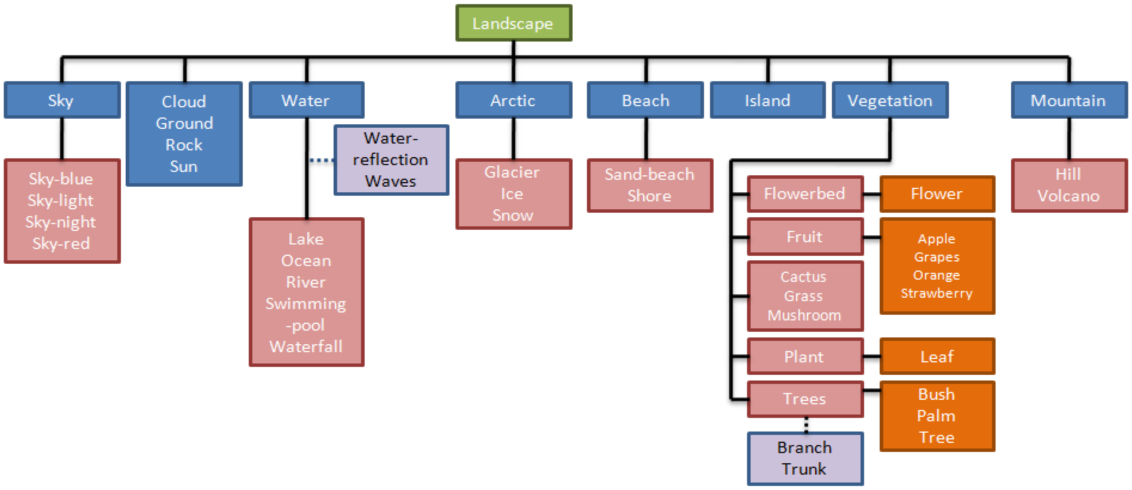

Figure 2. The *"Landscape"* branch from our hierarchical organization of the vocabulary of the SAIAPR TC12 collection.

between parentheses): *"sky-blue"* (5,722), *"man"*, (4,330), *"group-persons"* (4,232), *"ground"* (3,785), *"grass"* (3,211), *"cloud"* (3,128), *"rock"* (3,071), *"vegetation"* (2,763), *"trees"* (2,638), and *"sky"* (2,637). Sample images from the SAIAPR TC12 collection are shown in Figure 1.

Branch	Animal	Humans	Food	Man-Made	Landscape	Other
Frequency	1, 991	16, 950	705	34, 024	45, 308	622
Descendants	70	14	6	110	45	6
Leafs	56	12	5	88	33	6

Table 1. Statistics of the branches in the hierarchy of concepts. We show the number of labeled regions below each branch (row 2), the number of descendants per branch (row 3) and the number of leafs below each branch (row 4).

All of the data derived from our extension is publicly available from the official ImageCLEF website[11]; we also have a mirror website from our institution[12]. The extension we made to the IAPR TC12 has increased its number of applications and its scope in terms of the tasks that can be evaluated with it [15]; furthermore, the extension has been extremely helpful for the evaluation of the methods we developed and has attracted the interest from the scientific community [21, 17]. For a detailed description of our extension to the IAPR TC12 collection we refer the reader to the following reference [15].

5.2 Semantic cohesion for automatic image annotation

For AIA we propose an energy-based model that attempts to maximize the semantic cohesion among labels that have been assigned to adjacent regions in segmented images [21]. The model seeks to refine

[11] http://imageclef.org/SIAPRdata

[12] http://ccc.inaoep.mx/~tia/saiapr/

the initial labeling as provided by a multiclass classifier that is trained with purely visual information. The classifier (which can be built by using diverse learning algorithms) provides candidate labels for every region in an image; next, using information about the association between labels, the energy-based models selects the best combination of labels that should be assigned to the image.

Figure 3 depicts our approach to AIA. A multiclass classifier is used to obtain candidate labels for every region in the segmented image (initial labeling). For assigning a single label to each region the model selects the configuration of labels that maximizes the semantic cohesion among labels (semantic cohesion modeling); dotted squares in Figure 3 indicate the parts of the process where we have contributed.

