Data Fusion and Label Weighting for Image Retrieval based on Spatio-Conceptual Information

Carlos Hernández-Gracidas, Antonio Juárez, L. Enrique Sucar,
Manuel Montes-y-Gómez and Luis Villaseñor
National Institute of Astrophysics, Optics and Electronics (INAOE), México
{carloshg, antjug, esucar, mmontesg, villasen}@inaoep.mx

ABSTRACT

Using as experimental platform an image retrieval method based on a spatio-conceptual representation of images, in this paper we investigate two main concerns on annotation-based image retrieval: label weighting and data fusion. On the one hand, we analyze the influence of different weighting schemes on the quality of the retrieval performance, and, on the other hand, we study the application of fusion techniques for queries represented by more than one sample image. Particularly, we aim to compare fusion methods based on score information and on ranking information in order to determine the most adequate approach for the annotation-based image retrieval task.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*query formulation, search process*

General Terms

Algorithms, Experimentation

Keywords

Content-based image retrieval, data fusion, term weighting, spatial relations, conceptual graphs

1. INTRODUCTION

Image retrieval is traditionally performed in two main ways [4]: from a set of sample images, or from a textual query. We are particularly interested in the retrieval based on sample images, which is frequently regarded as content-based image retrieval (CBIR). A variation of this approach, known as annotation-based image retrieval (ABIR), consists of using low-level image features to identify objects in the images and associate labels to them. The main problem of using low-level features is that they fail at distinguishing between two visually similar (but actually different) concepts, a problem known as the "semantic gap". In order to tackle this problem, some recent methods have considered the use of high-level features such as spatial relations [2, 10].

Regarding the use of spatial information in image retrieval, in a previous work [5] we proposed a method that employs conceptual graphs to represent the objects appearing in the images as well as their spatial relations. The achieved results were encouraging; they showed that the inclusion of spatial information allows to improve the retrieval performance. Motivated by these results, in this paper we attempt to go a step forward by analyzing the influence of different weighting schemes on the quality of the retrieval performance. We mainly compare three traditional weighting schemes against a new one specially suited for ABIR, which is based on the idea that common labels (those co-occurring with other several labels) tend to be less important than specific ones (co-occurring with only a small subset of the labels), without mattering their frequency of occurrence in the entire image collection.

In addition, we study the application of fusion techniques for queries represented by more than one sample image. Fusion methods are based on two elements from the items retrieved: rankings and scores. Round robin, SUM, combSUM, combMNZ and Condorcet use rankings [8, 12]; while retrieval status value (RSV) and fuzzy Borda count use scores [6, 9]. Particularly, we aim to compare two different fusion methods, one based on scores [8] and other based on rankings [6]. Our main interest is to investigate which one of these approaches is the most adequate for the ABIR task.

The rest of the paper is organized as follows. Sections 2 and 3 describe the core image retrieval method based on the use of spatial information. Section 4 presents the experimental platform, indicating how data fusion was carried out and describing the used weighting schemes. Section 5 presents the experimental results. Section 6 shows our conclusions, and Section 7 presents our future work.

2. IMAGE REPRESENTATION

The representation of images is based on spatial relations and conceptual graphs (CGs) [11]. On the one hand, spatial relations were divided into three groups, namely: topological relations, horizontal relations, and vertical relations (refer to Table 1). Whereas, on the other hand, CGs are used to express the spatial relations among labels (objects) from an image. CGs are finite, connected, and bipartite graphs formed of two types of nodes: concepts (in our case labels) and relations (in our case spatial relations). Figure 1 shows

RIAO'10, 2010, Paris, France.

Table 1: Spatial relations used in this work.

			Directed	Undirected
Topological		1		Adjacent
		2		Disjoint
Order	Horizontal	3		Beside (left or right)
		4		H. aligned
	Vertical	5	Above	
		6	Below	
		7		V. aligned

an example of how the images in the database [3] are segmented and annotated, and how the spatial relations in the images can be represented by means of CGs. In accordance with Table 1, all of the relations are undirected except for the vertical relations *above* and *below*.

3. IMAGE RETRIEVAL

The similarity between a pair of images is measured using two different similarity measures: conceptual similarity (S_c) and relational similarity (S_r). S_c measures how similar two graphs are by counting how many concepts (labels) they have in common, while S_r measures how similar the relations among the concepts in common are. The equations for both of them are:

$$S_c = \frac{2n(G_c)}{n(G_1) + n(G_2)} \tag{1}$$

$$S_r = \frac{2m(G_{Tc}) + 2m(G_{Hc}) + 2m(G_{Vc})}{3m_{G_c}(G_1) + 3m_{G_c}(G_2)} \tag{2}$$

The similarity between two images is measured by S, which considers both S_c and S_r, giving each a weight depending on a constant α, which is shown in the following equation:

$$S = \frac{\alpha S_c + (1 - \alpha) S_r}{2} \tag{3}$$

For two images to be compared, they have to be preprocessed by segmenting and annotating them. After this process is done, spatial relations are computed in order to build their CGs. Once we have the CG for both images, they can be compared based on Equation 3. Finally, by repeating this process, one image can be compared against the whole image database, and, therefore, it is possible to determine a ranked list of similar images. More details on the retrieval method and how to obtain the CGs can be found in [5].

