Evaluating Term-Expansion for Unsupervised Image Annotation

Luis Pellegrin, Hugo Jair Escalante, and Manuel Montes-y-Gómez

Computer Science Department,
Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE),
Tonantzintla, Puebla, 72840, Mexico
{pellegrin,hugojair,mmontesg}@inaoep.mx

Abstract. Automatic image annotation (AIA) deals with the problem of automatically providing images with labels/keywords that describe their visual content. Unsupervised AIA methods are often preferred because they can annotate (virtually) any possible concept to images and do not require labeled data as their supervised counterparts. Unsupervised AIA methods use a reference collection of images with associated (unstructured, freeform) text to annotate images. Thus, this type of methods heavily rely on the quality of the text in the reference collection. With the goal of improving the annotation performance of unsupervised AIA methods, we propose in this paper a term expansion strategy that expands the text associated with images from the reference collection. The proposed method is based on term co-occurrence analysis. We evaluate the impact that the proposed expansion has in the annotation performance of a straight unsupervised AIA method using a benchmark for large scale image annotation. Two types of associated text are used and several image descriptors are considered. Experimental results show that, by using the proposed expansion, better annotation performance can be obtained, where the improvements depend on the type of associated text that is considered.

1 Introduction

Since the last decade, the development of multimedia devices has facilitated the generation of vast amounts of images and videos which are stored locally by users or shared via the internet; for instance, only in Facebook about 300 million images are uploaded every day[1]. The availability of such amounts of data makes necessary the development of methods and tools that can allow users to access the images they are interested on. In this context, a relevant task is that of Automatic Image Annotation (AIA), i.e., the task of assigning labels/keywords to images that describe their visual content [1]. The importance of AIA lies in that these type of methods allow users to search for images by using keywords [2].

[1] http://gigaom.com/2012/10/17/facebook-has-220-billion-of-your-photos-to-put-on-ice/

The AIA task has been traditionally faced with two main approaches: supervised and unsupervised ones [3]. The difference between these two approaches lies on whether they use manually labeled data or not, and on how the annotation process is performed. Supervised methods require a training data set with labeled images to learn the correspondence between images and labels. This type of methods have reported competitive performance, see e.g., [4]. However, an important limitation of these methods is that new images can be labeled only with concepts available in the training set, besides, they require of manually labeled data and have been applied in scenarios with a few labels (typically a few tens).

Unsupervised methods, on the other hand, assign concepts to images by processing the text available in a reference collection of images, where each image is associated to a free and unstructured text[2]. For labeling an image, unsupervised AIA methods apply text mining techniques to the texts associated to images (in the reference collection) that are most similar to the test/query image [5, 6]. By working with unsupervised AIA it is possible to consider a larger diversity of labels for annotation than with supervised methods, besides, no manually-labeled data is required and more scalable systems can be developed [7]. Since reference collections for unsupervised AIA are not "curated", the quality of text plays a key role on the performance of AIA.

This paper aims at improving the scope that images in the reference collection have into annotation process for unsupervised AIA. Specifically, we aim at expanding the textual information associated to images in such a way that not only concepts/labels present in original text can be used for annotation, but also, related terms. Our hypothesis is that by using term co-occurrence information we can discover related terms that can be used to label images. This working hypothesis has been proved to be helpful in related tasks [4, 8]. However, to the best of our knowledge, this form of expansion has not been used in the context of AIA. We introduce a term-expansion strategy based on co-occurrence statistics and evaluate the benefits of using the expanded text for unsupervised AIA. We report experimental results in a large scale image annotation benchmark and show that term expansion can be very helpful for improving the annotation performance of unsupervised AIA.

The rest of this paper is organized as follows. The next section reviews related work. Section 3 describes the considered AIA method and introduce the proposed expansion strategy. Section 4 presents experimental results using two different types of associated text. Finally, Section 5 presents conclusions and discusses directions for future work.

