Hindawi Publishing Corporation
ISRN Machine Vision
Volume 2013, Article ID 428746, 16 pages
http://dx.doi.org/10.1155/2013/428746

*Research Article*

# Multimodal Markov Random Field for Image Reranking Based on Relevance Feedback

**Ricardo Omar Chávez, Hugo Jair Escalante, Manuel Montes-y-Gómez, and Luis Enrique Sucar**

*Department of Computer Sciences, Instituto Nacional de Astrofśsica, Óptica y Electrónica, Luis Enrique Erro No. 1, 72840 Tonantzintla, PUE, Mexico*

Correspondence should be addressed to Hugo Jair Escalante; hugo.jair@gmail.com

Received 4 December 2012; Accepted 30 December 2012

Academic Editors: H. Erdogan, N. Grammalidis, N. D. A. Mascarenhas, and W. L. Woo



This paper introduces a multimodal approach for reranking of image retrieval results based on relevance feedback. We consider the problem of reordering the ranked list of images returned by an image retrieval system, in such a way that relevant images to a query are moved to the first positions of the list. We propose a Markov random field (MRF) model that aims at classifying the images in the initial retrieval-result list as relevant or irrelevant; the output of the MRF is used to generate a new list of ranked images. The MRF takes into account (1) the rank information provided by the initial retrieval system, (2) similarities among images in the list, and (3) relevance feedback information. Hence, the problem of image reranking is reduced to that of minimizing an energy function that represents a trade-off between image relevance and interimage similarity. The proposed MRF is a multimodal as it can take advantage of both visual and textual information by which images are described with. We report experimental results in the IAPR TC12 collection using visual and textual features to represent images. Experimental results show that our method is able to improve the ranking provided by the base retrieval system. Also, the multimodal MRF outperforms unimodal (i.e., either text-based or image-based) MRFs that we have developed in previous work. Furthermore, the proposed MRF outperforms baseline multimodal methods that combine information from unimodal MRFs.

## 1. Introduction

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 given rise to the existence of millions of digital images stored in many databases around the world. However, stored information is useless if we cannot access the specific data we are interested in. Thus, the development of effective methods for the organization and exploration of image collections is a crucial task [1–3].

In a standard image retrieval scenario one has available a collection of images and users want to access images stored in that collection, where images can be annotated (i.e., associated to a textual description). Images are represented by features extracted from them. Users formulate queries (which are associated to their information needs) by using either sample images, a textual description, or a combination of both. Queries are represented by features extracted from them and the retrieval process reduces to comparing the representations of documents in the collection to that of the queries. Images in the collection are sorted in descending order of similarity and are shown to users in response to their queries.

Image retrieval has been an active research area since more than two decades ago [1–6]. During that time, a wide variety of content-based (i.e., that use visual features derived from the image) [1], text-based (i.e., that use text associated to the image) [2], and multimodal [3, 7] (i.e., that combine visual and textual features) retrieval techniques have been proposed, which have proved to be effective in varied scenarios. Nevertheless, current retrieval methods still have problems for retrieving most of relevant images to a given query in the first positions. The latter is due to the fact that modeling user intention from queries is, in general,

a highly subjective and difficult task, hence, postprocessing and refinement strategies have been adopted [3, 4, 8–21].

Postretrieval techniques aim at refining retrieval results by feature reweighting, query modification, document reranking, and relevance feedback. The common idea is to interact with the user in order to learn or to improve a model of the underlying user's information need. Acceptable results have been obtained with such methods, however, they still have several limitations, including

(i) the need of extensive user interaction (One should note that, when available, user interaction should be included in postretrieval techniques as it is evident that information provided by the user is by far more reliable than that we would obtain with fully automatic approaches. Hence, we think that in general the goal of postprocessing methods should be the minimization of user interaction, instead of the development of fully automatic techniques.)

(ii) the multiple execution of retrieval models

(iii) the on-line construction of classification methods

(iv) the lack of contextual information in the postretrieval processing, which may be helpful for better modeling users' information needs

(v) the computational cost that involves processing the entire collection of documents for each feedback iteration and

(vi) the incapacity of methods to work with multimodal information.

This paper introduces an alternative postretrieval technique that aims at improving the results provided by an image retrieval system and that overcomes some of the limitations of current postretrieval methods. In particular, we face the problem of reranking the list of images as returned by an image retrieval system. This problem is motivated by the availability of retrieval systems that present high-recall and low-precision performance [7, 22], which evidences that the corresponding retrieval model is able to retrieve many relevant images but they are not placed in the right positions. Hence, given a list of ranked images, the problem we approach consists of moving relevant images to the first positions and displacing irrelevant ones to the final positions in the list.

We propose a solution to the reranking problem based on a multimodal Markov random field (MRF) that aims at classifying the ranked images as relevant or irrelevant. Each image in the list is associated to a binary random variable in the MRF (i.e., a node), and the value of each random variable indicates whether an image is considered relevant (when it takes the value 1) or not (when it takes the value 0). The MRF takes into account (1) the rank information provided by the base retrieval system, (2) similarities among images in the list, and (3) relevance feedback information. In this way, we reduce the problem of image reranking to that of minimizing an energy function that represents a trade-off between image relevance and interimage similarity.

Rank information provided by the retrieval system is the base of our method, which is further enriched with contextual (i.e., multimodal similarities among images) and relevance feedback information. The motivation for taking context into account is that relevant images to a query will be similar to each other and to the query, to some extent; whereas irrelevant images will be different among them. (One should note that irrelevant images will be somewhat similar to the query inevitably as that is why they were retrieved in the first place.) We consider relevance feedback as a seed generation mechanism for propagating the relevancy/irrelevancy status of nodes in the MRF. Our MRF is a multimodal as it can take advantage of both visual and textual information by which images are described. The proposed MRF does not require multiple executions of retrieval models nor training classification methods and it could work without user intervention. In consequence, our multimodal MRF overcomes the main limitations of current post-processing techniques, see Section 2.

We report experimental results in the IAPR TC12 collection [23] that show the validity of our formulation. This collection comprises 20, 000 images with manual annotations in three languages, and it is accompanied with sample queries and relevance judgements; this benchmark has been widely used for the evaluation of multimodal image retrieval systems [7, 22, 24, 25]. Experimental results show that our method is able to improve the ranking provided by a given retrieval system. Our multimodal MRF also outperforms unimodal (i.e., either text-based or image-based) MRFs that we have developed in previous work [26, 27]. Further, the proposed MRF outperforms baseline multimodal methods that combine information from unimodal MRFs. Our results motivate further research on the development of multimodal MRFs for related tasks, for example, for retrieval-result diversification [12, 25].

The contributions of this paper are as follows.

(i) We introduce a novel MRF model that incorporates multimodal information for image reranking. The proposed model is able to improve the ranking of the base retrieval system. The MRF relies on manual relevance feedback, thus it is a user adaptive technique, although it could work with automatic relevance feedback as well. Also, since our MRF works with a list of ranked images, it is not tied with a specific retrieval system nor with a particular architecture, image collection, or information modalities.

(ii) We propose an energy function for the MRF that incorporates information provided by the base retrieval system, relevance feedback, and interimage similarity. The energy function allows us to model the relationships among these sources of information. Also, the structure of the proposed MRF naturally allows us to take contextual information into account, which is often disregarded in usual post-retrieval techniques, although it proved very useful for our model.

(iii) We introduce baseline methods for multimodal image reranking based on ideas from late fusion and intermedia relevance feedback. The proposed formulations combine information from unimodal MRFs for image reranking and are able to improve the performance of the base retrieval system.

The rest of this paper is organized as follows. The next section reviews related work on image retrieval with emphasis on post-retrieval methods that incorporate relevance feedback. Section 3 presents background information that will be helpful for understanding the rest of the paper. Section 4 introduces the multimodal Markov random field for image reranking. Section 5 describes baseline techniques to which we compare our multimodal method. Section 6 describes the experimental setting we adopted and presents results on the IAPR TC12 collection. Section 7 presents the conclusions derived from this work and outlines future work directions.

## 2. Related Work

Given a database of images and a query formulated by a user, the goal of image retrieval systems is to return images in the database that are relevant to the query [1, 3]. The core of retrieval systems is the retrieval model which specifies how images/queries are represented, how they are compared, and how results are shown to users. However, because of the difficulty of the task and of the subjectivity of user intention modeling, retrieval models are often equipped with automatic or manual post-retrieval mechanisms that aim at refining and improving the outputs of retrieval systems [4, 10, 11].

*2.1. Relevance Feedback.* By far, the main post-retrieval technique used in image retrieval is relevance feedback (RF) and its variants thereof [3, 4, 7, 10, 11, 28–37]. RF aims to refine the retrieval results of an image retrieval system by taking advantage of information provided by the user. In each iteration of RF, the user indicates what images are relevant (or irrelevant) to her/his information need, then a specific criterion is adopted for modifying/adapting the original query with the goal of improving the preceding retrieval result.

RF was first introduced into image retrieval by Rui et al. [10] more than one decade ago, and nowadays it is a fundamental component of successful retrieval systems [3, 4, 7, 11, 24, 25, 37]. Usually, feedback information is used to modify the weights assigned to different features when computing similarity between queries and images in the database [10, 38, 39]. Alternatively, the distance between each document and the nearest relevant/irrelevant feedback image has been used to rerank the set of images in the collection [40–42]. Adapting the similarity function according to feedback images and then running again the retrieval model is another common approach in RF [32].

