Using a Markov Random Field for Image Re-ranking Based on Visual and Textual Features

Utilizando un Campo Aleatorio de Markov para el Reordenamiento de Imágenes Basado en Atributos Visuales y Textuales

R. Omar Chávez, Manuel Montes and L. Enrique Sucar
Coordinación de Ciencias Computacionales,
Instituto Nacional de Astrofísica Óptica y Electrónica,
Puebla, México
{romarcg,mmontesg,esucar}@ccc.inaoep.mx

Article received on January 15, 2010; accepted on June 29, 2010

Abstract. We propose a novel method to re-order the list of images returned by an image retrieval system (IRS). The method combines the original order obtained by the IRS, the similarity between images obtained with visual and textual features, and a relevance feedback approach, all of them with the purpose of separating relevant from irrelevant images, and thus, obtaining a more appropriate order. The method is based on a Markov random field (MRF) model, in which each image in the list is represented as a random variable that could be *relevant* or *irrelevant*. The energy function proposed for the MRF combines two factors: the similarity between the images in the list (*internal* similarity); and information obtained from the original order and the similarity of each image with the query (*external* similarity). Experiments were conducted with resources from the Image CLEF 2008 forum for the photo retrieval track, taking into account textual and visual features. The results show that the proposed method improves, according to the MAP measure, the order of the original list up to 63% (in the textual case) and up to 55% (in the visual case); and suggest future work using a combination of both kind of features.
Keywords: Image Re-ranking, Image Retrieval, Markov Random Field, Relevance Feedback.

Resumen. En este trabajo proponemos un método novedoso para re-ordenar una lista de imágenes recuperadas por un sistema de recuperación de imágenes (SRI). El método combina el orden original obtenido por el SRI, la similitud entre imágenes, obtenida con las características visuales y textuales, y un enfoque de retroalimentación de relevancia, todos ellos con el propósito de separar las imágenes relevantes de las irrelevantes, y así, obtener un orden más apropiado. El método está basado en el modelo de un campo aleatorio de Markov (CAM), en el que cada imagen en la lista fue representada como una variable aleatoria con dos posibles valores: relevante o irrelevante. La función de energía propuesta para el campo aleatorio de Markov combina dos factores: la similitud entre imágenes en la lista (similitud *interna*); y la información obtenida del orden original y la similitud de cada imagen con la consulta (similitud *externa*). Los experimentos fueron realizados con los recursos del foro Image CLEF 2008 para la tarea de recuperación de fotografías, tomando en cuenta los atributos textuales y visuales. Los resultados mostraron que el método propuesto mejora, de acuerdo con la medida MAP, el orden de la lista original hasta en un 63% (en el caso textual) y hasta un 55% (en el caso visual); y sugieren como trabajo a futuro el utilizar una combinación de ambos tipos de atributos.
Palabras clave: Re-ordenamiento de Imágenes, Recuperación de Imágenes, Campos Aleatorios de Markov, Retroalimentación de Relevancia.

1 Introduction

An image retrieval system (IRS) receives a query from a user, as keywords or sample images, and it is expected to return an ordered list of images that satisfies the user's request. Ideally, the IRS should return a list with all relevant images (Datta et al. 2008) ordered according to the user's request, so that the top images in the list are the ones *closer* to the user's expectations. The proper order in the list is important because it is easier and faster for the user to find images relevant to the query (Deselaers et al. 2008; Cui et al. 2008; Marakakis et al. 2008).

Current IRS, in general, tend to include several relevant images in the retrieved list. However, the images are not ordered properly, so there could be relevant images that are at the bottom of the list, and

many of the top images are not relevant. That is, IRS have a relatively good performance in terms of *recall*, but poor in terms of *precision*, in particular precision in the first 5, 10 and 20 images. Where *precision* indicates the percentage of relevant retrieved items, and *recall* the percentage of retrieved relevant elements from the total relevant items (Mani, 2001). Therefore an alternative to improve image retrieval is to re-order the list of images, so that the relevant ones are in the top of the list.

One way to improve the order of the results of an IRS is to use *relevance feedback*. That is, once the results are obtained by the IRS, a subset of relevant images is selected manually (by the user) or automatically. This subset is used to further refine the list obtained (Deselaers et al. 2008; Marakakis et al. 2008; Cui et al. 2008).

There are several approaches for relevance feedback. Some attempt to enrich the query to perform a new retrieval and obtain better results (Datta et al. 2008). This approach tends to be computationally expensive because it must re-use retrieval mechanisms to obtain the relevant images from the entire collection. This motivates to use only the retrieved list and reorder the images on the assumption that this list has relevant images, but not necessarily in the first positions.

Previous work (Lin et al. 2003; Cui et al. 2008; Marakakis et al. 2008) that tries to improve the order of a list of retrieved images do not use all the information available. Some just use a relevance feedback approach to locate images similar to the selected images as feedback. Other methods generate models that depend on the collection of images, assuming a certain number of search intentions. Some methods omit the use of information provided by the IRS, for example the original order. We consider that all the available information −the original order, the subset obtained via relevance feedback, the original query, and the entire list of retrieved images− is useful to improve the list order, and propose a re-ranking method that combines all this information.