2 Related work

In unsupervised AIA one has access to a reference collection of images with associated texts. Although this text is somehow related to images, it is important to emphasize that images are not labeled with and therefore supervised learning

[2] For instance, a reference collection might be a subset of images in the Web, where images are associated to the text in the webpage they are contained in.

is not an option. Instead, standard unsupervised AIA methods annotate images in two steps as follows. Given an image to annotate, first a content-based image retrieval (CBIR) module is used to obtain the $k-$most similar images to the query image; next, the texts associated to these k images are processed to obtain a set of keywords to annotate the image.

The works of Makadia et al. [5] and Villegas et al. [6] are two representative methods of unsupervised AIA, where the main difference between them is the way in which associated text is processed to obtain labels for an image. In [5] the authors propose a greedy strategy, where the most frequent terms in the first text are used as labels, if the number of desired labels is reached the method stops, otherwise, it keeps assigning label to images greedily. The method described in [6], on the other hand, assign to images the labels that mostly co-occur with terms in the retrieved texts[3]. In this paper, we consider an AIA method similar to that proposed in [6], however, we expand the terms associated to images before applying the annotation strategy. In this way, not only terms appearing in the text can be used for annotation, but also related terms.

Other interesting techniques have been proposed to obtain the annotation for images. For instance, in [9] it is proposed a measure to estimate the relevance of labels to images; this estimate is based on the difference of the distribution of the label to annotate in the k images returned by the CBIR and its distribution in a reference collection of images. In [10] BM25 (Okapi best matching 25) is used to assign weights to labels, and images are annotated with labels having the greatest scores. Both methods could be used in combination with our proposal, however, in this work we focused on the straight annotation process and postpone the use of more elaborated methods for future work.

3 Textual expansion for unsupervised annotation

This section describes the proposed term expansion technique and how we use it for AIA, see Figure 1. Before introducing the expansion method we describe the considered method for unsupervised AIA, which resembles the method in [6].

3.1 Unsupervised image annotation

Figure 1 shows the considered AIA method. Given an image to annotate: first a CBIR module is used to retrieve $k-$most similar images from a reference collection. Next, using the associated text of these k images, a text mining module is used to derive the labels to annotate. A standard AIA method (e.g., those in [6, 5]) comprises the modules within the box; the improvement reported in this paper comprises the modules below the box (see the next subsection).

The CBIR module involves the extraction of visual features for representing images and using a similarity/distance function to compare query and reference

[3] In [6] a set of reference labels is considered, in such a way that only labels in the reference set can be used for labeling images.

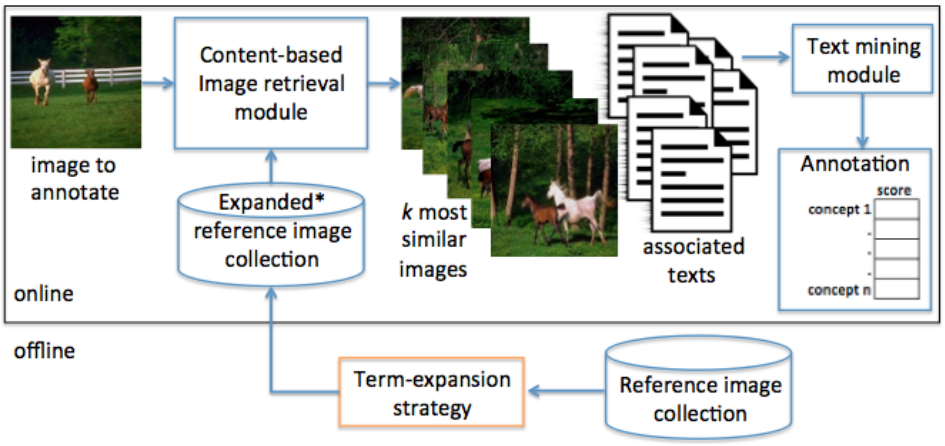

Fig. 1. Unsupervised AIA with the proposed term-expansion strategy. *Without expansion follows a traditional unsupervised AIA.

images [2]. We considered standard visual descriptors such as SIFT and color histograms, as provided with the considered data set (see Section 4). As distance function we used the L1 distance:

$$L1(\mathbf{x}, \mathbf{y}) = \frac{1}{D} \sum_{i=1}^{D} |x_i - y_i| \tag{1}$$

where $\mathbf{x}, \mathbf{y} \in \mathbb{R}^D$ are vectors of visual descriptors representing two images, x_i is the i^{th} element of vector $\mathbf{x}$, and D is the dimensionality of the input space.