Other researchers have adopted an active learning scenario for RF. The system asks the user to indicate whether *informative* images are relevant/irrelevant to her/his information need [30, 31, 43], where the *informativeness* of images depends on the active learning criterion. For example, Tong and Chang [31] ask the user to provide feedback on images lying at the margin of a support vector machine classifier that attempts to classify images as relevant/irrelevant to a query. Zhou et al. [30] rely on a cotraining mechanism for modifying the distance function for image retrieval.

Other variants are those based on supervised learning and information fusion. The former RF information is used for building a classifier, where relevant images are considered the positive examples and irrelevant images are considered the negative examples of a binary classification task [32, 44] or of a one-class classification problem [45]. For example, Yan et al. build a support vector classifier using as positive examples the query examples and as negative ones the most dissimilar video shots for content-based video retrieval [33, 34]. Conversely, late fusion techniques and dynamic list fusion methods have been used to combine information from multiple retrieval models with the goal of improving the performance of a single retrieval technique [8, 13, 14].

Most of the above-described strategies have been defined for content-based image retrieval, although they can be easily adapted for textual and multimodal systems. Recently, a multimodal version of RF has been proposed for image retrieval [15]. The so-called intermedia RF technique consists of performing two RF iterations in which the modalities used in the first and second iterations are different [15, 37]. For each RF iteration it can be adopted by any of the above-described variants of RF.

Whereas most of the above-cited works have reported acceptable performance, they present several limitations. Most of them require the multiple execution of a retrieval model or of the on-line construction of classification methods, and the latter formulation can be computationally expensive. Efficiency is a crucial factor in image retrieval because real time response is required [3, 20]. Techniques based on active learning require a large amount of user interaction in order to obtain acceptable performance; however, some users may not be willing to spend much time interacting with a system; thus, user interaction must be minimized or interactive systems must give support to *lazy* users. Additionally, most of the described methods do not take into account all of the available information (e.g., initial ranking and contextual information) for refining the retrieval results of the base system, which can be helpful for improving the effectiveness of post-retrieval techniques.

In this paper we propose a variant of RF that aims at alleviating some of the limitations of current methods. We propose a reranking technique for refining the output of an image retrieval system. Our approach does not require the multiple execution of a retrieval model (other than the included for the base retrieval method) nor the construction of a classifier. In fact, our method does not process the entire collection of documents in each iteration of feedback, but it focuses on the top-$k$ retrieved documents. Whereas some documents can be out of reach for a particular query or user, this restriction makes the post-retrieval process very fast. Opposed to active-learning-based approaches, our method

only requires minimal user interaction and it could even work without the need of a user at all (i.e., under a blind RF formulation). Additionally, the proposed model takes advantage of all of the information available during a retrieval session, namely, initial ranking as provided by the base retrieval system, multimodal similarity among the retrieved images (e.g., context), and RF information. A notable benefit of our approach is that it is not restricted to a particular retrieval system nor to a specific information modality or system architecture. Thus, our method offers advantages in terms of generality as it can be used with any retrieval system, efficiency and effectiveness.

*2.2. Reranking Methods.* Similar document reranking approaches have been proposed elsewhere [16–21, 40–42]. Giacinto and Roli have developed reranking methods that rely on the distance of each document in the collection to the nearest relevant and irrelevant image as marked by the user [40–42]. The intuition behind that method is that a relevant image will present a small distance to relevant images marked by the user and large distance to irrelevant images identified by the user. Our MRF is based on a generalization of that idea: relevant images will show a small distance to relevant images marked by the user and at the same time they will be similar to each other; irrelevant images will lie at a considerable distance from relevant images identified by the user and they can have low similarity among them. Thus, differently from the work of Giacinto and Roli, we consider contextual information for reranking the list of images (besides we do not process the entire collection at each feedback iteration).

Cui et al. describe a reranking approach based on visual features for Web image retrieval systems [20]. Under their approach, query images are categorized into one of five intention categories; then, depending on the category of the query, different feature weights (computed off-line) are used to compute a new ranking for images according to the RankBoost framework [46]. Whereas this approach is very efficient, it is limited to five intention categories; further, they rely on pretrained models for detecting user intention in query images, which are expensive to train and subject to uncertainty. The method does not consider multimodal information nor supports relevance feedback. The experimental evaluation presented in [20] is performed over a large number of images but restricted to a small vocabulary.

Lin et al. [21] and Marakakis et al. [16] introduce image reranking approaches. Lin et al. use the relevance model from Lavrenko and Croft [47] to rerank Web images using global information (i.e., textual information obtained from the HTML page that contains the image) [21]. The proposed model takes into account information from the base system, although it disregards context and relevance feedback information. Also, limited performance is reported in a small scale experimental setting. Marakakis et al. propose an alternative probabilistic reranking technique that attempts to model relevancy from relevance feedback [16]. As most of the reported work, this method requires to process the entire collection of documents in each feedback iteration; additionally, if the feedback images change drastically the model must be trained again.

Tian et al. describe a Bayesian formulation for reranking the list of videos obtained with textual queries [17]. Their goal is to infer the final list of ranking scores from the initial list of scores and similarities among videos, under the premise that visually similar videos must be assigned similar ranking scores. Jing and Baluja propose a formulation similar to that of Google's page rank for image reranking in the task of product image retrieval [18]. Important (*authority*) images are identified based on their similarity with other images under consideration. The latter model was evaluated on a large scale product data set. The above approaches are closely related to our proposal, although they focus on image reranking based exclusively on visual similarities; also they do not give support to multimodal queries. Further, these methods do not incorporate other information available such as query-document similarity and RF information.

Yao et al. introduce a multimodal co-reranking approach [19], where visual and textual similarities are used to rerank the list of images provided by a textual image retrieval technique. The proposed technique is based on random walks over visual and textual similarity graphs. As with other techniques, neither multimodal query similarities nor RF information is considered by this method.

In previous work [26, 27] we have explored a similar approach where a (unimodal) MRF model has been used with either textual or visual features for image reranking. We explored the combination of internal and external similarity for improving the ranking of images. The best results obtained from that work improved the retrieval performance of the initial list up to 66% in the textual case, and 51% in the visual case when 10 relevance feedback images were considered, and up to 20% and 8%, respectively, when a single image was taken as feedback. Results from [26, 27] showed that in most of the cases both methods identified complementary sets of relevant images, motivating the development of the multimodal technique introduced in this paper.

## 3. Markov Random Fields

Markov random fields (MRFs) are a type of undirected probabilistic graphical models that aim at modeling dependencies among variables of the problem in turn [48]. MRFs have a long history within image processing and computer vision [49, 50]. They were first proposed for denoising digital images [48] and since then a large number of applications and extensions have been proposed. Classical applications include image segmentation [51] and image filtering [52]; although, recently they have been successfully applied for image annotation [53], region labeling [54, 55], and information retrieval [56–58] with great success.

MRF modeling has appealing features for problems that involve the optimization of configurations of variables that present interdependencies among them. They rely on a strict probabilistic modeling, yet they allow the incorporation of prior knowledge by means of potential functions. For these reasons, we adopted an MRF model for reranking images

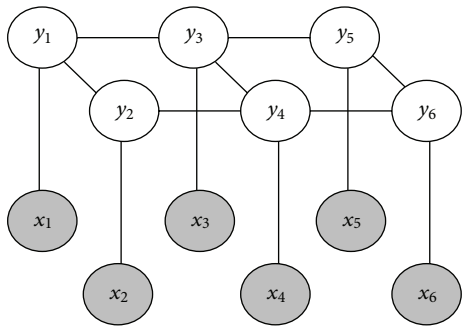


Figure 1: Graphical representation of a first-order MRF. Observed variables ($X$) are shaded.

listed by an image retrieval system. The rest of this section summarizes the formalism of MRFs.

An MRF is a set of random variables $F = \{f_1, \ldots, f_N\}$ indexed by nodes of a graph where the following conditions hold:

$$P(f_i) \geq 0, \quad \forall f_i \in F, \tag{1}$$

$$P(f_i \mid f_{s-\{i\}}) = P(f_i \mid \mathcal{N}(f_i)), \tag{2}$$

where $\mathcal{N}(f_i)$ is the set of neighbors of $f_i$ according to the neighboring system $\mathcal{N}$. Formula (1) is the so-called positivity condition and avoids negative probability values, whereas expression (2) states that the value of a random variable depends only on the set of neighbors of that variable, that is, the Markovian condition.

It has been shown that an MRF follows a Gibbs distribution [52], where a Gibbs distribution of the possible configurations of $F$ with respect to $\mathcal{N}$ has the following form:

$$P(f) = Z^{-1} \times e^{-(1/T)E(f)}, \tag{3}$$

where $Z$ is a normalization constant, and $T$ is the so-called temperature parameter (a common choice is $T = 1$) and $E(F)$ is an energy function of the following form:

$$E(F) = \sum_{c \in C} V_c(f) = \sum_{\{i\} \in C_1} V_1(f_i) + \sum_{\{i,j\} \in C_2} V_2(f_i, f_j) + \cdots, \tag{4}$$

where "$\cdots$" denotes possible potentials $V_c$ defined over higher order neighborhoods $C_3, C_4, \ldots, C_K$; each $C_i$ defines a neighborhood system of order $i$ between the nodes of the MRF. For example, Figure 1 shows an MRF with a neighborhood system of order 2. Often the set $F$ is considered as the union of two subsets of random variables $X \cup Y$; where $X$ is the set of observed variables and $Y$ is the set of output variables, which state we would like to predict. Potentials $V_c$ are problem dependent and commonly they are learned from data.