This paper proposes a method that combines the original order obtained by a IRS, the similarity between images obtained with textual and visual features and a relevance feedback approach, all of them with the purpose of separating the relevant images from does that are not relevant, and thus obtain a more appropriate order for the results generated by the base IRS. The method is based on a Markov random field (MRF) model, in which each image in the list is represented as a random variable that could be *relevant* or *irrelevant*. The relevance feedback is incorporated in the initialization of the model, making these images *relevant*. The energy function of the MRF combines two factors: the similarity between the images in the list (*internal* similarity); and information obtained from the original order and the similarity of each image with the query (*external* similarity). Taking these factors into account and assigning a weight to each, the MRF is solved (obtaining the more probable configuration) separating the relevant images from the rest. Based on this result, the list of images is reordered placing in the top positions the images selected as feedback and the images marked as relevant by the MRF, and at bottom the rest of images.

In this paper we consider an image collection that includes, for each query, three sample images and a textual description. Each image in the collection has an associated text. We consider for measuring similarity the textual and visual components.

Experiments were conducted using the resources of the forum ImageCLEF 2008 for the photo retrieval track (Arni et al. 2008). We used one of the IRS that participated in this forum as the base retriever to obtain the initial lists, and tested our method with each one of the 39 queries used in this competition. Each of the results obtained by our method improved the original list order; according to the MAP measure, the improvements obtained are up to 63% using textual features and 55% using visual features.

The rest of the document is organized as follows. Section 2 provides a brief review of related work. Section 3 describes the proposed method. Section 4 describes the textual and visual features used. Section 5 presents the experiments and the results. Finally, Section 6 gives the conclusions and suggests future work.

2 Related Work

The task of image retrieval consists of, given a user query, retrieve all the relevant images from an image collection. The query can be visual (one or more sample images), textual (a set of keyword or sentences) or a combination of both. The IRS returns a list of *relevant images* that should be ordered according to the user's request.

An IRS may use visual or textual features, or a combination of them, to index and retrieve images. The results obtained by the IRS depend largely on the features used to describe the images, so that most of the time the order of the results are not appropriate for all the queries. Some methods try to improve the order obtained by enriching the query and searching the image again, but because they have to rerun retrieval mechanisms this solution is computationally expensive.

In general, the lists returned by IRS have several relevant images, but these are not placed in the top of the list. As mentioned before, it is important that the relevant images are in the first positions in the retrieved list. One way to improve the order of the results of an IRS is to use *relevance feedback* (Clough & Sanderson, 2004). That is, once the results are obtained by the IRS, a subset of relevant images is selected and used to further refine the given list.

There are different approaches for image re-ranking that can be classified as follows:

1. According to the features used to describe the image content:
   a) Visual. Use visual image features such as color, texture, shape, SIFT (*Scale Invariant Feature Transform*), SURF (*Speeded Up Robust Features*) among others to describe the visual content of each image (Lowe, 2004; Cui et al. 2008; Bay et al. 2006; Berg, 2009; Jianjiang et al. 2007; Jégou et al. 2010). It is still an unsolved problem to find the visual features needed to fully represent the visual content of an image.
   b) Textual. Use descriptive text to represent the image content (Datta et al. 2008; Lin et al. 2003; Chong et al. 2009; Choochaiwattana & Spring, 2009). It is expected that the textual description of the image must be both specific and general to cover all search intentions.
   c) Multimodal. Combine visual and textual features to describe image content, trying to cover more search intentions (Awang Iskanda et al. 2006; Gong & Liu, 2009; Richter et al. 2010). An open problem is how to combine both types of features to take advantage of the selected features.
2. Based on the type of relevance feedback used:
   a) Manual. Ask a user to select some of the retrieved images that he considers relevant to his search intention (Datta et al. 2008; Jianjiang et al. 2007; Choochaiwattana & Spring, 2009). The greater the number of selected images, the IRS will have more information about the images relevant to the user, but it also implies more additional time and effort for the user.
   b) Automatic. Make assumptions about the user's search intention to automatically select some sample images (Cui et al. 2008; Marakakis et al. 2008; Chong et al. 2009; Bihong et al. 2007). It is a difficult task making assumptions that cover all search intentions. This approach reduces the user's effort.
3. According to the technique used to sort images:
   a) Classifiers.Select some relevant and irrelevant images to build a classifier and label the remaining images as relevant or irrelevant (Martin et al. 2004; Deselaers et al. 2008). These methods compare images from the collection with those selected as relevant (positive examples) without taking into account other external information or images labeled as relevant by the classifier itself.
   b) Similarity functions. Use similarity functions to determine the images that are more related with pre-established categories (Bihong et al. 2007; Cui et al. 2008; Choochaiwattana & Spring, 2009). These methods depend on the type of images and the number of categories in the collection.
   c) Probabilistic models. Use probabilistic models to order the retrieved images, Gaussian mixtures models to combine features, graph based models to describe the relation between images or spatial configuration models to improve the object retrieval precision (Ke et al. 2008; Lin et al. 2003; Richter et al. 2010; Gong & Liu, 2009). These methods need to process the entire image collection to get a more accurate model. If the images collection is changing constantly, is necessary to rebuild the model.
   d) List fusion. Merge result lists from several IRSs for the same query, in order to get a list to improve the order of any of the base lists taken into account (Escalante et al. 2008). Choose what lists and how merged these is the main problem of this approach.

Previous methods do not take into account all the available information to produce an adequate order for the retrieved list of images. In contrast, the method proposed in this paper, combines the original order obtained by an IRS, the similarity

between images based on textual or visual features, and a manual relevance feedback approach; all of them with the purpose of separating the relevant images from those that are not relevant, and thus obtaining a more appropriate order than that generated by the base IRS.