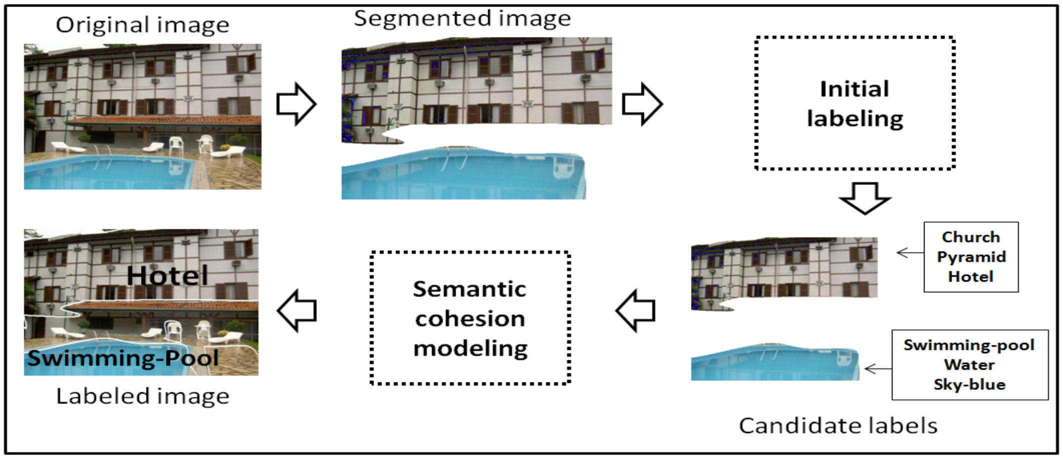

Figure 3. Our approach to AIA. An AIA method is used to assign candidate labels to each region in a segmented image; where each region can be assigned a single label. The combination of labels that maximizes the semantic cohesion is used to annotate the regions.

Figure 4 depicts the proposed energy-based model for AIA. The proposed model resembles a Markov random field with a predefined energy function (i.e., no learning phase must be performed) that incorporates the relevance weights as obtained from the multiclass classifier together with co-occurrence statistics. Inference in the model is performed via iterated conditioned modes (ICM). The code of the developed energy model is publicly available from the following link: `http://ccc.inaoep.mx/~hugojair/ebm`.

We report experimental results obtained with the proposed method over several benchmark image collections of heterogeneous characteristics. Table 2 shows the annotation accuracy, across the considered data sets, obtained by the initial classifier (column 3) and after applying our energy model (column 4); also, we show the best reported results for the corresponding data sets (column 2). For the initial classification we considered a random forest classifier (RF), since with this method we obtained the best labeling results in preliminary experiments, although several other classification techniques were evaluated. Our experimental results show the usefulness of the proposed method. First, the multiclass classification approach to AIA proved to be very effective. Second, the energy-based model improved the initial labeling for all of the considered collections (the difference was statistical significant according a Wilcoxon signed-ranks test with 95% of confidence). Third, the proposed method outperformed the best results reported in related works where the authors have used the same collections we did. Furthermore, we provide evidence that shows how the labels as generated with the energy-based model can be used to search for images by using single labels or by combining the labels with text by means of information fusion techniques. Summarizing, the main

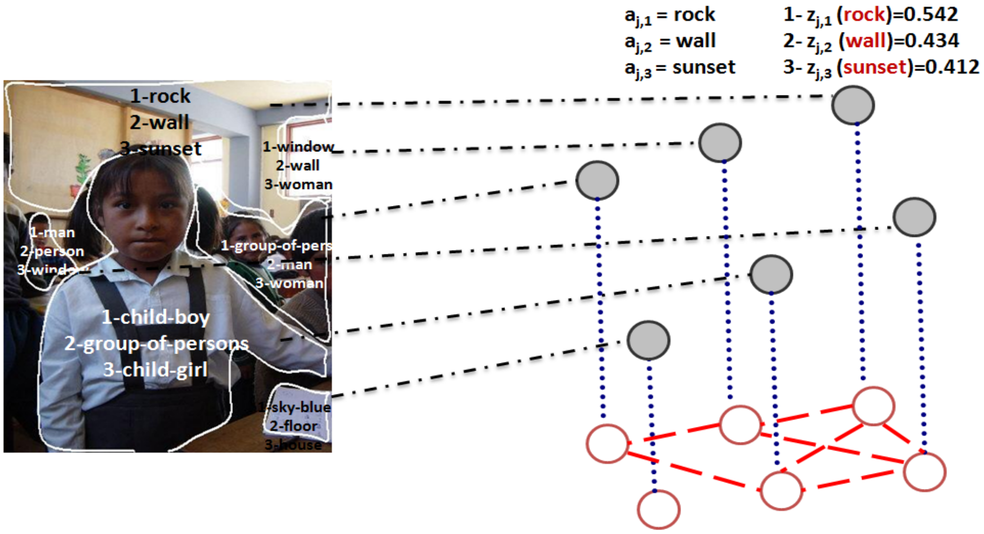

Figure 4. Illustration of the proposed energy-based model for AIA. Left: segmented image with candidate labels per region; a relevance weight is associated with each candidate label. Right: graph associated to the image according to the energy-based model. Unshaded nodes represent assignations of labels to regions, shaded nodes denote the confidence of classifiers in the candidate labels. We consider dependencies between spatially connected regions.