One of the main problems in MRFs is that of selecting the most probable configuration of $F$ (i.e., an assignment of values to each variable $f_i$ of the field). This configuration is obtained by minimizing Formula (4). For this purpose, a variety of optimization techniques have been used, including iterated conditioned modes (ICMs) [59], simulated annealing [60], and graph cuts [61]. ICM is one of the most used inference methods [48]; it is an iterative optimization procedure that performs local moves on the values of the nodes of the MRF. ICM fixes the value of all but one node in the MRF and determines the value of the remaining node by looking for the value minimizing Formula (4), and this process is repeated for all nodes and iterated several times. ICM does not guarantee finding the global optimum of Formula (4). However, it allows us to obtain acceptable locally optimal solutions; besides, it is a highly efficient method. Since efficiency is a crucial aspect in the considered reranking problem, we used ICM for inference in the multimodal MRF for image reranking.

## 4. Multimodal Markov Random Field for Image Reranking

The multimodal MRF we propose takes as input a list of $N$-ranked images, provided by an image retrieval system, and attempts to rerank the images in the list in such a way that relevant images are put before irrelevant ones. The proposed method can be added as a postprocessing stage for any image retrieval system, as it does not rely on information from a particular system. Figure 2 shows a schematic diagram of the proposed multimodal MRF. Besides the position of images in the list, the model incorporates interimage and query-image similarities as well as relevance feedback information. The rest of this section describes the multimodal MRF we propose.

We consider a MRF in which each node $F = \{f_1, \ldots, f_N\}$ corresponds to a document (image + text caption) in the list returned by a given retrieval system; each $f_i$ is binary random variable such that when $f_i = 1$ the $i$th-image is considered to be relevant to the search intention and when $f_i = 0$ the corresponding image is considered to be irrelevant. Figure 3 shows a diagram of the MRF for image reranking. The task of the multimodal MRF is to divide images in the list into relevant and irrelevant ones by varying the values of $F = \{f_1, \ldots, f_N\}$. Based on the final configuration of the MRF, we generate a new list of images by placing in the first positions those images $i$ for which $f_i = 1$, followed by the rest of images (i.e., images with $f_j = 0$); where we keep the relative position of images in the original list (i.e., images with $f_i = 1$ are put first in the order they appeared in the original list, followed by images for which $f_i = 0$ in the respective order).

We define an energy function for the multimodal MRF that attempts to model the relevancy status of images in terms of: (1) the information provided by the base retrieval system augmented with image-query similarity; (2) similarities among images in the list; and (3) relevance feedback information. Our hypothesis is, on the one hand, that relevant images must be very similar to the query (as they are supposedly relevant) and they must be similar to each other, as all relevant images are related to a common topic. On the other hand, irrelevant images must be less similar to the query than relevant ones and irrelevant images must be less similar

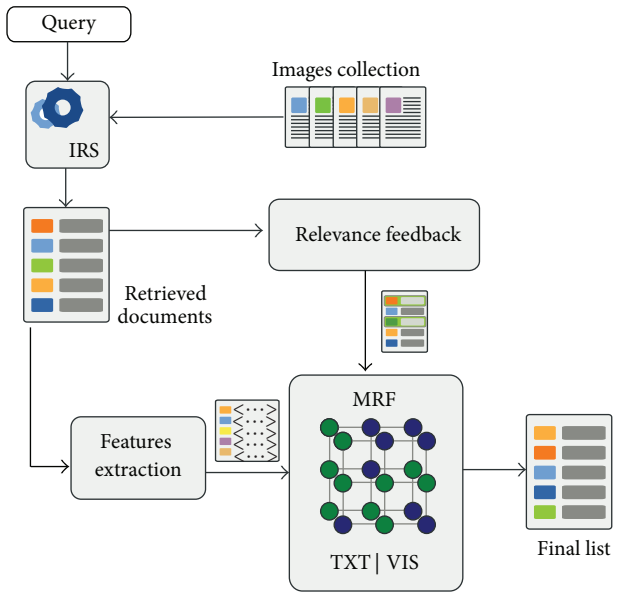


Figure 2: Schematic diagram for the proposed multimodal MRF. The MRF takes as input the list of images retrieved by an image retrieval system (IRS), relevance feedback information, and multimodal (textual and visual) features extracted from the images in the list. The structure of the MRF is depicted in Figure 3. The output of the model is a list of reranked images.

to each other, as they are not related to a common topic. The configuration of the MRF that minimizes the energy function is used to generate the new ranking.

*4.1. Energy Function.* The energy function of the multimodal MRF specifies how likely is that a configuration of the MRF (i.e., an assignment of values to each node $f_i$) is the best ranking for the images in the list. We define an energy function that incorporates interimage similarity, query-image similarity, rank information provided by the retrieval system, and relevance feedback information provided by the user. The underlying idea is that a combination of these information sources is beneficial for characterizing *good ranks*. Further, since the proposed model is multimodal, we take into account both textual and visual similarity. More specifically we propose an energy function of the following form:

$$U(F) = \sigma E_T(F) + (1-\sigma) E_V(F), \quad (5)$$

where $E_T(F)$ and $E_V(F)$ are individual energy functions defined for the textual and visual modalities, respectively, while the scalar $\sigma$ is introduced for weighting the influence of each modality. Since both $E_T(F)$ and $E_V(F)$ only differ in the way that similarity is estimated, we provide a general description of the form of the individual energy functions. In Section 4.2 we provide details on how we estimate the similarity for textual and visual features.

One should note that Formula (5) can be extended easily to incorporate information from more than two modalities. In that case, we would have an individual energy function per modality, which implies having means to compute similarity for each modality. While it is rather easy to extend the multimodal MRF to consider more modalities, we think that the inference process will become more difficult. This is because we will be optimizing an energy function with as many objectives as modalities. In that scenario the selection of $\sigma$ weights will be crucial. We would like to explore this research direction as future work.

Each individual energy function (i.e., $E_T(F)$ and $E_V(F)$) has the following form:

$$E(F) = \lambda \left( \sum_{f_i \in F} V_c(f_i, N_i) \right) + (1-\lambda) \left( \sum_{f_i \in F} V_a(f_i) \right), \quad (6)$$

where $N_i$ is the set of neighbors for node $f_i$, $V_c$ accounts for information of the association between neighboring images, whereas $V_a$ is the observation potential and it accounts for information that is associated to a single image, see Figure 3. $\lambda$ is a scalar that weights the importance of both $V_a$ and $V_c$.

We assume that each node in the multimodal MRF is connected to each other, that is, a fully connected graph. Since the number of images in the list is relatively small, considering a complete graph is not a computational issue and it allows us to consider the relations among all documents in the list. However, it is worth noting that both elements of the energy function in formula (6) are quadratic; that is, $O(V_c) = O(n^2)$ and $O(V_a) = O(n^2)$ being $n$ the number of documents contained in the retrieved list. Thus, $O(E) = O(n^2) + O(n^2) = O(\max(n^2, n^2)) = O(n^2)$ and since $\lambda$ is a constant, it is a depreciable element from the complexity calculation.

*4.1.1. Interaction Potential.* The interaction potential $V_c$ is defined as follows:

$$V_c(f_i, N_i) = \begin{cases} S\left(f_i, N_i^R\right) + \left(1 - S\left(f_i, N_i^I\right)\right), & \text{if } f_i = 0, \\ S\left(f_i, N_i^I\right) + \left(1 - S\left(f_i, N_i^R\right)\right), & \text{if } f_i = 1, \end{cases} \quad (7)$$

where $S(f_i, N_i^R)$ is the average similarity between image $f_i$ and its neighbors with relevant value $N_i^R$. Conversely, $S(f_i, N_i^I)$ represents the average similarity between the image associated with node $f_i$ and its neighbors with irrelevant value $N_i^I$. Thus, we divide the neighbors of node $f_i$ into two subsets: the neighbors with relevant value, $N_i^R$, and the neighbors with irrelevant value $N_i^I$ (i.e., $N_i = N_i^R \cup N_i^I$ and $N_i^R \cap N_i^I = \emptyset$).