The proposed method does not require accessing again the entire collection, considering only the list provided by IRS. The method is described in detail in the following sections.

3 Proposed Method

A general outline of the proposed method is given in Fig. 1. Given a query, the IRS retrieves from a given collection of images (that includes text captions) a list of files sorted according to a relevance criteria. From this list, some relevant images are selected based on a relevance feedback approach. For each image in the list, textual and visual features are extracted. The textual and visual feature description of each image in the list, the query given by the user, and a subset of images selected via relevance feedback, are combined to produce a re-ordered list. This re-ranking is obtained based on a Markov random field (MRF) model that separates the relevant images from irrelevant ones, generating a new list by positioning the relevant images first, and the others after. Next we give a brief review of MRFs, and then we describe in detail each component of the proposed method.

3.1 Markov Random Fields

Markov Random Fields (Li, 2004) are probabilistic models which combine *a priori* knowledge given by some observations and knowledge given by the interaction with neighbors.

Let $F = \{F_1, F_2, \ldots, F_n\}$ be random variables on a set *S*, where each F_i can take a value f_i in a set of labels *L*. This *F* is called a random field, and the instantiation of each of these $F_i \in F$ as an f_i is what is called a configuration of *F*, so, the probability that a random variable F_i takes the value f_i is denoted by $P(f_i)$, and the joint probability is denoted as $P(F_1 = f_1, F_2 = f_2, \ldots, F_n = f_n)$. A random field is said to be an MRF if it has the property of *locality*, i.e., if the field satisfies the following property:

$$P\left(f_i \middle| f_{\{S-\{i\}}\right) = P(f_i | f_{N_i}) \quad (1)$$

where $S - \{i\}$ represents the set *S* without the i^{th} element, $f_{N_i} = \{f'_i | i' \in N_i\}$, and N_i is the set of neighboring nodes of the node f_i. The joint probability can be expressed as:

$$P(f) = \frac{e^{-U_p(f)}}{Z} \quad (2)$$

where *Z* is called the partition function or normalizing constant, and $U_p(f)$ is called the energy function.

The optimal configuration is found by minimizing the energy function $U_p(f)$, obtaining a value for every random variable in *F*.

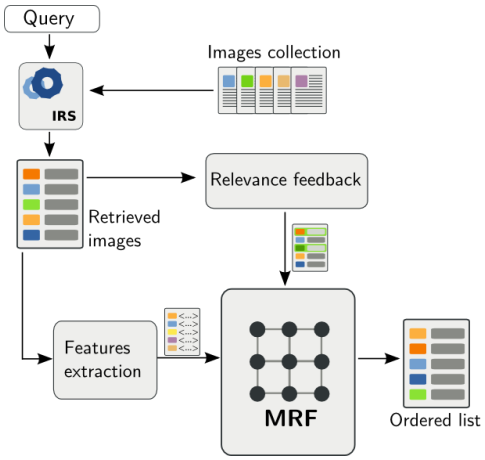


Fig. 1. Block diagram of the proposed method

3.2 Model

In our case we consider a MRF in which each node corresponds to a document (image + text caption) in the list. Each document is represented as a random variable with 2 possible values: *relevant* and *irrelevant*. We consider a fully connected graph, such that each node (document) is connected to all other nodes in the field; that is, we defined a neighborhood scheme in which each variable is adjacent to all the others. Given that the number of images in the list is relatively low (100 in the experiments), to consider a complete graph is not a problem computationally, and allows us to consider the relations between all documents in the list.

For representing the images we consider visual features from the image or textual features from the caption. To describe the images, in the textual case, we used a binary bag of words representation, in which each vector element represents a textual word

from the collection vocabulary; and the query is represented in the same manner. In visual case, the retrieved images, like the query images, are described by SIFT (Scale-Invariant Feature Transform) features. The internal and external similarities are considered via the energy function described next.

3.3 Energy Function

The energy function of the MRF combines two factors: the similarity between the images in the list (*internal* similarity); and external information obtained from the original order and the similarity of each image with the query (*external* similarity).

The internal similarities correspond to the interaction potentials and the external similarities to the observation potentials.

The proposed energy function takes into account both aspects and is defined as follows:

$$U(I_i) = V_c(I_i) + \lambda V_a(I_i) \tag{3}$$

Where $V_c(I_i)$ is the interaction potential and it considers the similarity between random variable I_i and its neighbors, representing the support that neighboring variables give to I_i. $V_a(I_i)$ is the observation potential and represents the influence of external information on variable I_i. The weight factor λ favors V_c ($\lambda < 1$), V_a ($\lambda > 1$), or both ($\lambda = 1$).

V_c is defined as:

$$V_c(I_i) = \begin{cases} \sum_j^m sim(I_i, I_j) + (1 - \sum_j^n sim(I_i, I_j)) \\ \quad if\ I_i = irrelevant \\ \sum_j^n sim(I_i, I_j) + (1 - \sum_j^m sim(I_i, I_j)) \\ \quad if\ I_i = relevant \end{cases} \tag{4}$$

Where $\sum_j^m sim(I_i, I_j)$ represents the average similarity between variable I_i and its neighbors with irrelevant value. $\sum_j^n sim(I_i, I_j)$ represents the average similarity between variable I_i and its neighbors with relevant value. Where *n*+*m* is the total of images in the original list. V_a is defined as follows:

$$V_a(I_i) = \begin{cases} (1 - simq(I_i, q)) \times g(posinv(I_i)) \\ \quad if\ I_i = irrelevant \\ simq(I_i, q) \times g(pos(I_i)) \\ \quad if\ I_i = relevant \end{cases} \tag{5}$$

The V_a potential is obtained by combing two factors. The first indicates how similar, $simq(I_i, q)$, or different, $1 - simq(I_i, q)$ is the I_i variable with the query *q*. The second is a function $g(x)$ that converts the position in the list given by a base IRS to a real value. The function $pos(I_i)$ returns the position of the image I_i in the original list, $posinv(I_i)$ returns the inverse position of the I_i variable in this list.