Once that $k-$images have been retrieved, the annotation process consists of mining the text associated to these images to extract the terms to label the query image. For this process we considered a straightforward strategy that accounts for the frequency of terms in the k retrieved texts. Candidate terms to be used as labels for the query image are sorted in descending order of their frequency, and we assign to the image the top $q-$concepts with highest scores.

We have described a standard/basic AIA method. One should note that this AIA strategy can be used to label images with both: the standard setting, i.e., without term expansion, (the modules in the box in Figure 1) and the proposed term-expansion approach (the whole approach depicted in Figure 1).

3.2 Term-expansion strategy

This section describes the proposed term expansion technique, which aims at improving the quality of the reference collection, see Figure 1. Our approach aims to associate images with similar contexts to each other, i.e., images that are associated to terms that co-occur with similar terms across the reference

collection. Thus, the text associated to images is augmented with related terms; or, from a different perspective, one can see this expansion as expanding the set of images that contain each term, see Figure 4.

More formally, let $\mathcal{C}$ denote the reference collection formed by image-text pairs, i.e., $\mathcal{C} = \{(I_i, T_i)\}_{1,\ldots,M}$, where each image I_i is represented by a vector of visual features $\mathbf{x}_i \in \mathbb{R}^D$ and its associated text T_i is represented by a bag-of-words $\mathbf{a}_i \in \mathbb{R}^{|V|}$, that is, $\mathbf{a}_i = \langle a_{i,1}, \ldots, a_{i,|V|} \rangle$ is a $|V|-$dimensional vector of reals where the j^{th} element indicates the (normalized) frequency of occurrence of term t_j in text T_i, where V is the vocabulary in the reference collection (i.e., the set of different terms in texts from the reference collection). The proposed method expands each T_i component (i.e., the vector $\mathbf{a}_i$), resulting in an expanded collection $\mathcal{C}^* = \{(I_i, T_i^*)\}_{1,\ldots,M}$, where T_i^* is a modified version of T_i that entails the expanded information (i.e., some terms in $\mathbf{a}_i^*$ that were zero in $\mathbf{a}_i$ can have now non-zero values). The rest of this section describes the way in which bag-of-words vectors are expanded.

The expansion strategy relies on term co-occurrence analysis. As a first step, we quantify the degree of association between terms by estimating co-occurrence statistics in the texts from the reference collection. Specifically, we consider the following term-relatedness measure which aims at approximating the conditional probability of occurrence of a term given another one:

$$P(t_k|t_j) \approx \frac{O(t_j, t_k)}{O(t_j)} \tag{2}$$

where $O(t_k, t_j)$ is the number of documents[4] in which terms t_k and t_j co-occur, and $O(t_j)$ is the number of documents in which term t_j appears. Let $\mathbf{a}_i^* = \langle a_{i,1}^*, \ldots, a_{i,|V|}^* \rangle$ denote the expanded textual representation T_i^* associated to image I_i. Then, the expansion for term $a_{i,k}^*$ is given by:

$$a_{i,k}^* = \sum_{j=1}^{n} a_{i,j} \cdot P(t_k|t_j) \tag{3}$$

where t_j, $j = \{1, \ldots, n\}$, denote the terms that appear in the original text T_i and $P(t_k|t_j)$ expresses the relatedness of the term t_j to the term t_k. This expansion is performed over the whole vocabulary (e.g., every element $a_{i,k}$ of $\mathbf{a}_i$ is expanded). In this way, the whole vector of terms is expanded by looking at the context of the corresponding associated text. One should note, that the expanded textual vector increases the number of terms associated to each image, which latter is used for AIA.