Under the proposed MRF the minimization of formula (6) leads to improved rankings; hence, we seek configurations of the MRF with low values of $V_c(f_i, N_i)$. Accordingly, if an image is being considered relevant (i.e., $f_i = 1$), low values of $V_c(f_i, N_i)$ are obtained when the hypothetically relevant image is not too similar to irrelevant images and highly similar to relevant images. On the other hand, when a node is being considered irrelevant (i.e., $f_i = 0$), low values of $V_c(f_i, N_i)$ are obtained when the hypothetically irrelevant image is not too similar to relevant images and highly similar

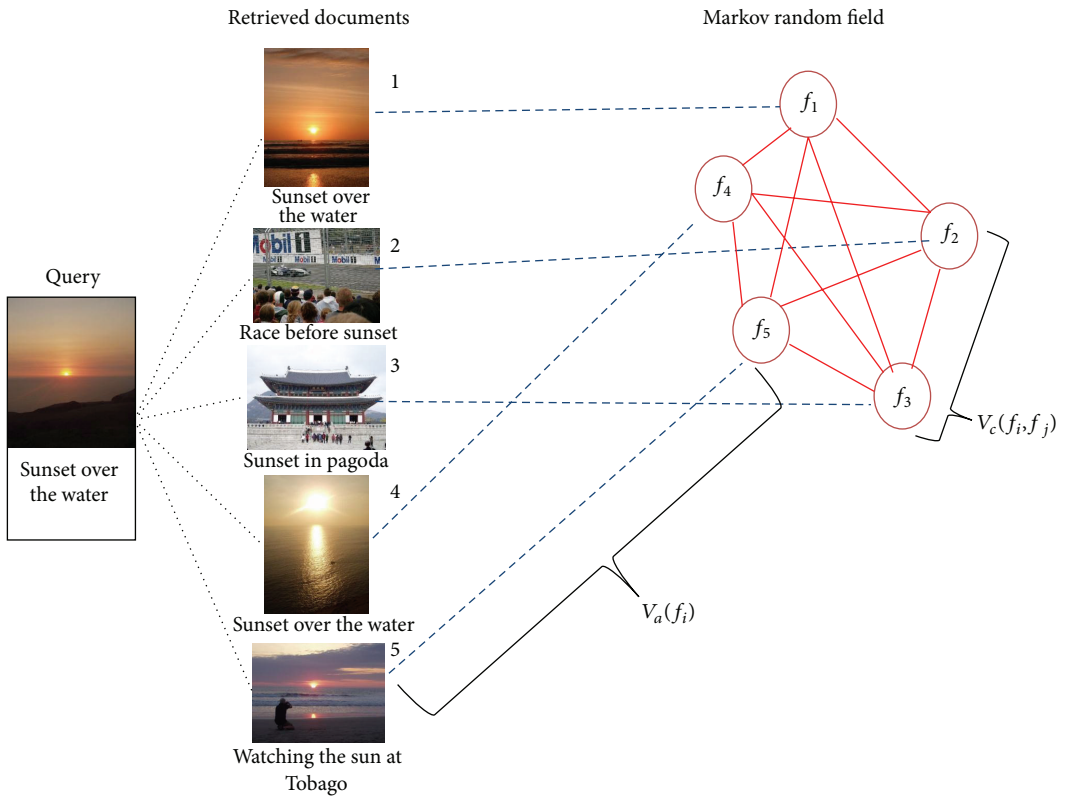


FIGURE 3: Diagram of the proposed MRF for image reranking. Each image in the original list is associated to a node in the MRF. We incorporate in the model the position of the document in the original list, the similarity of images to the query and the similarity among images. Single node information is incorporated through the $V_a(f_i)$ potential (dashed lines), while contextual information is introduced via the $V_c(f_i, f_j)$ potential (solid lines).

to irrelevant ones. Intuitively, $V_c$ assesses how much support we give same-valued images to keep the current value, and how much support we give oppose-valued images to change to the contrary value.

We compute the similarity $S(f_i, f_j)$ between a pair of images $I_i$ and $I_j$, associated to nodes $f_i$ and $f_j$, respectively, by comparing their visual or textual features. The particular similarity functions we consider are defined in such a way that they always return a normalized quantity (i.e., $S(f_i, f_j) \in [0, 1]$), see Section 4.2.

*4.1.2. Observation Potential.* The observation potential $V_a$ is defined as follows:

$$V_a(f_i) = \begin{cases} S_q(f_i, q) \times \delta\left(r(f_i)^{-1}\right), & \text{if } f_i = 0 \\ \left(1 - S_q(f_i, q)\right) \times \delta\left(r(f_i)\right), & \text{if } f_i = 1. \end{cases} \quad (8)$$

$V_a$ captures the affinity between the image associated to node $f_i$ and the query $q$, measured by a similarity term and by using information in the original list. On the one hand, $S_q(f_i, q)$, indicates how similar is the image associated to node $f_i$ to the query $q$. On the other hand, $\delta$ is a function that transforms the positions, $r(f_i)$ (resp., inverse of the position, $r(f_i)^{-1}$) of image $f_i$ in the original list into a real value.

The transformation $\delta$ is described as follows:

$$\delta(x) = \frac{\exp(x/20)}{\exp(5)} \quad (9)$$

when $f_i = 1$, $\delta(x)$ takes values proportional to the position of the image in the list and when $f_i = 0$, $\delta(x)$ takes values proportional to the inverse of the position in the list. The position (inverse of the position) of images is weighted exponentially because we want the position (resp., inverse of the position) to have more influence for the top-ranked (resp., bottom-ranked) images. The values 20 and 5 in expression (9) were fixed by trial and error on preliminary experimentation.

The observation potential incorporates information from the initial retrieval system; however, one should note that we only use the position of images in the list, which is independent of the retrieval system that was used to obtain the initial list. $V_a$ is based on the assumption that relevant images are very similar to the query and at the same time it is very likely that they appear in the top positions; on the other hand, irrelevant images are less similar to the query and it is very likely that they appear in the bottom positions of the list.

*4.1.3. Relevance Feedback.* We use relevance feedback information as a seed for building the initial configuration of the MRF; that is, we set $f_i = 1$ for relevance feedback images and we set $f_j = 0$ for the rest. In this way, the multimodal MRF

starts the energy minimization process knowing what images are potentially relevant to the query. Thus, the inference process consists of identifying further relevant images in the list by propagating through the MRF the relevance feedback information. Since we assume that relevance feedback images are indeed relevant to the query, they are considered to be relevant during the whole energy minimization process (i.e., the corresponding nodes are never set to 0).

Relevance feedback can be performed either automatically or manually. In previous work we have studied the use of manual and automatic relevance feedback for unimodal MRFs [26, 27]. We found that under both forms of feedback the unimodal MRFs can improve the retrieval performance of the initial list. However, as expected, improvements were larger when using manual relevance feedback. Therefore, in this work we limit ourselves to explore the performance of the multimodal MRF using manual relevance feedback and we postpone to future work experiments with pseudo-relevance feedback.

*4.1.4. Inference in the Multimodal MRF.* As stated before, the configuration that minimizes expression (5) is used for generating the new image ranking. In this work, such configuration is obtained via heuristic optimization using the iterated conditioned modes (ICMs) algorithm [59]. We also performed experiments with simulated annealing (SA). However, SA increases significantly the computational burden of the optimization process and the quality of solutions is comparable to that obtained with ICM. Hence we preferred ICM over SA. Graph cuts are another option to explore for future work, although it is also a time-consuming method.

*4.2. Similarity Estimation.* In Section 4.1 we described the form of the unimodal energy functions we consider. In this section we describe how similarity is estimated for textual and visual features. The proposed similarity measures are basically the normalized number of common words or SIFT features in the images that are compared. Before describing the similarity measures we describe the textual and visual features we extract from annotated images.

*4.2.1. Visual and Textual Features.* We consider images that are annotated with textual descriptions; specifically, we consider images from the IAPR TC12 collection [23], which was used extensively in the context of ImageCLEF [22]. Figure 4 shows a sample image from the IAPR TC12 collection. For representing images we consider visual features, extracted from the image itself, and textual features, extracted from its corresponding caption. Regarding Web-based collections, we can use a preprocessing step to extract the text from HTML files containing the analyzed image.

As textual features we use a binary bag of words representation, where each image is represented by a binary vector that indicates the presence (1) or absence (0) of words from the collection vocabulary in the document.

As visual features we consider the set of SIFT (Scale-Invariant Feature Transform) features extracted from the image [62]. In preliminary experiments we have explored other types of textual (e.g., *n*-grams and the *tf-idf* weighting scheme) and visual (e.g., color, texture, and shape) features; however, the best results with our multimodal MRF were obtained by using the binary bag of words together with the SIFT features. Images in the retrieved list and the corresponding queries are represented using the above multimodal features.

⟨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 parked in front of the house; a woman and a child are walking over the square; ⟨/DESCRIPTION⟩

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

*4.2.2. Similarity for Textual Features.* For the estimation of similarity in terms of textual features we propose the following functions.

(i) The similarity between two images $I_i$ and $I_j$ associated to nodes $f_i$ and $f_j$ in terms of their textual representation is defined as follows:

$$S\left(f_i, f_j\right) = \frac{\left|I_i \cap I_j\right|}{\left|I_i \cup I_j\right|}, \tag{10}$$

where $|I_i \cap I_j|$ and $|I_i \cup I_j|$ are the number of words that occur in the intersection and union, respectively, of the textual representations of $I_i$ and $I_j$. This similarity function is the ratio between the number of words that are common for both documents and the number of different words that occur in either document.

(ii) The similarity between an image $I_i$ associated to node $f_i$ and query $q$ using textual information is defined as follows:

$$S_q\left(f_i, q\right) = \frac{\left|I_i \cap q\right|}{\left|q\right|}. \tag{11}$$

The difference between $S_q(f_i, q)$ and $S(f_i, f_j)$ is the denominator $|q|$, which represents the number of words in the query. This difference is justified because we want $S_q(f_i, q)$ to be independent of the length of the documents.

*4.2.3. Similarity for Visual Features.* For the estimation of similarity in terms of SIFT features we propose the following functions.

(i) The similarity between two images $I_i$ and $I_j$ associated to nodes $f_i$ and $f_j$ in terms of their visual representation is defined as follows:

$$S\left(f_i, f_j\right) = \frac{2 \times \text{match}\left(f_i, f_j\right)}{\text{count}\left(f_i\right) + \text{count}\left(f_j\right)}, \tag{12}$$

where match $(f_i, f_j)$ is the number of similar descriptors between images $I_i$ and $I_j$ as described by Lowe [62] and count $(f_i)$ is the number of SIFT features found in image $I_i$.