The initial configuration of the MRF is obtained by relevance feedback. That is, the subset of images selected via relevance feedback are initialized as relevant, and all other images as irrelevant. Then, the MRF configuration of minimum energy (MAP) is obtained via stochastic simulation using the ICM algorithm. We experimented using also Simulated Annealing with similar results (Chellappa, 1993). At the end of this optimization process, each variable (image) has a value of relevant or irrelevant. Based on these values, a new re-ordered list is produced, by positioning first the relevant images according to the MRF, and then the irrelevant ones.

4 Features Extraction

Each image is represented by its textual or visual features; for each of them a word based representation is used: textual words and visual words.

In the case of textual features we used the words of the textual description of the image, representing each image by a word vector. Each element of the vector indicates the absence (0) or presence (1) value of a textual word in the image description. To obtain the textual description of each image we follow the next steps:

1. Stopword removal. Stopwords were removed from the description of the retrieved images and the query.
2. Vocabulary extraction. We obtained the vocabulary from all textual descriptions of images and query.
3. Vector construction. We identified the occurrence or absence of the vocabulary words in the description of each of the images and the query and the vector was constructed for each image.

In the case of visual features, SIFT features were used to represent objects within the image. These features are taken as visual words and, as in the case of textual features, we built a representation with these words. To find the SIFT features in an

image we performed the following steps (Lowe, 2004):

1. Scale-space extrema detection. The first stage of computation searches over all scales and image locations.
2. Keypoint localization. At each candidate location, a detailed model is fit to determine location and scale.
3. Orientation assignment. One or more orientations are assigned to each keypoint location based on local image gradient directions.
4. Keypoint descriptor. The local image gradients are measured at the selected scale in the region around each keypoint.

The SIFT features are invariant to image scale and rotation, change in 3D viewpoint, additive noise, and change in illumination. To calculate the SIFT points and common points between images we used the implementations proposed in (Lowe, 2004).

Fig. 2 shows the textual words selected from the descriptive fields and the SIFT keypoints of an image in the collection. Only the textual words showed in the Fig. 2 have the value 1 in the textual vector from this image. The SIFT keypoints are specified by 4 floating point numbers giving subpixel row and column location, scale, and orientation, the invariant descriptor vector for the keypoint is given as a 128 integers in the range [0,255] (Lowe, 2004).

5 Experimental Evaluation

We conducted a series of experiments with the following objectives: (i) to test the results of the proposed method compared with the original list, (ii) to asses the level of improvement when using textual or visual features to measure similarity, (iii) to evaluate the sensitivity of the method to the model parameters and (iv) to compare the improvement between the use of visual or textual features.

5.1 Experimental setup

To perform the experiments we used the resources of the forum Image CLEF 2008, which consists of the image collection IAPR TC-12 (Arni et al. 2008), a set of queries for the photo retrieval track and a list of results from one of the participants (TIA-TXTIMG). This collection was chosen because it has relevance judgments for each query, this allowed a comparison with results obtained by the proposed method; it also includes for each image, a textual description of its visual content.

< machu, picchu, rear, view, llama, terraces, ruins, bald, mountain,
range, clouds, background >

Fig. 2. SIFT *keypoints* and textual words obtained after feature extraction of an image in the collection

<title> church with more than two towers</title>
<narr> Relevant images will show a church, cathedral or a mosque with three or more towers.</narr>

Fig. 3. Query example for the photo retrieval track from the Image CLEF 2008 forum. Each query includes, among other fields, a field *title* summarizing the query objective, and a field *narrative* specifying textually which images are relevant to the query

The TIA-TXTIMG SRI retrieves a list of images by combining the results of several retrieval methods (Escalante et al. 2008). We considered the best results obtained by this group as the input for our method.

The proposed method was tested with the 39 queries from Image CLEF 2008 forum. These queries have 3 sample images and a textual description that includes a narrative about the images relevant to the query. For experiments using only textual information, the sample images were not considered. For experiments using visual information the proposed method used only the three sample images for each query. Fig. 3 shows a sample query for the photo retrieval track.

Each of the images in the IAPR TC-12 collection has assigned a set of descriptive fields, we included the words in the title and in the textual description to represent the images. Fig. 4 shows an example of an image of the IAPR TC-12 collection and its descriptive fields.

As input data for the proposed method, we selected the first 100 images retrieved by TIA-TXTIMG IRS for each of the 39 queries.

For evaluation we used MAP and Precision at the first *N* retrieved items (Mani, 2001). MAP is defined as follows:

$$MAP = \frac{1}{Q}\sum_{i=1}^{Q}\left[\frac{\sum_{r=1}^{m} P_i(r) \times rel_i(r)}{n}\right] \quad (6)$$

Where $P_i(r)$ is the precision at the first *r* documents, $rel_i(r)$ is a binary function which indicates if document at position *r* is relevant or not for the query *i*; *n* is the total number of relevant documents for the query *i*, *m* is the number of relevant documents retrieved and *Q* is the set of all queries. Precision at *N* is defined as the percentage of retrieved relevant items at the first *N* positions of the result list.