Data set	Reference	OVA - RF	OVA-RF + EBM
COREL-AN	45.64% [31]	57.90% (26.86%)	**58.97% (29.21%)**
COREL-AG	50.50% [24]	56.56% (12.00%)	**57.23% (13.33)**%
COREL-BN	39.50% [24]	46.65% (18.10%)	**48.74% (23.39%)**
COREL-BG	43.00% [24]	46.08% (7.16%)	**46.87%(9.00%)**
COREL-CN	42.50% [24]	54.13% (27.36%)	**55.59% (30.1%)**
COREL-CG	47.50% [24]	52.28% (10.06%)	**52.71% (10.97%)**
SCEF-I	**60.94%** [42]	59.99% (-1.55%)	60.35% (-0.96%)
SCEF-II	78.73% [42]	81.55% (3.58%)	**82.92% (5.32%)**
MSRC-I	**93.94%** [50]	86.60% (-7.81%)	88.82% (-5.45%)
MSRC-2	70.50% [47]	70.60% (0.14%)	**76.03% (7.84%)**
VOGEL	71.70% [49]	70.78% (-1.28%)	**72.54% (1.17%)**

Table 2. Comparison of the labeling accuracy obtained by the initial classifier (column 3), the proposed energy-based model (column 4) and the best reported result for the corresponding image collection (column 2). For OVA-RF and OVA-RF+EBM we show in parenthesis the relative improvement of our methods over the corresponding references.

benefits of the proposed method are the generality of the approach, its easiness of implementation, its effectiveness and its high efficiency. Our work on image annotation with the energy-based model is described in detail in the following reference [21].

5.3 Semantic cohesion for multimodal image retrieval

For image retrieval we propose methods based on the semantic cohesion among labels and text to represent multimodal documents. Specifically, we propose two forms of representing images based on distributional term representations (DTRs) that have been widely used in computational linguistics [34]. Under the considered DTRs each term is represented by a vector of statistics of occurrence over the documents in the collection or co-occurrences over terms in the vocabulary. In this way, the representation of a term will be influenced by the terms it mostly co-occurs with (capturing dependencies between terms) or by the documents in which it occurs (capturing dependencies between terms and documents). For example, Figure 5 shows the multimodal term co-occurrence representation of a selected term.

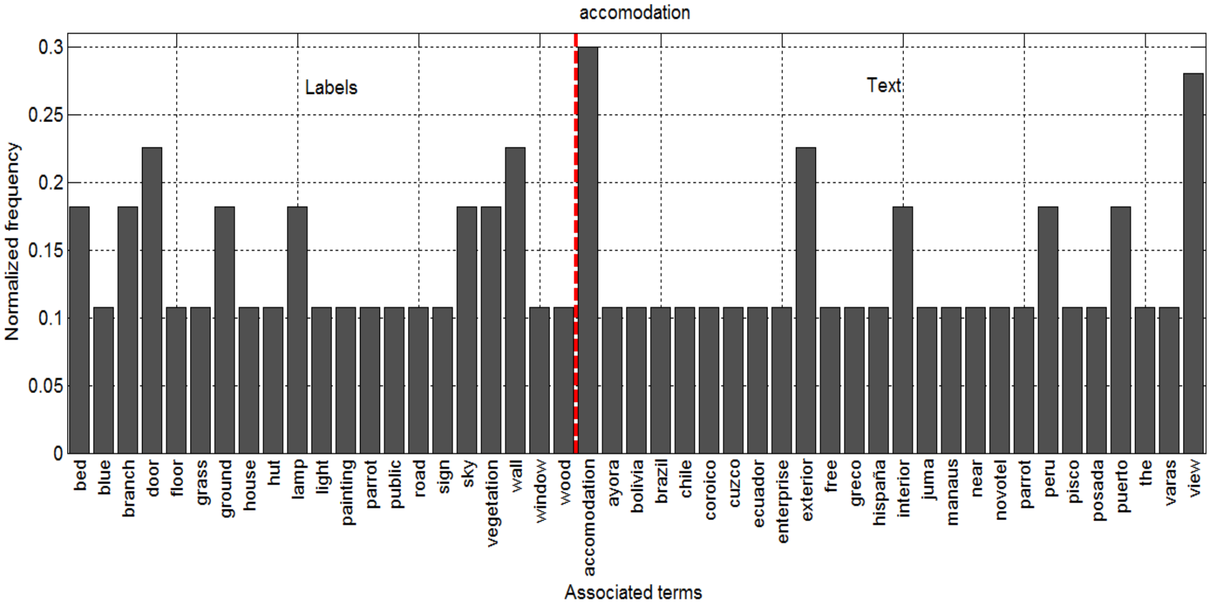

Figure 5. Representation of the term "accommodation" according the multimodal term co-occurrence representation, see [22]. The term is represented by the frequency of co-occurrence of this term with other terms in the multimodal vocabulary (we only show those terms that co-occur with "accommodation" at least one time). The vertical line separates labels form textual terms.