4. EXPERIMENTAL SETUP

4.1 Database and Topics

The image database used for our experiments is the manually segmented and annotated IAPR-TC12 [3], consisting of 20,000 images of sports events, people, animals, cities and landscapes. It was chosen because it provides a reliable dataset that allows focusing more on image retrieval, than on the effects of the automatic segmentation and annotation. However, it must be highlighted that our method is applicable to automatic segmentation and annotation as well. Figure 2 shows some of the images from the IAPR-TC12.

For an objective evaluation of our method, we considered the 39 topics developed for the ImageCLEF 2008 photo retrieval task [1]. These topics are classified as visual (20 top-

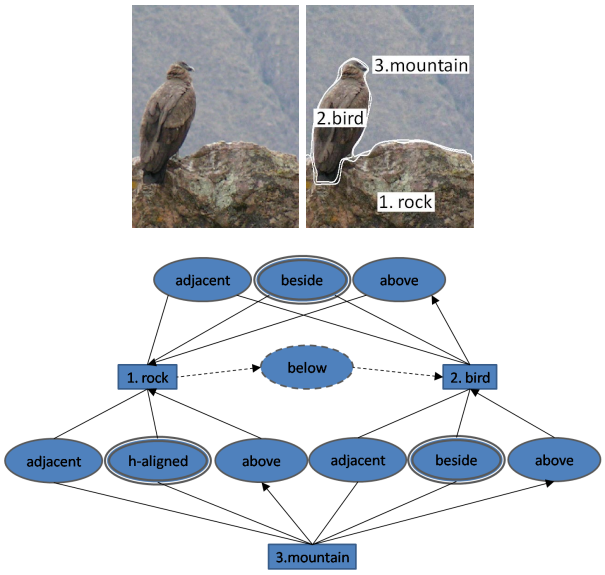

Figure 1: Top-left: One of the images in the IAPR-TC12. Top-right: The same image, segmented and annotated. Bottom: Conceptual graph indicating the spatial relations in the image. Topological relations are shown with filled nodes, horizontal relations appear with double-lined border, and vertical relations appear with single-lined border.

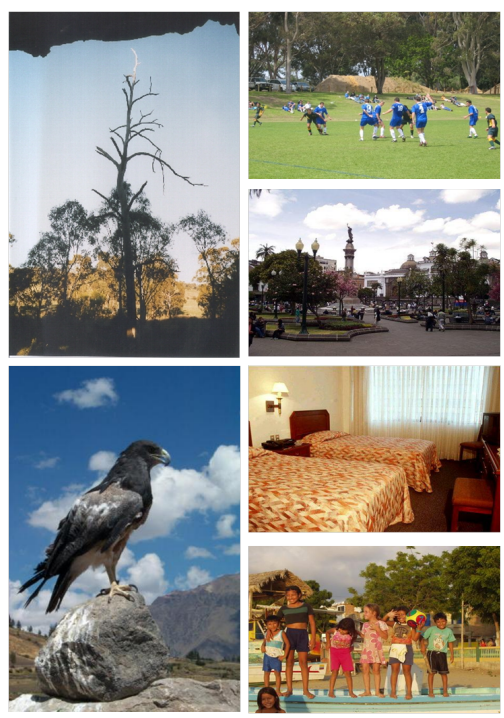

Figure 2: Sample images from the IAPR-TC12.

ics) and non-visual (19 topics), and are expressed by means of a textual description and three example images. Given that the proposed method is exclusively based on visual information we only considered for the experiments the subset of visual topics with their corresponding three sample images.

4.2 Label Weighting

We considered three traditional weighting schemes for the experiments:

- Equal weights: $w_i = 1$. It gives an equal weight to each label.
- Frequency-based weight: $w_i = \frac{1}{|\{I:l_i \in I\}|}$. A simple schema, where $|\{I : l_i \in I\}|$ is the total number of images where label l_i appeared.
- $TFIDF_{ij} = TF_{ij} \times IDF_i$ [7]. It is the traditional TFIDF measure, where $TF_{ij} = \frac{n_{ij}}{N_j}$ (the occurrences of label l_i in image I_j divided by the number of labels in I_j), and $IDF_i = log\frac{|D|}{\{I:l_i \in I\}}$ (the number of images in the collection is divided by the number of images containing label l_i).