Although it is important to expand every term in the vocabulary in terms of the context of the associated text, it is often possible to introduce noisy information that may degrade the annotation performance. Hence, we considered a more controlled form of expansion in which only terms that co-occur with

[4] We consider as a document the pair of image and associated text in the reference collection.

high frequency will be expanded. We use Equation (4) to obtain the $P(t_k|t_j)$ component in Equation (3), to expand only a subset of terms as determined by a threshold u.

$$P(t_k|t_j) = \begin{cases} 0 & \text{if } P(t_k|t_j) < u \\ P(t_k|t_j) & \text{if } P(t_k|t_j) \geq u \end{cases} \quad (4)$$

In order to define u we took into consideration several options and we found that, in general, it is difficult to fix an optimal cutoff value for every possible situation. Therefore, instead of using a fixed threshold we adopted a dynamic value that depends on the mean and standard deviation on the co-occurrence values of each term:

$$u = \mu(P(t_{1,\ldots,n}|t_j)) + \sigma(P(t_{1,\ldots,n}|t_j)) \quad (5)$$

The next section presents an experimental study that aims at evaluating the benefits offered by the proposed expansion strategy.

4 Experimental results

This section reports results of experiments that evaluate the proposed expansion strategy in the task of unsupervised AIA. First we describe the considered data set, next we present the experimental settings, then we report the experimental results.

4.1 Scalable concept annotation data set

For the evaluation of the proposed expansion strategy we considered a benchmark used in the scalable concept annotation subtask at ImageCLEF2013 [11, 6]. It was created by filtering out over 31 million images that were obtained by querying three popular search engines; a subset of 250,000 images (together with the webpages that contained the images) was selected to be used as the reference collection. A subset of 1000 images was manually selected and labeled with n concepts, taken from a vocabulary of 107 possible concepts that are used for the evaluation of the annotation performance. The reference collection includes the 250,000 images and their corresponding associated texts. For the associated text we have used two resources: the complete WebPage that contained each image and the keywords that were used to retrieve the image. Images are represented by 7 variants of visual descriptors: SIFT, color histogram, GETLF, GIST and three subtypes of color SIFT descriptors (C, RGB and OPPONENT SIF), all of them are histograms that account for the occurrence of representative visual descriptors taken from a codebook (bag-of-visual words). These descriptors were made available by the owners of the data set [11, 6].

4.2 Experimental settings

The text in the reference collection was indexed[5] to obtain the bag-of-words representations, $\mathbf{a}_i$, for each text T_i. This indexing procedure was applied separately for the two types of textual information (i.e., keywords and WebPages). In order to make more manageable the vocabulary of terms from the WebPages stop words and terms with very low/high frequency were removed.

For the evaluation of our method we adopted the same protocol from the scalable concept annotation task [6]: given an input image and a set of $c-$specific concepts the system decides which of them are present in the image and which ones are not. In fact, the annotation methods provide a ranking of the $c-$concepts in descending order of their relevance to describe the query image. The evaluation is performed by estimating the Average Precision (AP) of the ranked list of concepts as follows:

$$AP = \frac{1}{|G|} \sum_{g=1}^{|G|} \frac{t}{rank(g)} \tag{6}$$

where G is the ordered set of the ground truth annotations, and $rank(g)$ is the ordered position, in the ranked list provided by the AIA system, of the g^{th} ground truth annotation. Thus, lager values of AP indicate better annotation performance. For all of our experiments we associate each image with the top-10 concepts/terms with higher frequency (see Section 3). Hence, our system always return a list of 10 concepts, sorted in descending order of relevance.

4.3 Experiment 1: WebPages vs. keywords annotation

The aim of this experiment was to evaluate the annotation performance before applying the expansion strategy by using two different types of associated text in the reference collection: 1) WebPages, and 2) keywords. We considered all of the visual descriptors for the CBIR module and compare the annotation performance for different values of k (the number of nearest images considered for AIA, see Section 3.). The results of this experiment are shown in Figure 2.