Once each term in the multimodal vocabulary is represented through DTRs, documents are represented by adding the DTRs of terms that appear in the document. Intuitively, each document is represented by the context associated to the terms that occur in the document. Figure 6 shows a sample document from the SAIAPR TC12 collection and Figure 7 shows its multimodal term co-occurrence representation. The representation of a document can be considered an expansion of the terms that are contained in the document. The expansion will be influenced by either: a) the terms that mostly co-occur with the terms that occur in the document (under the term-co-occurrence-representation), capturing second order relations between terms; or b) the documents in which mostly occur the terms in the document (under the document-occurrence-representation), capturing the association among terms through the representations of the terms. Additionally, we developed several standard techniques for combining information from labels and text.

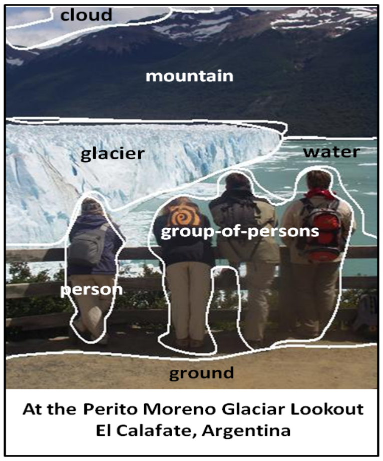

Figure 6. A sample image from the SAIAPR TC12 collection.

We report experimental results with the developed techniques on the SAIAPR TC12 collection. Table 3 compares the retrieval performance of our proposals (multimodal document occurrence representation, MDOR, and multimodal term co-occurrence representation, MTCOR) with unimodal (text-only and labels-only) and standard multimodal techniques (late fusion, early fusion and inter-media relevance feedback), using the SAIAPR TC12 collections over two sets of topics (ImageCLEF2007 and ImageCLEF2008). Experimental results obtained with the standard methods show that the combination of labels and text can be helpful for improving significantly the performance of unimodal strategies. However, the proposed representations achieve better performance than the standard techniques. The difference in performance is statistically significant for MDOR according to a pairwise t-student test with 95% of confidence. Furthermore, the content of multimodal images is better represented with our techniques, when compared to unimodal or standard multimodal strategies. In summary, we provide evidence showing that the combination of labels and text can be very helpful for image retrieval and we prove that the proposed representations provide an effective solution to the multimodal image retrieval task. Our developments on multimodal image retrieval with distributional term representations are explained in detail in [22].

6 Contributions

This section elaborates on the contributions derived from our work, which are as follows:

- We have developed a new method for region-level automatic image annotation, based on the idea of the semantic cohesion, which is easy to implement, highly efficient, generic and very effective; such method has been evaluated in several image collections obtaining superior performance than that reported in related works. Also, according to our knowledge, the labeling refinement method we proposed was the first of its kind [20, 19, 21]; currently, several researchers are adopting similar formulations [37, 41, 31, 45, 46, 42, 26].