Experiments were conducted using only textual or visual features. For this, similarity functions were defined for the potentials V_c and V_a.

For the similarity based on textual features, the following similarity functions were proposed:

- The similarity function used to measure the similarity between variables is defined as: $sim(I_i, I_j)$, where:

$$sim(I_i, I_j) = 1 - \frac{2\,|I_i \cap I_j|}{|I_i \cup I_j|} \quad (7)$$

- The similarity function between an image I_i and the query *q* was defined as $simq(I_j, q)$, where:

$$simq(I_j, q) = 1 - \frac{|I_i \cap q|}{|q|} \quad (8)$$

For the similarity based on visual features, the following similarity functions were proposed:

- The similarity between neighboring variables was measured using $sim(I_i, I_j)$, where I_i and I_j are variables represented by its SIFT features. The similarity function $sim(I_i, I_j)$ is defined as:

$$sim(I_i, I_j) = 1 - \frac{match(I_i, I_j)}{num(I_i) + num(I_j)} \quad (9)$$

- The function $match(I_i, I_j)$ finds the common number of SIFT features between the variable I_i and I_j. Function $num(I_i)$ calculates the number of SIFTS features found for I_i image.
- Because the query is composed by 3 example images, the similarity function between an image I_i and query *Q* was defined as:

$$simq(I_i, Q) = 1 - \max\left(\frac{match(I_i, q_1)}{num(q_1)}, \frac{match(I_i, q_2)}{num(q_2)}, \frac{match(I_i, q_3)}{num(q_3)}\right) \quad (10)$$

```
<TITLE>The Plaza de Armas</TITLE>
<DESCRIPTION>a yellow building
with white columns in the
background; two palm trees in
front of the house; cars parkedin
front of the house; a woman and a
child are walking over the
square</DESCRIPTION>
```

Fig. 4. Example of an image from the IAPR TC-12 collection and its set of descriptive fields: *title* and *description*

The features for textual and visual similarity functions can be defined based on 2 general operations. The first is the intersection of either textual or visual words, and determines the words in

common between two images. The second is the coverage, and determines how many words in an image are contained in the other.

The function used to map the order from the original list is $g(x) = {e^{x/20}}/{e^5}$, the intuitive idea of this function is such that it first increases slowly so that the top images have a small potential, and then it increases exponentially to amplify the potential for those images in the bottom of the list.

5.2 Experimental results

Five experiments were conducted for each type of feature, visual or textual, varying λ, see Table 1. Each of the 5 experiments were made taking into account 1, 3, 5, 8 and 10 images as relevance feedback.

A simulated user feedback technique was used to perform the experiments. The collection used contains, in addition to the queries and images, relevance judgments indicating which images are relevant to each of the proposed queries, given that it is known beforehand which images are relevant in the retrieved list, hence some of this images are taken as feedback. This type of feedback is known as simulated user feedback.

When the MRF converges, the images selected by the relevance feedback are placed in the top of the new list, then are placed those with *relevant* value respecting the order in the original list. Images with *irrelevant* value are placed after placing all the images with *relevant* value.

Table 1. Meaning of each experiment conducted to evaluate the proposed method based on the value of λ

Value of λ	*Description*
λ = 1.5	More importance to Va
λ = 1	Equal importance to Vc and Va
λ = 0.5,0.3	More importance to Vc
λ = 0	Cancel the contribution of Va
λ = ∞	Cancel the contribution of Vc

A comparison between the results of the original list retrieved by the TIA-TXTIMG IRS and the results obtained with the proposed method (using visual or textual features) for the different configurations of parameters and features are shown in Fig. 5, where the results indicated the average of the MAP values obtained for the 39 queries. The graph shows that better values are obtained with $\lambda = 0.3$ consistently for different number of images selected as feedback.

Note that all variants of the proposed method showed in Fig. 5 improved the results of the original list.

When the value of λ is small (eg. $\lambda = 0.3$) the proposed method yields the best results. So it seems that, at least for this collection, the information from the neighbors is more valuable than the information from the original order and the similarity with the query.

Table 2 shows a comparison between the best results obtained by the proposed method and the results obtained by the TIA-TXTIMG IRS. This table also shows the values for precision measure at the first 5, 10 and 20 retrieved images, where the proposed method also overcomes the TIA-TXTIMG IRS.

Table 2. A comparison between the best results obtained by some variants of the proposed method and the results obtained by the TIA-TXTIMG IRS. The number after the letter *F* indicates the number of images taken for relevance feedback, the number following the letter *L* indicates the value of λ. P5, P10 and P20 indicate the precision to 5, 10 and 20 retrieved images respectively. MAP is the *Mean Average Precision*

	Experiment	*P5*	*P10*	*P20*	*MAP*
Baseline	TIA TXT-IMG	0.4769	0.4538	0.3910	0.2359
Textual features	F1-L0.3	0.6103	0.5410	0.4833	0.2902
	F3-L0.3	0.7846	0.6128	0.4962	0.3070
	F5-L0.3	1.0000	0.7154	0.5474	0.3358
	F8-L0.3	1.0000	0.8821	0.6218	0.3706
	F10-L0.0	1.0000	0.9744	0.6551	0.3858
Visual features	F1-L1.0	0.6667	0.5282	0.4090	0.2584
	F3-L0.3	0.8410	0.6179	0.4423	0.2795
	F5-L0.3	1.0000	0.7051	0.4846	0.3019
	F8-L0.3	1.0000	0.8795	0.5679	0.3379
	F10-L0.5	1.0000	0.9744	0.6205	0.3580

The results show that for simulated user feedback and considering only textual features to measure the similarity of images, an improvement of up to 63% in the MAP is obtained selecting 10 images as feedback, and an improvement of 23% in the MAP when only one image is selected as feedback. As more images are given as feedback, the performance improves.