We can clearly observe that better annotation performance was obtained with WebPages using any of the seven visual descriptors (see Figure 2 (a)). We believe that the WebPages include more information that can be used to describe the visual content of images, so that the WebPages offer better annotation performance than keywords, which only include few words. Thus, despite being cleaner, keywords do not contain enough useful-information and, therefore, it seems this type of information is more appropriate to be expanded with our proposed technique. On the other hand, it can be seen that, in terms of visual descriptors (see Figure 2 (b)), better results were obtained with SIFT descriptors and its color-variants. Also, it is clear that, in general, better performance is obtained with larger values of k, this can be due to the fact that, in this way, more term-frequency is accumulated, which can be beneficial for AIA.

[5] We used the Text to Matrix Generator (TMG 5.0) in Matlab [12].

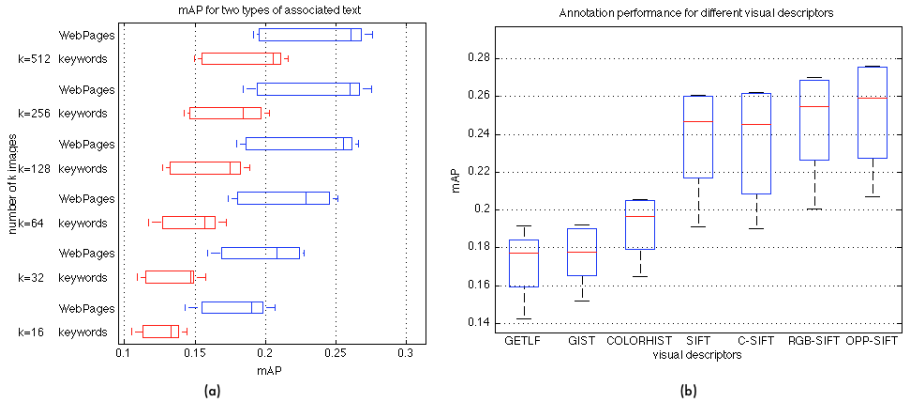

Fig. 2. mean Average Precision (mAP): (a) using keywords and WebPages as associated text in the reference collection using all the 7 different visual descriptors, and (b) the 7 different visual descriptors using keywords and WebPages.

4.4 Experiment 2: direct vs. expanded annotation

In a second experiment we evaluated the annotation performance when using the expanded text and compared this performance with the traditional AIA approach. For clarity, and because of results from the previous section, we only show results obtained with the OPPONENT-SIFT visual descriptor.

The results of this experiment are shown in Figure 3; we compare performance without expansion, with expansion of all terms (*exp-total*), and with controlled expansion (*exp-(M+std)*), see Section 3, for WepPages and keywords.

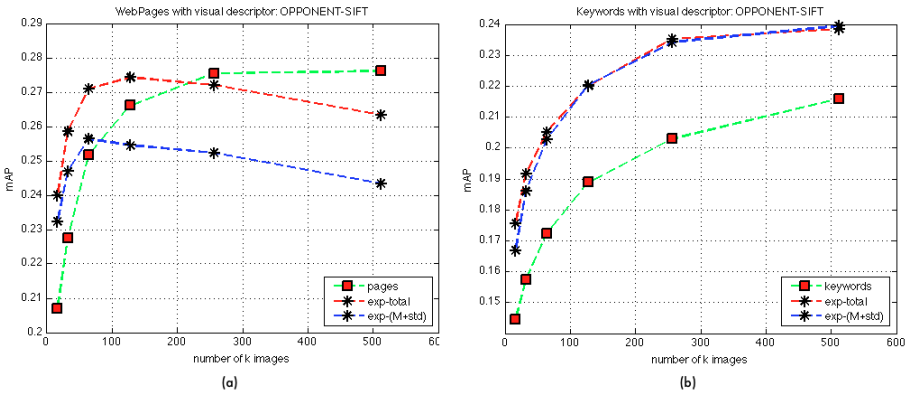

Fig. 3. mAP using the original text in the reference collection vs. expanded variants for WebPages (a) and keywords (b); OPPONENT-SIFT was used as visual descriptor.