(ii) The similarity between an image $I_i$ associated to node $f_i$ and the sample images $q_1, \ldots, q_H$ that compose the query $q$ is defined as follows (One should note that queries in the IAPR TC12 collection are composed of 3 sample images, i.e., $H = 3$, see Section 6.1.)

$$S_q\left(f_i, q\right) = \max_h \left( \frac{\text{match}\left(I_i, q_{1,\ldots,H}\right)}{\text{count}\left(q_{1,\ldots,H}\right)} \right), \tag{13}$$

where $H$ is the number of available query images. Thus we take the maximum similarity of the image $I_i$ to any query image $q_h$, as with the textual features, we want the similarity function to be independent of the number of SIFT features in image $I_i$.

## 5. Baseline Multimodal Methods

The multimodal MRF introduced in Section 4 is a way of combining the information from the unimodal energy functions $E_T$ and $E_V$, which take into account either textual or visual information, respectively. While in previous work we have studied the benefits of $E_T$ and $E_V$ for image reranking separately [26], it is clear that there are other alternatives for combining information from $E_T$ and $E_V$. The rest of this section describes two of such alternative methods based on MRFs that incorporate multimodal information, these techniques are considered baselines to which we compare our multimodal MRF.

The first strategy is a late fusion approach in which we run two unimodal MRFs, one using $E_V$ and the other using $E_T$, for reranking the list provided by the image retrieval system. The two reranked lists generated with the unimodal methods are combined to obtain a new list of images. The new list is generated by applying the CombMNZ technique, which combines the lists of results through redundancy and position information to reallocate the elements in a new order [63]. We call this strategy late fusion.

Figure 5 shows a diagram of this strategy. The intuitive idea of the late fusion approach is that since each unimodal MRF is based on information from different modalities, the resultant lists may contain complimentary ranking information, thus the combination of both lists may result in an even better ranking than those obtained through the unimodal MRFs.

We developed another alternative technique that is inspired by the so-called intermedia relevance feedback approach widely used in multimodal image retrieval [37]. The proposed technique consists of the serial application of unimodal MRFs, where a different energy function ($E_V$ or $E_T$) is used each time. The initial list is reranked with a unimodal MRF method using either $E_T$ or $E_V$, next the resultant reranked list is used as input for a second unimodal MRF, this time using a different energy function than that used in the first stage. The resultant list is the output of the method. We call this strategy intermedia relevance feedback.

Figure 6 shows the scheme of the intermedia relevance feedback strategy and its two possible configurations (i.e., using first $E_V$ then $E_T$ and vice versa). Intuitively, we assume that the list reranked by an initial (unimodal) MRF can be further improved by reranking it using another MRF, but this time using different information.

## 6. Experiments and Results

We conducted several experiments for evaluating the performance of the multimodal MRF described in Section 4 and that of the two alternative strategies introduced in Section 5. The goals of our experiments were (i) to evaluate the performance of the multimodal MRF under different parameter settings, (ii) to compare the performance of the multimodal MRF to the base retrieval system and to unimodal MRFs, and (iii) to compare the performance of the proposed multimodal MRF to that of the alternative multimodal methods for image reranking. The rest of this section describes the experimental results and highlights our main findings.

*6.1. Experimental Setup.* For our experiments we used the IAPR TC12 collection with topics and ground-truth data as used in ImageCLEF2008 [23], see Figure 4 for a sample image. This collection has been widely used by the image retrieval community, see for example [7, 13, 14, 22–27, 37, 54, 64]. The benchmark comprises 20,000 images with manual annotations in three languages; it is accompanied with sample queries and relevance judgements. Complexity of queries ranges from very simple (e.g., *destinations in Venezuela*) to extremely difficult (e.g., *creative group pictures in Uyuni* or *church with more than two towers*) and most queries require the combination of visual and textual information in order to obtain relevant documents. The considered ground-truth data contains 39 queries, where each query is composed of 3 sample images and a textual description. Figure 7 shows a sample query as considered in our experiments.

As base image retrieval system we considered the system that our group developed for the photographic retrieval task at ImageCLEF2008 and selected a particular list of images

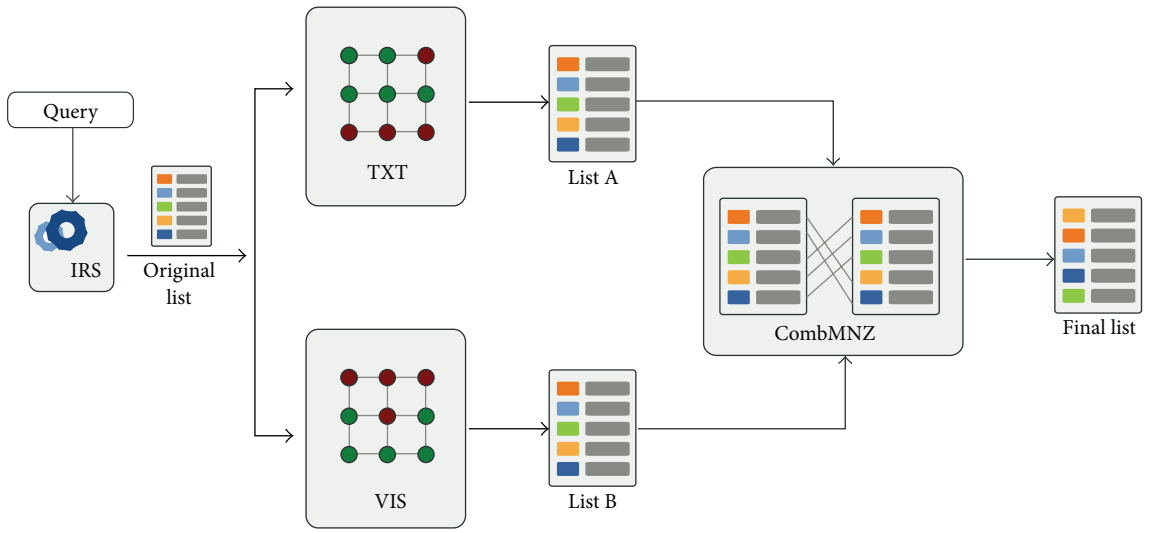


Figure 5: Diagram of the late fusion strategy. Two unimodal MRFs are run separately and the resultant lists are combined using the CombMNZ technique.

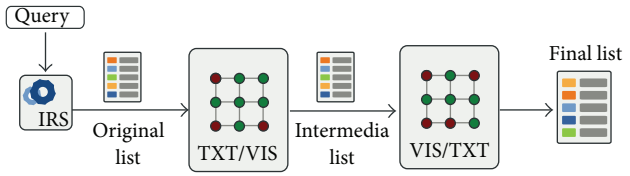


Figure 6: Diagram of the intermedia relevance feedback strategy. One unimodal MRF is applied after the other using different information (i.e., textual or visual).

⟨title⟩ animal swimmimg ⟨/title⟩
⟨narr⟩ Relevant images will show one or more animals (fish, birds, reptiles, etc.) swimming in a body of water. Images of people swimming in water are note relevant. Images of animals that are not swimming are not not relevant, ⟨/narr⟩

Figure 7: Query example for the photo retrieval track from the ImageCLEF2008 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. Three sample images are associated to each query.

that was evaluated in that forum. The specific list of images we considered (referred to as SRI TIA-TXTIMG) was ranked third in terms of recall in ImageCLEF2008 [64], which means that most of the relevant documents were retrieved. However, documents in that list were not placed in the right positions as it was ranked 77th in terms of mean average precision (MAP). Hence, there is a considerable potential of improvement for our reranking techniques. As input data for the proposed method we considered the top 100 images retrieved by the image retrieval system for each of the 39 queries. For evaluation we used the average over the 39 topics of the MAP as leading measure [65], although we also report precision (P) at different numbers of retrieved documents.

Each trial of our experiments proceeds as follows. We provide the original list as input to the technique under evaluation. Next, using ground truth information, we simulate a manual relevance feedback session by identifying $k$ relevance feedback images in the list (i.e., we randomly select $k$-relevant images, as indicated in the ground-truth information, that are contained in the list), for different values for $k$. These $k$-images are used as initial configuration for the method under evaluation. The ICM algorithm is run for 50 iterations (as maximum) trying to minimize the corresponding energy function, although in average it took only 7 iterations to converge. The final configuration of the method is used to generate a new list of ranked images. Finally, the retrieval performance of the reranked list is evaluated.

In the following, when mentioning a statistically significant difference between results we will refer to a paired $t$-Student test using a $\alpha = 0.05\%$ confidence level [66, 67]. We consider this test because it is one of the mostly used for evaluating information retrieval systems.

*6.2. Multimodal MRFs versus Unimodal MRFs.* In the first experiment we compared the performance of the multimodal MRF to that obtained by the base image retrieval system and that reported with unimodal MRFs. The goal of this experiment was to (a) quantify the improvement offered by the multimodal MRF for image reranking over the initial list and (b) to evaluate the difference in performance when using multimodal or unimodal information for the MRF.