The results for visual features also show that proposed method improved by to 55% the MAP the

original list when 10 images were selected as feedback, and by 9% when only one image was used as feedback.

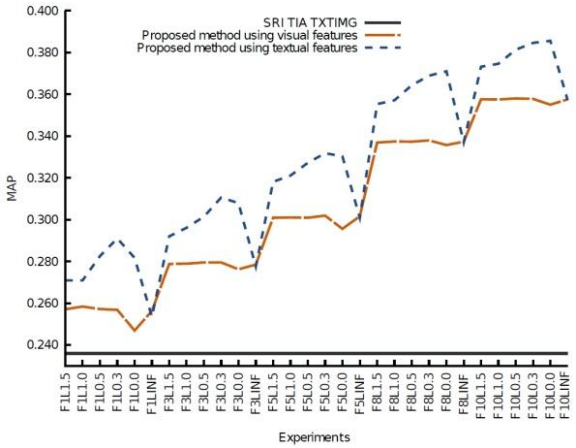


Fig. 5. Comparison between the results obtained by the TIA-TXTIMG IRS and the results obtained by the proposed method, using visual or textual features, for simulated user feedback varying the value of λ and the number of selected images as feedback (the bottom line shows the MAP for the original list)

Fig. 6 shows the first 20 images from the original list obtained by the IRS for the query: *straight road in the USA* as well as the list obtained after the application of the proposed method. In this example the proposed method, using textual or visual features, placed in the top of the list some images that were not in the original list, in fact all images placed in the first 10 positions are relevant. For this query, textual features were sufficient to properly order (with the proposed method) the recovered images.

Experiments showed cases where the textual description is insufficient to locate relevant images. Fig. 7 shows the 20 first images obtained by the IRS and the first 20 obtained by our method, using textual features only, for the query: *church with more than two towers*. For this example, few of the images at the top are relevant, because in its textual description is not mentioned information about the number of towers in the building, which is an important factor that determines the relevance of the images to this query. In addition, some relevant images that were not selected by the method using textual features only, were identified using visual features. These results motivate the use of a combination of features that allows to use the advantage of both.

To show that the results obtained by the proposed method are significant with respect to the results from the base IRS, we used the *paired t-student* test, and showed that using a confidence level $\alpha = 0.01$ the results are statistically significant (Kanji, 1993; Dietterich, 1998).

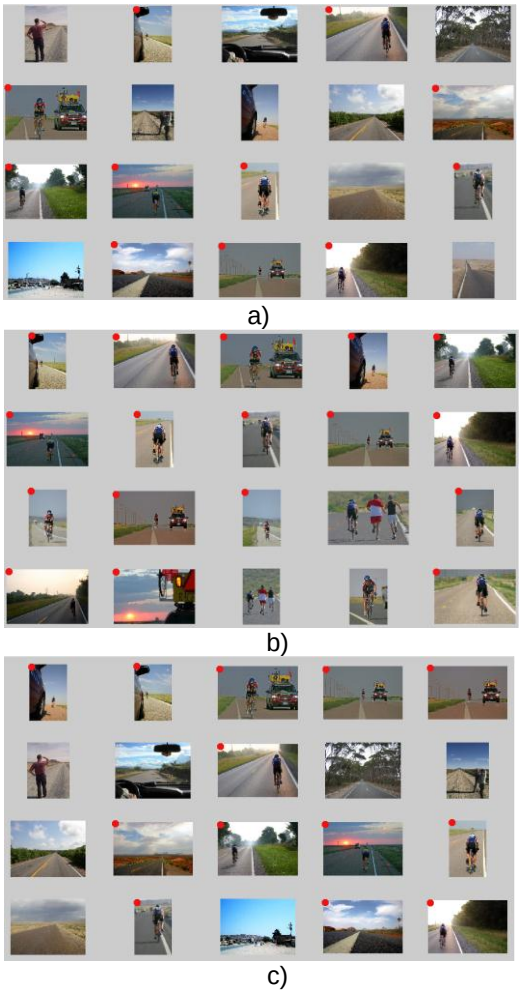


Fig. 6. First 20 images in the list obtained by the IRS (a) and the list sorted by our method using textual (b) or visual (c) features for the query *straight road in the USA*. The relevant images to the query are indicated with a red dot in the upper left corner

6 Conclusions and Future Work

This paper proposed a method for improving the order of a list of images retrieved by an IRS. Based on relevance feedback, the proposed method integrates, via a MRF, to separate the relevant and irrelevant images from the original list: the similarity between the images in the list (*internal* similarity); and information obtained from the original order and the query (*external* similarity).