Without expansion, we can observe that by using keywords the annotation performance tends to increase slowly up to $k \approx 512$ (Figure 3 (b)), and the performance using WebPages has a lapse of convergence after $k \approx 256$ images are used (Figure 3 (a)). We believe that the convergence in these two cases is due to a saturation of information, that is, there exist a majority of images that contain the concepts to annotate. On the other hand, using a complete expansion, we can observe that the performance with the keywords improved, whereas the performance with WebPages started to drop as the number of images increased. This behavior is somewhat expected as: (1) having rich enough information (WebPages) suggests an expansion will not may have an impact (instead, noise removal mechanisms may be helpful for this type of data); and (2) having scarce information (keywords) is indicative of the need for an expansion.

Finally, we can observe that if we perform a controlled expansion we can improve the annotation performance of WebPages using a less images (k) for the CBIR module. In fact, we can achieve similar performance when using $k = 512$ (without expansion) and $k = 128$ (with controlled expansion). This is an interesting result because we can use 4-times less images and still obtain quite competitive performance. One should note that when considering larger values of k the annotation performance decreases. This can be due to the fact that when estimating the threshold u for too many texts, non-relevant/noisy terms may be the more frequent ones.

In order to give insights into the quality of the expansion that can be generated with our method we show in Figure 4 some images with their associated keywords before and after applying the expansion procedure. For each image, we show at most the top-10 words most relevant from the expansion strategy.

It is interesting that we can find in the expanded text many related terms to the original keywords that can be used to label images. For example, using as original 'chortle', 'effectively' and 'exposes' (refer to third image in the first row in Figure 4), in the expansion we can see terms like 'comedian' and 'guffaw' that are related terms to 'chortle'. Another observation is that some terms are ambiguous like 'siberian' and these can express different concepts, as consequence the expansion includes different concepts (like 'cat' or 'purring') and not all are related to the visual content of the image (refer to second image of the first row).

On the other hand, we evaluate qualitatively the images related to a given concept before and after the expansion. In Figure 5, we present the 10-top related images to the concept 'cloud'; we can observe that before the expansion there are images not related to the concept but after the expansion we can find three related topics: 1) 'cloud' like meteorology definition, 2) 'cloud' related to computation and, 3) 'cloud' related to a video game.

In the Figure 6 we present another example of the expansion; we show 10-top related images to the concept 'traffic', we can observe an increase in the number of images related to 'traffic' like transportation definition as well as an increase in the number of images in 'traffic' related to information at the Internet.

Fig. 4. Annotation before and after the expansion. The terms as original are those extracted from the keywords.

5 Conclusions

In this paper we have introduced a new term expansion strategy for unsupervised image annotation. This strategy is based on term co-occurrence analysis; it aims to expand the text associated to images from the reference collection with related terms that could contribute to improve the annotation performance.

Our research on the expansion strategy is ongoing and requires an in depth evaluation and analysis with a larger pool of concepts/labels. However the results presented are encouraging. The set of experiments we discussed showed that, by using the proposed expansion, better annotation performance can be obtained, where the improvements depend on the type and amount of associated text that is considered. Particularly our results suggest that the expansion will not may have an impact when original textual information is rich enough (as in the case of full web pages), but that it could be really useful when this information is scarce (such as in the case of having a bunch of keywords).

For future work we are interested in expanding not only the textual information associated to images from the reference collection but also their visual representation. The idea is to enhance the performance of the CBIR module