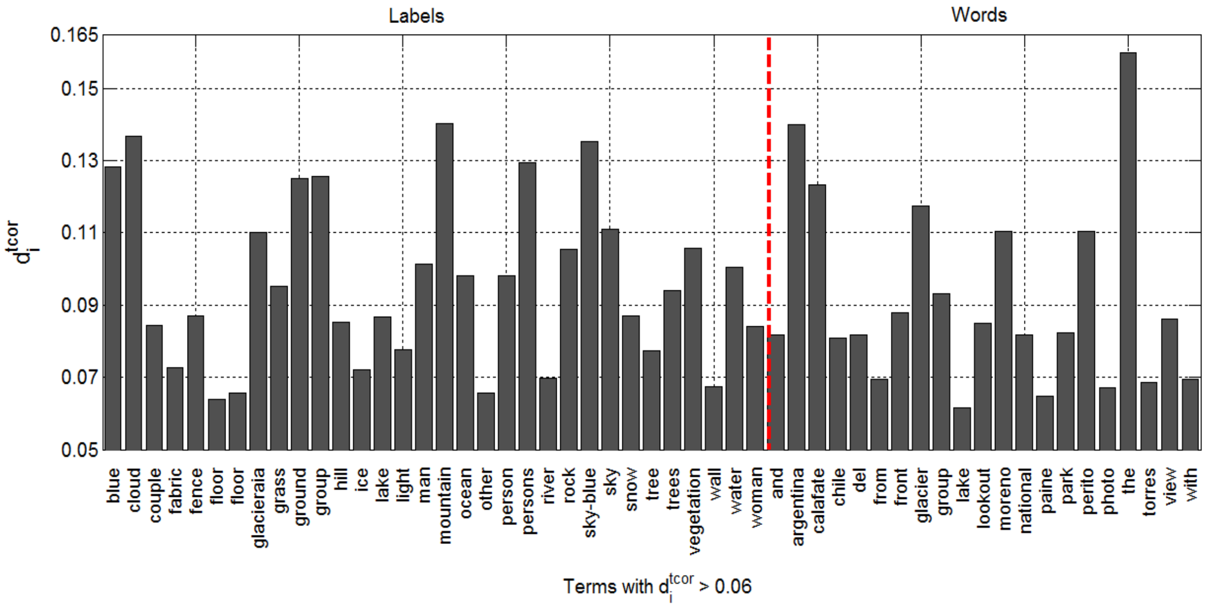

Figure 7. Representation based on term co-occurrence statistics for the document shown in Figure 6.

Topics	ImageCLEF2007				ImageCLEF2008			
Method	**MAP**	**P20**	**R20**	**RR**	**MAP**	**P20**	**R20**	**RR**
MDOR	**0.2141**	0.2425	**0.3295**	**2507**	**0.1958**	**0.25**	**0.3333**	**1679**
MTCOR	0.1935	0.2392	0.2951	2379	0.1763	0.2564	0.2893	1632
Late-fusion	0.1348	0.1858	0.1879	1703	0.1126	0.1936	0.1759	1139
Early-fusion	0.189	**0.2508**	0.2996	2226	0.1565	0.2372	0.2695	1416
IM-relevance feedback	0.1659	0.2142	0.262	1952	0.1326	0.1987	0.2205	1302
Labels-only	0.0587	0.1417	0.1066	1201	0.053	0.141	0.1133	727
Text-only	0.1241	0.1767	0.1694	1424	0.1033	0.1795	0.1534	1014

Table 3. Retrieval results by using the proposed representations (rows 3-4), standard techniques (rows 5-7) and unimodal methods (rows 8-9). We report the mean-average precision (MAP), precision (P20) and recall (R20) at 20 documents and number of relevant retrieved documents (RR).

- We proposed the use of AIA labels for image retrieval in both annotated and un-annotated image collections [12, 13]. For combining labels and text we developed methods for representing documents that take advantage of the semantic cohesion among terms from multiple modalities [22]. We gave evidence that shows that the use of AIA labels can be helpful to improve the image retrieval performance of unimodal techniques [22]; furthermore, the proposed representations for multimodal documents outperform baseline methods that proved to be very effective. According to our knowledge, our work is the first that explores the approach of combining information from labels and text.

- In collaboration with the TIA research group, we designed, developed and released the SAIAPR TC12 benchmark, a new resource that allows the evaluation of image annotation methods as well as studying the impact of those methods into multimodal image retrieval [15, 17].

In addition to the above described contributions, during the development of our research we accomplished other important achievements [14, 16, 9, 18, 23]. In particular, we would like to emphasize the development of particle swarm model selection, a novel technique for the automated selection of classifiers, such work was motivated by the need of the development of highly effective classification methods [18, 23], we refer the reader to [18] for further details.

7 Conclusions and future work

Interesting findings were derived from our work, the most important are listed below:

- We have provided experimental evidence that shows that the idea of semantic cohesion can be effectively exploited for modeling multimodal information. The proposed methods for image annotation and image retrieval that are based on such idea obtained superior performance than that reported in related works; furthermore, our techniques offer additional benefits. Thus, we can conclude that the semantic cohesion modeling, and more specifically, that a modeling based on co-occurrence statistics offer important benefits in terms of effectiveness, efficiency and representation power.

- On automatic image annotation we found that a one-vs-all approach with a random forest (RF) classifier is particularly helpful for region labeling. In this aspect, the annotation accuracy that can be obtained by selecting the correct label for each region, starting from its set of candidate labels is very important, hence, we think the development of refinement techniques (similar to ours) is an important research topic.

  The energy-based model (EBM) that we proposed can improve the output of the RF classifier, obtaining superior performance to that reported in the state-of-the-art on a variety of image collections. Since the EBM requires information that is easy to obtain its range of application is much more wide than that of other methods; also, the EBM is more efficient than other techniques. Moreover, the combination of labels generated with the EBM and text can be helpful for outperforming unimodal and standard multimodal retrieval techniques. In consequence, we can conclude that our annotation approach is highly competitive (in terms of both efficiency and efficacy) and advantageous over similar annotation techniques.