Figure 8 compares the average MAP obtained with the multimodal MRF to that obtained by the base image retrieval

Table 1: MAP, precision at 5 (P5), 10 (P10), and 20 (P20) documents obtained by different parameter settings of the proposed multimodal MRF (columns 6–9), compared to the performance one would obtain with relevance feedback alone (columns 2–5). Column 1 indicates the parameter settings for the corresponding row as follows: the first value is $k \in \{1, 3, 5, 8, 10\}$, the number of relevance feedback images, $\lambda_V$ is the value of $\lambda$ for the visual energy function (i.e., $E_V$), and $\lambda_T$ is the value of $\lambda$ for the textual energy function (i.e., $E_T$).

| Base IRS | P5 | P10 | P20 | MAP | | | | |
|---|---|---|---|---|---|---|---|---|
| SRI TIA-TXTIMG | 0.4769 | 0.4538 | 0.3910 | 0.2359 | | | | |
| $\sigma = 1$ | | Only relevance feedback | | | | Multimodal MRF | | |
| Configuration | P5 | P10 | P20 | MAP | P5 | P10 | P20 | MAP |
| $k = 1; \lambda_V = 0.5; \lambda_T = 1.0$ | 0.5641 | 0.4897 | 0.4038 | 0.2486 | 0.6051 | 0.5436 | 0.4756 | 0.2852 |
| $k = 3; \lambda_V = 0.5; \lambda_T = 1.0$ | 0.7846 | 0.5872 | 0.4359 | 0.2742 | 0.7897 | 0.6282 | 0.5141 | 0.3123 |
| $k = 5; \lambda_V = 0.5; \lambda_T = 1.0$ | 1.0000 | 0.6795 | 0.4782 | 0.2984 | 1.0000 | 0.7256 | 0.5551 | 0.3344 |
| $k = 8; \lambda_V = 0.5; \lambda_T = 1.0$ | 1.0000 | 0.8692 | 0.5577 | 0.3352 | 1.0000 | 0.8846 | 0.6359 | 0.3752 |
| $k = 10; \lambda_V = 0.5; \lambda_T = 1.0$ | 1.0000 | 0.9744 | 0.6128 | 0.3560 | 1.0000 | 0.9744 | 0.6628 | 0.3919 |
| $\sigma = 0.7$ | | Only relevance feedback | | | | Multimodal MRF | | |
| Configuration | P5 | P10 | P20 | MAP | P5 | P10 | P20 | MAP |
| $k = 1; \lambda_V = 0.5; \lambda_T = 1.0$ | 0.5641 | 0.4897 | 0.4038 | 0.2486 | 0.6462 | 0.5744 | 0.5038 | 0.2960 |
| $k = 3; \lambda_V = 1.0; \lambda_T = 1.0$ | 0.7846 | 0.5872 | 0.4359 | 0.2742 | 0.8000 | 0.6282 | 0.5205 | 0.3153 |
| $k = 5; \lambda_V = 0.7; \lambda_T = 1.0$ | 1.0000 | 0.6795 | 0.4782 | 0.2984 | 1.0000 | 0.7256 | 0.5615 | 0.3375 |
| $k = 8; \lambda_V = 0.0; \lambda_T = 0.5$ | 1.0000 | 0.8692 | 0.5577 | 0.3352 | 1.0000 | 0.8718 | 0.5897 | 0.3470 |
| $k = 10; \lambda_V = 0.7; \lambda_T = 1.0$ | 1.0000 | 0.9744 | 0.6128 | 0.3560 | 1.0000 | 0.9744 | 0.6731 | 0.3936 |
| $\sigma = 0.5$ | | Only relevance feedback | | | | Multimodal MRF | | |
| Configuration | P5 | P10 | P20 | MAP | P5 | P10 | P20 | MAP |
| $k = 1; \lambda_V = 0.7; \lambda_T = 1.0$ | 0.5641 | 0.4897 | 0.4038 | 0.2486 | 0.6513 | 0.5821 | 0.5077 | 0.2956 |
| $k = 3; \lambda_V = 0.7; \lambda_T = 1.0$ | 0.7846 | 0.5872 | 0.4359 | 0.2742 | 0.8103 | 0.6487 | 0.5282 | 0.3154 |
| $k = 5; \lambda_V = 1.0; \lambda_T = 1.0$ | 1.0000 | 0.6795 | 0.4782 | 0.2984 | 1.0000 | 0.7256 | 0.5718 | 0.3358 |
| $k = 8; \lambda_V = 1.0; \lambda_T = 1.0$ | 1.0000 | 0.8692 | 0.5577 | 0.3352 | 1.0000 | 0.8923 | 0.6500 | 0.3801 |
| $k = 10; \lambda_V = 1.0; \lambda_T = 1.0$ | 1.0000 | 0.9744 | 0.6128 | 0.3560 | 1.0000 | 0.9744 | 0.6885 | 0.3966 |
| $\sigma = 0.3$ | | Only relevance feedback | | | | Multimodal MRF | | |
| Configuration | P5 | P10 | P20 | MAP | P5 | P10 | P20 | MAP |
| $k = 1; \lambda_V = 1.0, \lambda_T = 0.7$ | 0.5641 | 0.4897 | 0.4038 | 0.2486 | 0.6718 | 0.5923 | 0.5167 | 0.3004 |
| $k = 3; \lambda_V = 1.0, \lambda_T = 1.0$ | 0.7846 | 0.5872 | 0.4359 | 0.2742 | 0.8051 | 0.6436 | 0.5359 | 0.3142 |
| $k = 5; \lambda_V = 1.0, \lambda_T = 1.0$ | 1.0000 | 0.6795 | 0.4782 | 0.2984 | 1.0000 | 0.7462 | 0.5987 | 0.3432 |
| $k = 8; \lambda_V = 1.0, \lambda_T = 1.0$ | 1.0000 | 0.8692 | 0.5577 | 0.3352 | 1.0000 | 0.9000 | 0.6833 | 0.3851 |
| $k = 10; \lambda_V = 1.0, \lambda_T = 1.0$ | 1.0000 | 0.9744 | 0.6128 | 0.3560 | 1.0000 | 0.9744 | 0.7218 | 0.4031 |
| $\sigma = 0$ | | Only relevance feedback | | | | Multimodal MRF | | |
| Configuration | P5 | P10 | P20 | MAP | P5 | P10 | P20 | MAP |
| $k = 1; \lambda_V = 0.3, \lambda_T = 1.0$ | 0.5641 | 0.4897 | 0.4038 | 0.2486 | 0.6513 | 0.5282 | 0.4128 | 0.2569 |
| $k = 3; \lambda_V = 0.7, \lambda_T = 1.0$ | 0.7846 | 0.5872 | 0.4359 | 0.2742 | 0.8410 | 0.6205 | 0.4423 | 0.2795 |
| $k = 5; \lambda_V = 0.0, \lambda_T = 1.0$ | 1.0000 | 0.6795 | 0.4782 | 0.2984 | 1.0000 | 0.7000 | 0.4846 | 0.3016 |
| $k = 8; \lambda_V = 0.5, \lambda_T = 1.0$ | 1.0000 | 0.8692 | 0.5577 | 0.3352 | 1.0000 | 0.8821 | 0.5654 | 0.3374 |
| $k = 10; \lambda_V = 0.7, \lambda_T = 1.0$ | 1.0000 | 0.9744 | 0.6128 | 0.3560 | 1.0000 | 0.9744 | 0.6205 | 0.3580 |

Table 2: Relative improvements (%) in terms of MAP of relevance feedback alone (RF), textual MRF ($t$-MRF), visual MRF ($v$-MRF), multimodal MRF (MM-MRF), the two intermedia relevance feedback variants (IRF ($T > V$) and IRF ($V > T$)), and the late fusion (LF) method over the base image retrieval system.

| $k$ | RF | $t$-MRF | $v$-MRF | MM-MRF | IRF $T > V$ | IRF $V > T$ | LF |
|---|---|---|---|---|---|---|---|
| 1 | 5.38 | 20.90 | 8.90 | **27.34** | 24.16 | 23.31 | 16.53 |
| 3 | 16.24 | 32.39 | 18.48 | 33.70 | 34.29 | **34.17** | 27.64 |
| 5 | 26.49 | 41.75 | 27.85 | **45.49** | 42.81 | 42.48 | 36.12 |
| 8 | 42.09 | 59.05 | 43.03 | **63.25** | 59.60 | 59.52 | 52.40 |
| 10 | 50.91 | 66.13 | 51.76 | **70.88** | 66.77 | 66.47 | 59.81 |

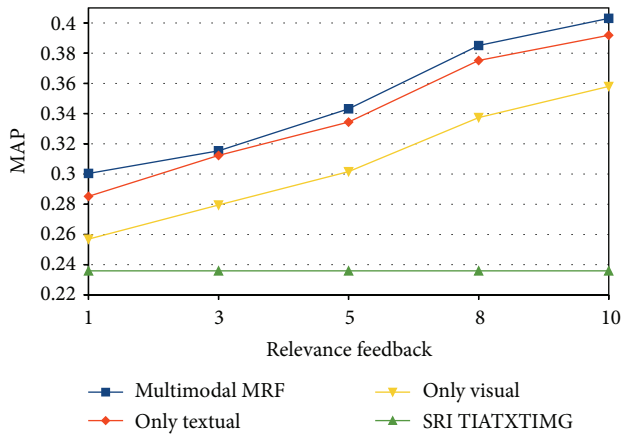


FIGURE 8: Comparison of the multimodal MRF (blue squares) with the base image retrieval system (green upwards-triangles), the textual MRF (red rhombus), and the visual MRF (yellow downwards-triangles). The plot shows the average MAP of each method ($y$-axis) against the number of relevant feedback images considered ($x$-axis).

system [64], the textual and the visual unimodal MRFs [26, 27]. We performed experiments with the following numbers of relevance feedback images: $k \in \{1, 3, 5, 8, 10\}$, which are used as seed for the reranking methods, see Section 4. The parameters of all of the methods were set empirically, in Section 6.3 we compare the performance of the multimodal MRF for different values of $\sigma$, $\lambda$, and $k$. The results for unimodal methods were taken from previous work [26, 27].