In addition, we proposed and applied a modified version of the TFIDF weighting scheme, which is defined as follows:

$$MTFIDF_{ij} = TF_{ij} \times MIDF_i \quad (4)$$

In this case, TF_{ij} is the same as in $TFIDF_{ij}$, and $MIDF_i$ is defined as:

$$MIDF_i = log\frac{|D|}{\sum_{\{I:l_i \in I\}} \times N} \quad (5)$$

This is a modified version of IDF_i, where the number of images in the database ($|D|$) is divided by the sum of the number of labels in each image (N) where label l_i appeared ($\{I : l_i \in I\}$). The idea of this new weighting scheme is that labels co-occurring with many other labels are less important than those appearing with only a few labels.

4.3 Fusion Scheme

As we mentioned in Section 1, there are situations in which there is more than one sample image available per topic (for instance, in the ImageCLEF evaluation forum [1]). To take advantage of all this information, we considered the fusion of the ranked lists obtained per sample image (using the same retrieval method). The fusion process was carried out as follows: first, independent retrieval processes were performed using the IR method described in Section 3, obtaining the result lists L_1, L_2 and L_3 (one for each sample image). Then, we fused the three resulting lists, obtaining a new list L_f. After this, the 1,000 more similar images were kept and presented as the final result. This process is illustrated in Figure 3. In particular, in the experiments we compared two data fusion methods: SUM[1], which is a method based on rankings, and fuzzy Borda count[2], which is a method based on scores.

9[1]Although combMNZ is considered the baseline method in fusion tasks, in preliminary experiments we compared SUM and combMNZ, with a better performance from SUM, which is why we chose it.

9[2]Fuzzy Borda count is chosen, since it claims to outperform combMNZ in several tasks it has been used for.

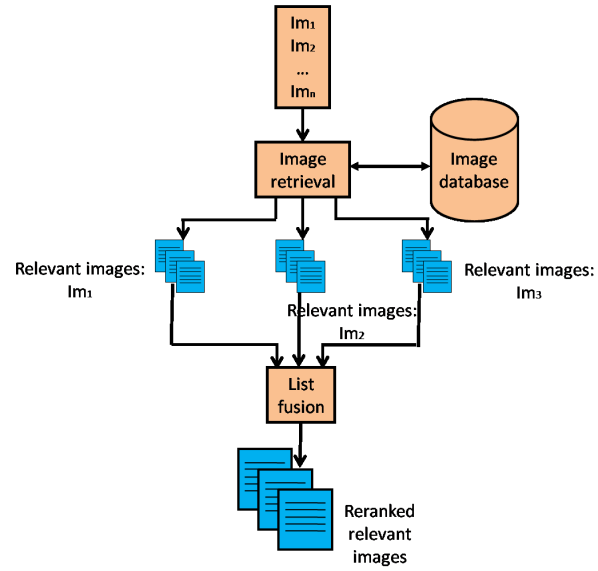

Figure 3: Block diagram of the list fusion method.

5. EXPERIMENTAL RESULTS

5.1 Importance of Label Weighting

This section presents the results of the comparison of the weighting schemes presented in 4.2. We used the SUM fusion method to perform the experiments, varying α in Equation 3 from 0 to 1, with increments of 0.1. We evaluated the retrieval performance using the MAP measure, and considered the application of equal weights to all of the labels as the baseline. Figure 4 shows the results of these experiments.

According to these results, it is clear that giving the labels a different weight is more adequate than using the same weight for all of them. It is also clear that, from the approaches giving different weights to labels, $MTFIDF_{ij}$ is the one that works better, outperforming the others for every value of α. Observing the figure, it was around $\alpha = 0.5$ where the best results were reached. On the basis of these results, the next section only shows the comparison of data fusion methods using the $MTFIDF_{ij}$ weighting scheme.

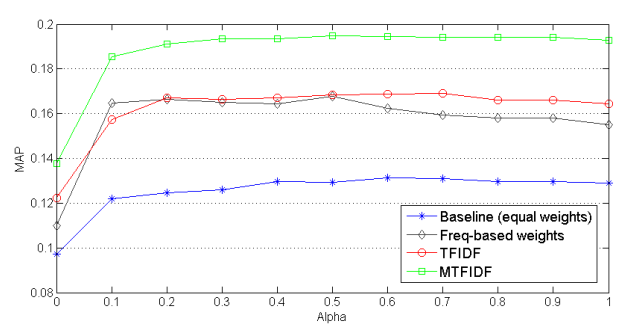

Figure 4: Comparison of the retrieval performance achieved by four different label weighting schemes.

5.2 Data Fusion: Rankings vs. Scores

The experiments reported in this section focused on the analysis of the behavior and the comparison of the performance of two fusion methods. Particularly, we compared the SUM method (based on rankings), and the fuzzy Borda count (based on scores). We intended to analyze which information, scores or rankings, is more relevant for the task at hand. We evaluated the retrieval performance based on the MAP measure, and considered the average of the three individual runs (using each of the three sample images) as the baseline. Figure 5 shows the results using $MTFIDF_{ij}$ and varying α in the [0,1] interval.