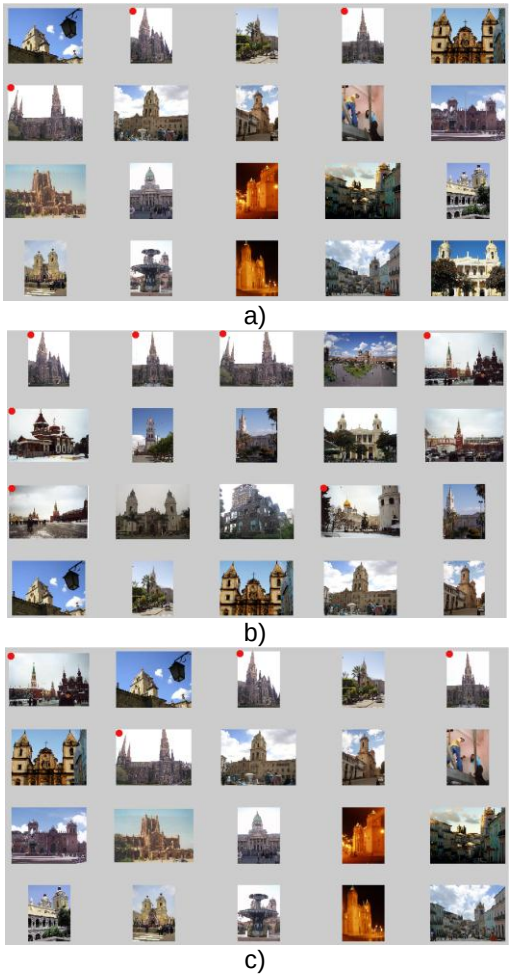


Fig. 7. First 20 images in the list obtained by the IRS (a) and the list sorted by the MRF using textual (b) or visual (c) features for query *church with more than two towers*. The relevant images to the query are indicated with a red dot in the upper left corner

Also we proposed similarity measures based on textual and visual features that allowed to differentiate relevant and irrelevant images. From the results of the experiments we identified some cases in which a feature type provides more information than the other.

Experiments were conducted using the resources of the forum ImageCLEF 2008 for the photo retrieval track. The results showed that, in the best case, the proposed method improved the MAP up to 63% compared with the original list selecting 10 images as feedback, and 23% selecting only one image as feedback when using textual features, and up to 55 % selecting 10 images as feedback and 9% when using one image as feedback using visual features only. These differences are statistically significant according to the *paired t-student* test at a $\alpha = 0.01$ confidence level.

The best results are obtained by giving greater importance to information from neighbors, obtained from the textual or visual similarity between images. We are currently experimenting with other IRS to analyze the level of improvement of the proposed method.

In summary, the main contributions of this work are: (i) a novel image re-ranking method based on a MRF that integrates a relevance feedback approach, the similarity between the images in the list and external information obtained from the original order and the query, (ii) a common representation for visual and textual features and the corresponding similarity measures.

For some queries the textual or visual description does not provide relevant information about the visual content of the image, therefore it proposed as future work to include a combination of textual and visual features to exploit the advantages of both.

References

1. **Arni, T., Clough, P., Sanderson, M., & Grubinger, M. (2008)**. Overview of the ImageCLEFphoto 2008 Photographic Retrieval Task. *9th Cross-language evaluation forum conference on Evaluating systems for multilingual and multimodal information access (CLEF 2008),* Aarhus, Denmark, 500-511.
2. **Awang, D. N. F., Pehcevski, J., Thom, J. A., & Tahaghoghi, S. M.M. (2006)**. Combining image and structured text retrieval. Advances in XML Information Retrieval and Evaluation. *Lecture Notes in Computer Science,* 3977, 525-539.
3. **Bay, H., Tuytelaars, T., & Van Gool, L. (2006)**. Surf: Speeded up robust features. *9th European Conference on Computer Vision ,* Graz, Austria, 404-417.
4. **Berg, T. L. (2009)**. Finding Iconic Images. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops,* Miami, FL, USA, *2009*, 1-8.