From this figure we can see that the three reranking techniques outperformed the base retrieval system, thus confirming that postprocessing techniques based on MRFs are beneficial for image reranking. As expected, improvements are larger for higher values of $k$, although there are important improvements even when $k = 1$, which is a very positive result as providing a single relevance feedback image is an easy task for the user.

The unimodal MRF based on textual features outperformed the corresponding MRF based on visual features. This result can be due to the fact that textual methods have traditionally reported much better performance than visual techniques in the IAPR TC12 collection [22, 64]. Another possibility is that we have not used the right visual features for effectively representing images. We performed preliminary experiments using color and texture features used in previous work [68], in addition to SIFT features, nevertheless our best results in image reranking were obtained with SIFT features. We will explore other types of visual features for representing images in future work.

The multimodal MRF consistently outperformed the results obtained by unimodal MRFs for all values of $k$, giving evidence that the combination of textual and visual features results in better reranking performance than using such modalities separately. The improvement was more important for the visual MRF, although the multimodal MRF outperforms the textual one by a significant margin, with exception of $k = 3$. We are not sure of what causes this result, but it could be related to the fact that we have 3 query images that are considered for estimating similarity in individual energy functions, see Section 4.2. Anyways, one should note that small improvements in MAP are very difficult to obtain in this collection [22]. The differences between the unimodal MRFs and the multimodal MRF were statistically significant, see Section 6.1.

It is important to mention that (although results are not directly comparable (Results are not directly comparable as the entries evaluated in ImageCLEF2008 included the top 1000 ranked documents, whereas the results in this paper were computed with the top 100 documents only. Please note that using the top 1000 ranked documents would increase the performance of the multimodal MRF. For example, the initial list (i.e., SRI) achieved an MAP of almost 0.24 considering the top-100 images, while we obtained a MAP of 0.3066 when the top 1000 images were considered [64].)) the results in terms of MAP, obtained with the multimodal MRF as shown in Figure 8 would be ranked in positions 5th, 8th, 22nd, 29th, and 32th of the results (1042 entries) evaluated in ImageCLEF2008 [25], when using 10, 8, 5, 3, and 1 relevance feedback images, respectively. Recall the list of the retrieval system we considered was ranked 77th in terms of MAP. This result gives us an idea on the effectiveness of the proposed approach when compared to state-of-the-art systems.

One should note that since relevance feedback images are always considered relevant, regardless of the values of the energy function, they will be always relevant to the query in turn (as relevance feedback images are obtained via simulated relevance feedback [32]). Therefore, it is necessary to make a comparison of the results obtained by using relevance feedback alone (i.e., putting relevant images in the first positions and leaving the rest of the list unchanged) versus results from the multimodal MRF method. Accordingly, in the rest of the paper we also compare the performance of our reranking methods to that one we would obtain by using relevance feedback alone.

*6.3. Parameter Settings for the Multimodal MRF.* In a second experiment we evaluated the performance of our multimodal MRF under different parameter settings. The goal of the experiment is to get insights about the importance of each information modality and the information provided by the observation and interaction potentials.

The parameters that have a direct impact in the performance of the reranking methods based on MRFs are the following: $\lambda$, the scalar weighting the contribution of contextual and individual information into the energy functions (see Formula (6)); $\sigma$, the scalar weighting the contribution of textual and visual information into the multimodal MRF (see Formula (5)); and $k$, the number of relevance feedback images that are used as seed for the MRFs. Table 1 shows the performance of our multimodal MRF for different configurations of the just-mentioned parameters. The values of $\sigma$, $\lambda_V$, and $\lambda_T$ were varied in the range $[0, 1]$; eleven equally spaced values in such interval were tried for each parameter. We show the best results obtained by each configuration of values for $k$, $\lambda_V$, and $\lambda_T$ for the following fixed values of $\sigma = \{1, 0.7, 0.5, 0.3, 0\}$.

From Table 1 we can see that the multimodal MRF outperformed relevance feedback in all of the reported cases in terms of the MAP; all of the differences between columns 5 and 9 were statistically significant. When $\sigma = 0$ (i.e., no textual information was considered) we obtained worse results than those obtained with other values, confirming that visual features alone are not very useful for image reranking. Results with $\sigma = 1$ (i.e., no visual information was considered) were acceptable, although they were worse than those obtained with $\sigma = 0.3$, $\sigma = 0.5$, and $\sigma = 0.7$. The best results were obtained with $\sigma = 0.3$, thus giving evidence that multimodal information, under the considered features, is indeed helpful for image reranking and reflecting the fact that a higher weight (penalty) to the visual potential results in larger improvements.

From Table 1 we can also see that low values of $\lambda_V$ were used for obtaining better retrieval results, which means that for the visual energy function ($E_V$) single node information was more helpful than contextual information (recall that each configuration of parameters, a row in Table 1, is the best configuration of values for $\lambda_T$ and $\lambda_V$ that was obtained for a fixed value of $k$ and $\sigma$). The latter can be due to the fact that relevant images are not very similar to each other, giving evidence that relevant images may be *visually diverse* among them for ImageCLEF2008 topics. On the other hand, the values of $\lambda_T$ were very close to 1 in general, which means that for the textual energy function ($E_T$) the contextual information was highly informative, while similarity between queries and images was not very helpful. The latter result indicates that relevant images have similar annotations. Hence, an interesting finding of our work is that it seems that relevant images are related to a common semantic topic (as they have similar high-level annotations) and at the same time, images are not visually similar (as there is a broad visual diversity among relevant images). Thus, a promising direction for jointly optimizing relevance and diversity may be the combination of both contextual textual information and individual visual information, we will explore this research direction as future work.

Regarding the number of relevance feedback images, we found that the larger the number of selected images the better the performance of the method. However, we would like to emphasize that even when a single image is provided by the (simulated) user, the multimodal MRF was able to significantly outperform the results obtained with relevance feedback alone. In fact, the differences in performance between relevance feedback and our multimodal MRF are larger when $k = 1$.

*6.4. Comparison of Multimodal MRFs to Multimodal Alternatives.* In our third experiment we compared the performance of the multimodal MRF to that obtained by the baseline reranking methods described in Section 5. The goal of this experiment is to determine whether the multimodal MRF is able to improve the performance of alternative multimodal image reranking techniques. Figure 9 shows the average MAP obtained by all of the developed methods. Note that for the intermedia relevance feedback approach we performed

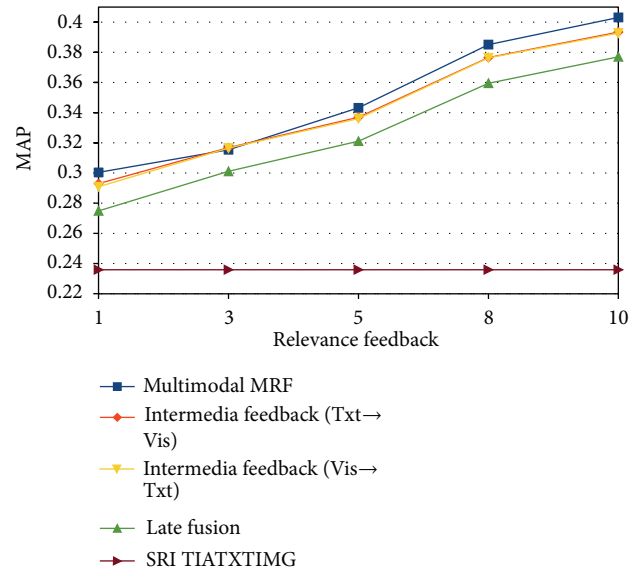


Figure 9: Comparison of the multimodal MRF (blue squares) and baseline methods described in Section 5 for different numbers of relevant documents. We show the performance of the base retrieval system (dark-red rightwards-triangles), the late fusion technique (green upwards-triangles), and intermedia relevance feedback in its two variants TXT > VIS (red rhombus) and VIS > TXT (yellow downwards-triangles). The plot shows the average MAP of each method ($y$-axis) against the number of relevant feedback images considered ($x$-axis).

experiments with both variants: using the textual MRF and then the visual MRF (TXT > VIS) and vice versa (VIS > TXT).

The plot shows that the multimodal MRF outperformed the other two techniques, and the differences were statistically significant for all but for the intermedia relevance feedback approach with $k = 3$ relevant documents. The late fusion approach achieved the smallest improvement over the base retrieval system, which can be due to the fact that we are weighting equally the contribution of visual and textual modalities for generating a new list of images. Hence, the influence of the list obtained with the visual MRF may affect the performance of the fusion.

The results obtained with both variants of the intermedia relevance feedback technique were very close to each other. Actually, it is very difficult to distinguish these results in Figure 9. This can be due to the fact that no matter how the ranking starts, there is a dominant MRF that generates the final ranking regardless of the stage in which it is applied; we assume that the textual MRF is the dominant technique for this experiment. Our assumption is based on the fact that the performance of intermedia relevance feedback is very close to that obtained by the textual MRF, see Figure 8.

With exception of $k = 3$, the multimodal MRF consistently outperformed the other baseline methods, hence proving that the specific form of the multimodal energy function is a better way to combine textual and visual information for image reranking with MRFs. Thus, it is beneficial to combine information in the interaction process (i.e., similar to early fusion technique [69]).