Results are clear; they show that both fusion methods significantly outperformed the results achieved by using sample images individually. Surprisingly, the simple SUM method outperformed fuzzy Borda in every case, indicating that the fusion based on scores performed better than that based on rankings for this dataset.

Experiments were run on a Centrino 2 computer, with a dual 2.0 GHz processor and 4 GB in RAM, using MATLAB. The average time for comparing an image to the 20,000 images in the database is 25.27 seconds, while the average time for comparing the three images and fusing their results is 68.49 seconds. It is likely that the average fusion time for the three images is less than three times the average comparison time of an individual image, given that some operations are performed only once for the fusion.

6. CONCLUSIONS

In this paper we evaluated label weighting and list fusion using a high-level spatio-conceptual representation of images. A spatio-conceptual retrieval method was used to retrieve lists of relevant images. Three different label weighting schemes were experimented in order to determine their influence on the retrieval performance. Also, two fusion methods were considered to investigate the usefulness of score and ranking information in the fusion process. The achieved results indicated the pertinence of the proposed weighting scheme ($MTFIDF_{ij}$). They also showed that, for the used result lists, scores were more useful than rankings.

7. FUTURE WORK

Regarding the data fusion process, as future work we plan to consider the situation where result lists are obtained from different retrieval methods, and also to study the relation between the redundancy among the lists and the performance of the fusion. In addition, we consider using the label hierarchy available for the used image database, and combining this fusion with other text-based methods.

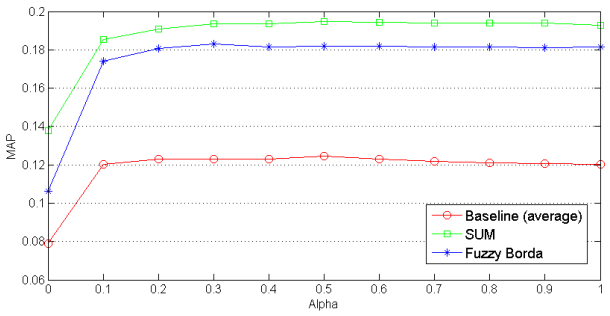

Figure 5: Rankings (SUM) vs. scores (fuzzy Borda), measured by MAP and using $MTFIDF_{ij}$.

8. ACKNOWLEDGMENTS

This research was partially supported by CONACyT under project grant 61335 and Ph.D. fellowships 205839 and 165499.

9. REFERENCES

[1] T. Arni, M. Sanderson, P. Clough, and M. Grubinger. Overview of the imageclef 2008 photographic retrieval task. *Working Notes of the CLEF*, 2008.

[2] S. Berretti, A. D. Bimbo, and E. Vicario. Weighted walkthroughs between extended entities for retrieval by spatial arrangement. *IEEE Transactions on Multimedia*, 5(1):52–70, 2003.

[3] H. Escalante, C. Hernández-Gracidas, J. González, A. López, M. Montes, E. Morales, L. Sucar, L. V. nor, and M. Grubinger. The segmented and annotated iapr-tc12 benchmark. *Computer Vision and Image Understanding, in press*, 2009.

[4] A. Goodrum. Image information retrieval: An overview of current research. *Informing Science*, 3:2000, 2000.

[5] C. Hernández-Gracidas, L. Sucar, and M. M. y Gómez. Modeling spatial relations for image retrieval by conceptual graphs. In *First Chilean Workshop on Pattern Recognition*, 2009.

[6] G.-L. J.L. and M.-P. M. A fuzzy borda count in multi-person decision making. *Multiple Objective and Goal Programming. Recent Developments. Advances in Soft Computing*, pages 46–60, 2002.

[7] K. S. Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28:11–21, 1972.

[8] K. Paul. Decision level data fusion for routing of documents in the trec3 context: A best case analysis of worst case results. In *Proceedings of The Third Text REtrieval Conference (TREC-3)*, 1995.

[9] J. Perea, L. Ureña, D. Buscaldi, and P. Rosso. Textmess at geoclef 2008: Result merging with fuzzy borda ranking. In *9th Int. Cross-Language Evaluation Forum CLEF-2008 working notes*, 2008.

[10] V. Rathi and A. K. Majumdar. Content based image search over the world wide web. In *Proceedings of the Indian Conference on Computer Vision, Graphics and Image Processing*, 2002.

[11] J. F. Sowa. *Conceptual structures: information processing in mind and machine*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1984.

[12] S. Wu and S. McClean. Performance prediction of data fusion for information retrieval. *Information Processing and Management*, 42(4):899–91, 2006.