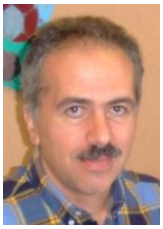
5. **Bihong, G., Bo, P., & Xiaoming, L. (2007)**. A Personalized Re-ranking Algorithm Based on Relevance Feedback. *Advances in Web and Network Technologies, and Information Management. Lecture Notes in Computer Science,* 4537, 255-263.
6. **Chellappa, R., & Jain, A. (1993)**. *Markov Random Fields: Theory and Application.* Boston: Academic Press.
7. **Chong, T., Yanxiang, H., Donghong, J., Guimin, L., & Zhewei, M. (2009)**. A Study on Pseudo Labeled Document Constructed for Document Re-ranking. *International Conference on Artificial Intelligence and Computational Intelligence*, Shanghai, China, 377-380.
8. **Choochaiwattana, W., & Spring, M. B. (2009)**. Applying Social Annotations to Retrieve and Re-rank Web Resources. *International Conference on Information Management and Engineering*, Kuala Lumpur, Malaysia, 215-219.
9. **Clough, P., & Sanderson, M. (2004)**. Relevance Feedback for Cross Language Image Retrieval. *Advances in Information Retrieval*: *26th European Conference on IR Research, ECIR 2004*. *Lecture Notes in Computer Science,* 2997, 238-252.
10. **Cui, J., Wen, F., & Tang, X. (2008)**. Real time google and live image search re-ranking. *16th ACM international conference on multimedia MM '08,* Vancouver, Canada, 729-732.
11. **Datta, R., Joshi, D., Li, J., & Wang, J. Z. (2008)**. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computer Surveys , 40 (2)*, 1-60.
12. **Deselaers, T., Paredes, R., Vidal E., & V, E. (2008)**. Learning Weighted Distances for Relevance Feedback in Image Retrieval. 19th *International Conference on Pattern Recognition, Tampa, Florida, USA,* 1-4.
13. **Dietterich, T. G. (1998)**. Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. *Neural Computation*, 10 (7), 1895-1923.
14. **Escalante, H. J., Gonzalez, J. A., Hernandez, C., Lopez, A., Montes, M., Morales, E., et al. (2008)**. TIA-INAOE's Participation at ImageCLEF2008. *Working Notes of the CLEF 2008 Workshop*, Aarhus, Denmark.
15. **Escalante, H. J., Hernandez, C. A., Sucar, L. E., & Montes, M. (2008)**. Late fusion of heterogeneous methods for multimedia image Retrieval. *1st ACM international conference on Multimedia information retrieval (MIR '08)*, Vancouver, Canada, 172-179.
16. **Gong, Z., & Liu, Q. (2009)**. Improving keyword based web image search with visual feature distribution and term expansion. *Knowledge and Information Systems, 21 (1)*, 113-132.
17. **Jégou, H., Douze, M., & Schmid, C. (2010)**. Improving bag-of-features for large scale image search. *International Journal of Computer Vision,* 87 (3) , 316-336.
18. **Jianjiang, L., Zhenghui, X., Ran, L., Yafei, Z., & Jiabao, W. (2009)**. A Framework of CBIR System Based on Relevance Feedback,. *Third International Symposium on Intelligent Information Technology Application (IITA 2009),* 175-178.
19. **Kanji, G. K. (1993)**. *100 statistical tests.* Newbury Park, California: Sage Publications.
20. **Gao, K., Lin, S., Zhang, Y., & Tang, S. (2008)**. Object-based Image Retrieval with Attention Analysis and Spatial Re-ranking. *IFIP Advances in Information and Communication Technology*, 288, 118-128.
21. **Li, S.Z. (1994)**. Markov random field models in computer vision. *Computer Vision ECCV '94, Lecture Notes in Computer science*, 801, 361-370.
22. **Lin, W. H., Jin, R., & Hauptmann, A. (2003)**. Web Image Retrieval Re-Ranking with Relevance Model. *IEEE/WIC International Conference on Web Intelligence (WI '03)*, Halifax, Canada, 242-248
23. **Lowe, D. G. (2004)**. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60 (2), 91-110.
24. **Mani, I. (2001)**. *Automatic Summarization .* Amsterdam ; Philadelphia: John Benjamins Publishing Co.
25. **Marakakis, A., Galatsanos, N., Likas, A., & Stafylopatis, A. (2008)**. Application of Relevance Feedback in Content Based Image Retrieval Using Gaussian Mixture Models*: 20th IEEE International Conference on Tools with Artificial Intelligence* (*ICTAI '08),* Dayton, Ohio, USA, 141-148.
26. **Martin, B., Dirk, P., & Josiah, P. (2004)**. Application of Machine Learning Techniques to the Re-ranking of Search Results *KI 2004: Advances in Artificial Intelligence*. *Lecture Notes in Computer Science, 3238*, 67-81
27. **Richter, F., Romberg, S., Horster, E., & Lienhart, R. (2010)**. Multimodal ranking for image search on community databases. *International conference on multimedia information retrieval (MIR '10)*, Philadelphia, Pennsylvania, USA, 63-72.

R. Omar Chávez

He is a Master in Computational Sciences student at the National Institute of Astrophysics, Optics and Electronics, located in Puebla, Mexico. He obtained his bachelor degree in Computer Engeneering from the Universidad Tecnológica de la Mixteca. His main research interests include probalistics models, computer vision, information retrieval and information re-ranking.

Manuel Montes y Gómez

Obtained his Ph.D. in Computer Science from the Computing Research Center of the National Polytechnic Institute of Mexico. Currently, he is a full-time lecturer at the Computational Sciences Department of the National Institute of Astrophysics, Optics and Electronics, located in Puebla, Mexico. His research is on automatic text processing. He is author of more than 100 international papers in the fields of information retrieval, question answering, information extraction and text mining. Dr. Montes has been visiting professor at the Polytechnic University of Valencia (Spain), at the University of Geneva (Italy), and from August 2010 he is visiting professor at the University of Alabama (USA). In addition, Dr. Montes is member of the Mexican National System of Researchers, the Mexican Society of Artificial Intelligence, the Mexican Association for Natural Language Processing and the International Web Intelligence Consortium.



L. Enrique Sucar

Has a Ph.D in computing from Imperial College, London, UK, 1992; a M.Sc. in electrical engineering from Stanford University, California, USA, 1982; and a B.Sc. in electronics and communications engineering from ITESM, Monterrey, México, 1980. He is currently a Senior Researcher at INAOE, Puebla, Mexico. Dr. Sucar is Member of the National Research System, the Mexican Science Academy, AAAI, SMIA and Senior Member of the IEEE. He has served as president of the Mexican AI Society, has been member of the Advisory Board of IJCAI, and is Associate Editor of the journals Computación y Sistemas and Revista Iberoamericana de Inteligencia Artificial. His main research interests are in graphical models and probabilistic reasoning, and their applications in computer vision, robotics and biomedicine.