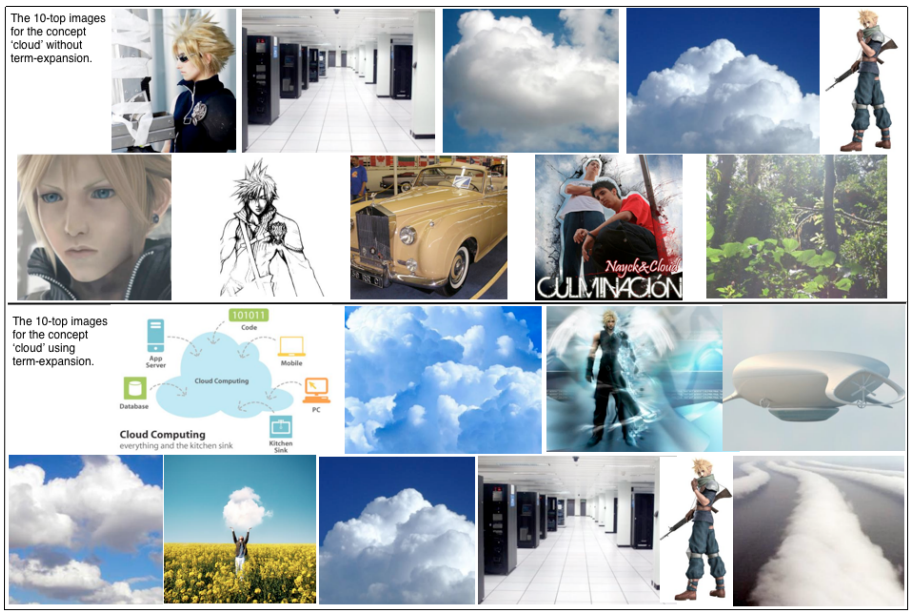

Fig. 5. 10-top related images to the concept 'cloud' before and after the expansion.

Fig. 6. 10-top related images to the concept 'traffic' before and after the expansion.

by allowing the retrieval of relevant images having related (although not equal) visual features.

Acknowledgments

This work was partially supported by CONACyT under scholarship No. 214764 and by the LACCIR programme under project ID R1212LAC006. Hugo Jair Escalante was supported by the internships programme of CONACyT under grant No. 234415. The authors would like to thank Mauricio Villegas and Roberto Paredes for their support on the considered data set.

References

1. Barnard, K., Duygulu, P., de Freitas, N., Forsyth, D., Blei, D., Jordan, M.: Matching words and pictures. Journal of Machine Learning Research **3** (2003) 1107–1135
2. Datta, R., Joshi, D., Li, J., Wang, J.: Image retrieval: Ideas, influences, and trends of the new age. ACM Computing Surveys (CSUR) **40** (2008)
3. Hanbury, A.: A survey of methods for image annotation. Journal of Visual Languages and Computing **19** (2008) 617–627
4. Escalante, H.J., Montes, M., Sucar, E.: An energy-based model for region labeling. Computer Vision and Image Understanding **115** (2011) 787–803
5. Makadia, A., Pavlovic, V., Kumar, S.: Baselines for image annotation. International Journal of Computer Vision **90** (2010) 88–105
6. Villegas, M., Paredes, R., Thomee, B.: Overview of the imageclef 2013 scalable concept image annotation subtask. In: CLEF 2013 Evaluation Labs and Workshop, Online Working Notes. (2013) 1–19
7. Carneiro, G., Chan, A., Moreno, P., Vasconcelos, N.: Supervised learning of semantic classes for image annotation and retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence **29** (2007) 394–410
8. Escalante, H.J., Montes, M., Sucar, E.: Multimodal document indexing based on semantic cohesion for image retrieval. Information Retrieval **15** (2012) 1–32
9. Uricchio, T., Bertini, M., Ballan, L., Del Bimbo, A.: MICC-UNIFI at ImageCLEF 2013 scalable concept image annotation. In: Working Notes for CLEF 2013 Conference, Valencia, Spain, September 23-26. (2013)
10. Reshma, I., Ullah, M., Aono, M.: KDEVIR at ImageCLEF 2013 image annotation subtask. In: Working Notes for CLEF 2013 Conference, Valencia, Spain, September 23-26. (2013)
11. Villegas, M., Paredes, R.: Image-text dataset generation for image annotation and retrieval. In Berlanga, R., Rosso, P., eds.: II Congreso Español de Recuperacion de Informacion, CERI'12. (2012) 115–120
12. Zeimpekis, D., Gallopoulos, E.: TMG: A MATLAB toolbox for generating term-document matrices from text collections. In: Grouping Multidimensional Data: Recent Advances in Clustering. Springer (2010) 187–210