- We provided experimental evidence that shows that the combination of labels and text can improve the retrieval performance of unimodal methods, even when standard information fusion techniques were used. Despite the latter is highly intuitive, according to our knowledge, there are not similar works that attempt to combine text and labels. Nevertheless, we found that in order to obtain satisfactory results with standard techniques a few parameters have to be tuned, which limits the generality of such methods. We can conclude in this aspect that labels can be helpful for multimodal image retrieval, although it is not trivial the development of techniques that can obtain acceptable results under this scheme.

- In image retrieval we also found that the proposed representations, which are based on semantic cohesion, can effectively model the content of multimodal documents by means of multimodal distributional term representations. In particular we found that by using the multimodal document occurrence representation we can obtain better retrieval performance than that of standard techniques. Both representations can capture aspects of multimodal documents that with current techniques would not be possible. Additionally, the proposed representations do not depend of an effective parameter tuning phase for obtaining acceptable performance. Therefore, we conclude that the proposed representations model effectively the content of documents and hence better retrieval performance can be obtained with such techniques.

Despite we obtained satisfactory results with the proposed methods we have identified the following limitations. On the one hand, in image annotation, the improvement due to our EBM is still small, even when the potential of improvement as provided by RF classifiers is considerable. Also, the EBM requires of segmented images which may be difficult to obtain and/or of limited quality. Further, the EBM relies on supervised learning techniques, which require of manually labeled examples that can be difficult to obtain. On the other hand, in image retrieval, we have found that the dimensionality of the document representations based on DTRs can be very high, which limits the applicability of such methods for large image collections. Also, the way in which the DTRs of terms are combined for representing documents can be further improved.

Because of the above limitations we would like to explore the following topics for future work. On image annotation we would like to study the use of external resources for computing co-occurrence statistics under the proposed EBM; we believe that this aspect may have a positive impact into the semantic cohesion modeling. We are also interested on incorporating global information extracted from the images into the energy function of the EBM; in this way, we will consider both region-level and image-level information, which we think will improve the effectiveness of the EBM. Alternatively, we can iteratively combine the application of an AIA image-level method followed by a region-level AIA technique, in such a way that the local method refines the predictions of the global technique. Further, we would like to modify the EBM so that it can be used with semi-supervised learning techniques with the goal of reducing the degree of supervision required by the EBM. Additionally, we would like to develop retrieval techniques that can take advantage of the region-level information as provided by our AIA techniques.

On image retrieval, we would like to explore dimensionality reduction techniques over the DTRs of terms so that our methods can be applied to large scale image collections. Also, we would like to explore alternative ways for representing documents according to the DTRs. Further, we would like to extensively study the redundancy and complementariness of information offered by the DOR and TCOR representations. Finally, we believe that the application of our methods to other related problems can be a fruitful research topic. For example, the EBM can be applied to other structured prediction problems (e.g., handwritten-word classification) and the representations based on DTRs can be used to obtain richer visual vocabularies in the tasks of object recognition and image categorization under the bag-of-visual-words approach.

Acknowledgements

We would like to thank the members of the *TIA* research group at *INAOE*. This work was partially supported by *CONACyT* under project grant No. 61335 and scholarship No. 205834.

References

[1] M. Allan and J. Verbeek. Ranking user-annotated images for multiple query terms. In *Proceedings of the 20th British Machine Vision Conference*, London, UK, 2009.

[2] T. Arni, M. Sanderson, P. Clough, and M. Grubinger. Overview of the ImageCLEFphoto 2008 photographic retrieval task. volume 5706 of *LNCS*, pages 500–511. Springer, 2008.

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

[4] K. Barnard, Q. Fan, R. Swaminathan, A. Hoogs, R. Collins, P. Rondot, and J. Kaufhold. Evaluation of localized semantics: Data, methodology, and experiments. *International Journal of Computer Vision*, 77(1–3):199–217, 2008.

[5] D. Blei. *Probabilistic Models of Text and Images*. PhD thesis, U.C. Berkeley, 2004.

[6] P. Clough, M. Grubinger, T. Deselaers, A. Hanbury, and H. Müller. Overview of ImageCLEF2006 photographic retrieval and object annotation tasks. volume 4730 of *LNCS*, pages 579–594. Springer, 2006.

[7] P. Clough, M. Grubinger, T. Deselaers, A. Hanbury, and H. Müller. Overview of the ImageCLEF 2007 photographic retrieval task. volume 5152 of *LNCS*, pages 433–444. Springer, 2007.

[8] R. Datta, D. Joshi, J. Li, and J. Z. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys*, 40(2):1–60, 2008.

[9] T. Deselaers, A. Hanbury, and V. Viitaniemi, et al. Overview of the imageclef 2007 object retrieval task. volume 5152 of *LNCS*, pages 445–471. Springer, 2008.

[10] P. Duygulu, K. Barnard, J. F. G. de Freitas, and D. A. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *Proceedings of the 7th European Conference on Computer Vision*, volume 2353 of *LNCS*, pages 97–112, London, UK, 2002. Springer.

[11] H. J. Escalante. *Cohesión Semántica para la Anotación y Recuperación de Imágenes*. PhD thesis, Computational Sciences Department, National Institute of Astrophysics, Optics and Electronics, Puebla, Mexico, 2010.

[12] H. J. Escalante, J. A. González, C. A. Hernández, A. López, M. Montes, E. Morales, L. E. Sucar, and L. Villaseñor. Annotation-based expansion and late fusion of mixed methods for multimedia image retrieval. volume 5706 of *LNCS*, pages 669–676. Springer, 2008.

[13] H. J. Escalante, C. Hernández, A. López, H. Marín, M. Montes, E. Morales, E. Sucar, and Luis Villaseñor. Towards annotation-based query and document expansion for image retrieval. volume 5152 of *LNCS*, pages 546–553. Springer, 2007.

[14] H. J. Escalante, C. Hernandez, E. Sucar, and M. Montes. Late fusion of heterogeneous methods for multimedia image retrieval. In *Proceedings of the 2008 ACM Multimedia Information Retrieval Conference*, pages 172–179, Vancouver, British Columbia, Canada, 2008. ACM Press.

[15] H. J. Escalante, C. A. Hernández, J. A. González, A. López, M. Montes, E. Morales, L. E. Sucar, L. Villaseñor, and M. Grubinger. The segmented and annotated IAPR TC-12 benchmark. *Computer Vision and Image Understanding*, 114(4):419–428, 2010.

[16] H. J. Escalante, M. Montes, and E. Sucar. An energy-based model for feature selection. In *WCCI08 Workshop on Causality Challenge*, Hong-Kong, China, 2008.

[17] H. J. Escalante, M. Montes, and E. Sucar. On the SAIAPR TC-12 benchmark. In *Proceedings of the 2009 Theseus/ImageCLEF Workshop*, pages 44–51, Corfu, Greece, 2009.

[18] H. J. Escalante, M. Montes, and E. Sucar. Particle swarm model selection. *Journal of Machine Learning Research*, 10:405–440, February 2009.

[19] H. J. Escalante, M. Montes, and L. E. Sucar. Improving automatic image annotation based on word co-occurrence. In *Proceedings of the 5th International Adaptive Multimedia Retrieval workshop*, volume 4918 of *LNCS*, pages 57–70, Paris, France, 2007. Springer.

[20] H. J. Escalante, M. Montes, and L. E. Sucar. Word co-occurrence and Markov random fields for improving automatic image annotation. In *Proceedings of the 18th British Machine Vision Conference*, volume 2, pages 600–609, Warwick, UK, 2007.

[21] H. J. Escalante, M. Montes, and L. E. Sucar. An energy-based model for region-labeling. *Computer Vision and Image Understanding*, submitted, 2010.

[22] H. J. Escalante, M. Montes, and L. E. Sucar. Multimodal indexing based on semantic cohesion for image retrieval. *Submitted to Information Retrieval Journal*, 2010.

[23] H. J. Escalante, M. Montes, and L. Villaseñor. Particle swarm model selection for authorship verification. In *Proceedings of the 14th Iberoamerican Congress on Pattern Recognition*, volume 5856 of *LNCS*, pages pp. 563–570, Guadalajara, Mexico, 2009.

[24] H. J. Escalante, E. Sucar, and M. Montes. *Applied Swarm Intelligence*, chapter Multi-class Particle Swarm Full Model Selection for Automatic Image Annotation, Accepted. Springer Series in Studies in Computational Intelligence. Springer, 2010.

[25] S. Feng, R. Manmatha, and V. Lavrenko. Multiple bernoulli relevance models for image and video annotation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1002–1009, Washington, DC, USA, 2004. IEEE.

[26] C. Galleguillos, A. Rabinovich, and S. Belongie. Object categorization using co-occurrence, location and appearance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, Anchorage, Alaska, 2008.

[27] A. Goodrum. Image information retrieval: An overview of current research. *Journal of Informing Science*, 3(2), 2000.

[28] M. Grubinger. *Analysis and Evaluation of Visual Information Systems Performance*. PhD thesis, School of Computer Science and Mathematics, Faculty of Health, Engineering and Science, Victoria University, Melbourne, Australia, 2007.

[29] M. Grubinger, P. Clough, H. Müller, and T. Deselaers. The IAPR TC-12 benchmark: A new evaluation resource for visual information systems. In *Proceedings of the International Workshop OntoImage'2006 Language Resources for Content-Based Image Retrieval*, Genova, Italy, 2006.

[30] A. Hanbury. A survey of methods for image annotation. *Journal of Visual Languages and Computing*, 19(5):617–627, 2008.

[31] C. Hernandez and L. E. Sucar. Markov random fields and spatial information to improve automatic image annotation. In *Proceedings of the 2007 Pacific-Rim Symposium on Image and Video Technology*, volume 4872 of *LNCS*, pages 879–892, Santiago, Chile, 2007. Springer.

[32] M. Inoue. On the need for annotation-based image retrieval. In *Proceedings of the Workshop on Information Retrieval in Context*, pages 44–46, Sheffield, UK, 2004.

[33] J. Jeon, V. Lavrenko, and R. Manmatha. Automatic image annotation and retrieval using cross-media relevance models. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 119–126. ACM Press, 2003.

[34] A. Lavelli, F. Sebastiani, and R. Zanoli. Distributional term representations: An experimental comparison. In *Proceedings of the International Conference of Information and Knowledge Management*, pages 615–624. ACM Press, 2005.

[35] M. S. Lew, N. Sebe, C. Djeraba, and R. Jain. Content-based multimedia information retrieval: State of the art and challenges. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 2(1):1–19, 2006.

[36] Y. Liu, D. Zhang, G. Lu, and W. Ma. A survey of content-based image retrieval with high-level semantics. *Pattern Recognition*, 40(1):262–282, 2007.

[37] A. Llorente and S. Rüger. Using second order statistics to enhance automated image annotation. In *Proceedings of the 31st European Conference on Information Retrieval*, volume 5478 of *LNCS*, pages 570–577, Toulouse, France, 2009. Springer.

[38] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. Miller. Introduction to wordnet: An on-line lexical database. *International Journal of Lexicography*, 3(4):235–244, 1990.

[39] 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)*, Orlando, FL, USA, 1999.

[40] H. Müller, S. Maillet-Marchand, and T. Pun. The truth about corel - evaluation in image retrieval. In *Proceedings of the International Conference on Image and Video Retrieval*, volume 2383 of *LNCS*, pages 38–49, London, UK, 2002. Springer.

[41] G. Papadopoulos, V. Mezaris, I. Kompatsiaris, and M. Strintzis. Combining global and local information for knowledge-assisted image analysis and classification. *EURASIP Journal on Advances in Signal Processing*, Article ID 45842, 15 pages, 2007.

[42] G. Papadopoulos, C. Saathoff, M. Grzegorzek, V. Mezaris, I. Kompatsiaris, S. Staab, and M. Strintzis. Comparative evaluation of spatial context techniques for semantic image analysis. In *Proceedings of the 10th International Workshop on Image Analysis for Multimedia Interactive Services*, pages 161–164, London, UK, 2009. IEEE.

[43] Y. Rui, T. Huang, and S. Chang. Image retrieval: current techniques, promising directions and open issues. *Journal of Visual Communication and Image Representation*, 10(4):39–62, april 1999.

[44] Y. Rui, T. Huang, M. Ortega, and S. Mehrotra. Relevance feedback: A power tool for interactive content-based image retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 8(5):644–655, 1998.

[45] C. Saathoff, M. Grzegorzek, and S. Staab. Labeling image regions using wavelet features and spatial prototypes. In *Proceedings of the 3rd International Conference on Semantic and Digital Media Technologies: Semantic Multimedia*, volume 5392 of *LNCS*, pages 89–104, Koblenz, Germany, 2008. Springer.

[46] C. Saathoff and S. Staab. Exploiting spatial context in image region labeling using fuzzy constraint reasoning. In *Proceedings of the 9th International Workshop on Image Analysis for Multimedia Interactive Services*, pages 16–19, Klagenfurt, Austria, 2008. IEEE.

[47] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling appearance, shape and context. *International Journal on Computer Vision*, 81(1):2–24, 2008.

[48] A. W.M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1349–1380, 2000.

[49] J. Vogel and B. Shiele. Semantic modeling of natural scenes for content-based image retrieval. *International Journal on Computer Vision*, 72(2):133–157, 2007.

[50] J. Winn, A. Criminisi, and T. Minka. Object categorization by learned universal visual dictionary. In *Proceedings of IEEE International Conference on Computer Vision*, pages 1800–1807, Beijing, China, 2005. IEEE.