In average, the multimodal MRF took about 30 milliseconds to rerank a list of images. Hence, the efficiency of the proposed method is adequate for the standard image retrieval scenario that requires subsecond response times. One should note that the multimodal MRF is implemented in Matlab, therefore, lower response times would be expected when using programming languages as Java or C++, which are known to be more efficient than Matlab. Regarding the baseline methods, the late fusion approach obtained a similar average processing time to the multimodal MRF (i.e., ≈20 ms), whereas the intermedia relevance feedback method took an average of ≈50 ms to rerank a list of images.

Finally, Table 2 shows the relative improvements obtained by each method over the base image retrieval system. While these results are consistent with those reported in Figure 9, in this table we can appreciate the differences in terms of percentage of MAP improvement. The multimodal MRF obtained improvements over the base retrieval method that were superior to those achieved by the other methods. All of the differences reported in Table 2 were statistical significant. The multimodal MRF outperformed relevance feedback by 19.91% in average and the most competitive baseline (IRF ($T > V$)) by 2.6%.

## 7. Conclusions and Future Work

We have introduced a multimodal Markov random field for improving the order of a list of images as provided by an image retrieval system. The model incorporates similarity between images in the list, similarity between images and query, information obtained from the original order, and relevance feedback information. The reranking problem is faced as one of separating relevant from irrelevant images in the list. Our work included the development of potentials and energy functions based on textual and visual features that allowed us to differentiate relevant from irrelevant images. Additionally, we developed alternative multimodal strategies for image reranking based on MRFs.

Experimental results obtained in a benchmark collection composed of 20,000 annotated images showed the effectiveness of our method. In particular, results showed that the proposed method outperformed the base image retrieval system and a simple relevance feedback strategy. The proposed baselines achieved very competitive results, although the best results overall were obtained with our multimodal MRF. We studied the performance of the proposed model under different parameter settings and found that the combination of visual and textual information is more helpful than using unimodal MRFs separately. We also found that in the multimodal MRF the contextual potential was very important for the textual information, while image-query similarity was more helpful for the visual information. Thus, a future work direction is to study hybrid energy functions.

The contributions of this work are as follows. We proposed a novel image reranking technique that incorporates visual and textual information. To the best of our knowledge there are no similar works on (a) using MRFs for image result reranking nor on (b) image reranking methods that face the problem as one of combinatoric optimization. The proposed model was able to improve the ranking of the base retrieval system and the performance of the base list modified with relevance feedback information. Our approach incorporates contextual information, which is often disregarded in usual postretrieval techniques. The MRF relies on manual relevance feedback, hence it is a user adaptive mechanism. Moreover, since our MRF works with a list of ranked images, it is not tied with a specific retrieval system nor with a particular architecture, image collection or information modality.

As future work we would like to modify the multimodal MRF so that the reranking of images can be guided by the maximization of retrieval performance and result diversification. Another direct research direction would be extending the proposed MRF to take into account more than two modalities. Also, we would like to extend the multimodal MRF to incorporate levels of relevance instead of a binary relevance formulation. We think that a promising future work direction is the development of dynamic multimodal MRFs that can be used through retrieval sessions instead of reranking a single queries. Other research directions are exploring other types of visual features for representing images, performing experiments with other image retrieval systems, and testing the proposed method with other images collections, for example, unstructured collections like results from a Web search engine.

## Acknowledgments

The authors are grateful to the reviewers, whose comments have helped them to improve significantly this paper. This work was supported by CONACyT under project Grant 61335 and scholarships 205834 and 205834.

## References

[1] 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*, vol. 22, no. 12, pp. 1349–1380, 2000.

[2] A. Goodrum, "Image information retrieval: an overview of current research," *Journal of Informing Science*, vol. 3, no. 2, pp. 63–66, 2000.

[3] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Image retrieval: ideas, influences, and trends of the new age," *ACM Computing Surveys*, vol. 40, no. 2, article 5, 2008.

[4] Y. Liu, D. Zhang, G. Lu, and W. Ma, "A survey of content-based image retrieval with high-level semantics," *Pattern Recognition*, vol. 40, no. 1, pp. 262–282, 2007.

[5] 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*, vol. 2, no. 1, pp. 1–19, 2006.

[6] Y. Rui, T. Huang, and S. Chang, "Image retrieval: current techniques, promising directions and open issues," *Journal of Visual Communication and Image Representation*, vol. 10, no. 4, pp. 39–62, 1999.

[7] P. Clough, M. Grubinger, T. Deselaers, A. Hanbury, and H. Müller, "Overview of ImageCLEF 2006 Photographic retrieval

and object annotation tasks," in *Proceedings of the 7th Workshop of the Cross-Language Evaluation Forum (CLEF '07)*, vol. 4730 of *Lecture Notes in Computer Science*, pp. 579–594, Springer, 2007.

[8] P. K. Atry, M. A. Hossain, A. E. Saddik, and M. S. Kankanhalli, "Multimodal fusion for multimedia analysis," *Multimedia Systems*, vol. 16, no. 6, pp. 345–379, 2010.

[9] M. Broilo and F. G. B. De Natale, "A stochastic approach to image retrieval using relevance feedback and particle swarm optimization," *IEEE Transactions on Multimedia*, vol. 12, no. 4, pp. 267–277, 2010.

[10] 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*, vol. 8, no. 5, pp. 644–655, 1998.

[11] X. Zhou and T. Huang, "Relevance feedback in image retrieval: a comprehensive review," *Multimedia Systems*, vol. 8, pp. 536–544, 2003.

[12] T. Deselaers, T. Gass, P. Dreuw, and H. Ney, "Jointly optimising relevance and diversity in image retrieval," in *Proceedings of the ACM International Conference on Image and Video Retrieval (CIVR '09)*, pp. 296–303, ACM Press, July 2009, paper 39.

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

[14] A. Juàrez, M. Montes, L. Villaseñor, D. Pinto, and M. Pérez, "Selecting the N-top retrieval result lists for an effective data fusion," in *Proceedings of the 11th International Conference on Intelligent Text Processing and Computational Linguistics*, vol. 6008 of *Lecture Notes in Computer Science*, pp. 580–589, Springer, 2010.

[15] Y. Chang, W. Lin, and H.-H. Chen, "Combining text and image queries at ImageCLEF 2005," in *Working Notes of the CLEF Workshop*, Vienna, Austria, 2005.

[16] A. Marakakis, N. Galatsanos, A. Likas, and A. Stafylopatis, "Application of relevance feedback in content based image retrieval using gaussian mixture models," in *Proceedings of the 20th IEEE International Conference on Tools with Artificial Intelligence (ICTAI '08)*, pp. 141–148, November 2008.

[17] X. Tian, L. Yang, J. Wang, Y. Yang, X. Wu, and X. S. Hua, "Bayesian video search reranking," in *Proceedings of the 16th ACM International Conference on Multimedia*, pp. 131–140, ACM Press, Vancouver, Canada, 2008.

[18] Y. Jing and S. Baluja, "PageRank for product image search," in *Proceedings of the International World Wide Web Conference Committee*, pp. 307–315, ACM Press, Beijing, China, 2008.

[19] T. Yao, T. Mei, and C. W. Ngo, "Co-reranking by mutual reinforcement for image search," in *Proceedings of the ACM International Conference on Image and Video Retrieval*, pp. 34–41, ACM Press, Xian, China, 2010.

[20] J. Cui, F. Wen, and X. Tang, "Real time google and live image search re-ranking," in *Proceedings of the ACM Multimedia Information Retrieval Conference*, pp. 729–732, ACM Press, Vancouver, Canada, 2008.

[21] W. Lin, R. Jin, and A. Hauptmann, "A web image retrieval re-ranking with relevance model," in *Proceedings of the IEEE International Conference on Web Intelligence*, p. 242, 2003.

[22] H. Müller, P. Clough, T. Deselaers, and B. Caputo, *ImageCLEF: Experimental Evaluation in Visual Information Retrieval*, Springer Series on Information Retrieval, 2010.

[23] M. Grubinger, *Analysis and evaluation of visual information systems performance [Ph.D. thesis]*, School of Computer Science and Mathematics, Faculty of Health, Engineering and Science, Victoria University, Melbourne, Australia, 2007.

[24] P. Clough, M. Grubinger, T. Deselaers, A. Hanbury, and H. Müller, "Overview of the ImageCLEF 2007 photographic retrieval task," in *Proceedings of the 8th Workshop of the Cross-Language Evaluation Forum (CLEF '08)*, vol. 5152 of *Lecture Notes in Computer Science*, pp. 433–444, Springer, 2008.

[25] T. Arni, M. Sanderson, P. Clough, and M. Grubinger, "Overview of the ImageCLEF 2007 photographic retrieval task," in *Evaluating Systems for Multilingual and Multimodal Information Access*, vol. 5706 of *Lecture Notes in Computer Science*, pp. 500–511, Springer, 2009.

[26] R. O. Chàvez, M. Montes, and E. Sucar, "Using a markov random field for image re-ranking based on visual and textual features," *Computación y Sistemas*, vol. 14, no. 4, pp. 393–404, 2011.

[27] R. O. Chàvez, M. Montes, and E. Sucar, "Image Re-ranking based on relevance feedback combining internal and external similarities," in *Proceedings of the 23rd International FLAIRS Conference*, pp. 140–141, Daytona Beach, Fla, USA, 2010.

[28] I. J. Cox, M. L. Miller, T. P. Minka, T. V. Papathomas, and P. N. Yianilos, "The Bayesian image retrieval system, PicHunter: theory, implementation, and psychophysical experiments," *IEEE Transactions on Image Processing*, vol. 9, no. 1, pp. 20–37, 2000.

[29] C. Zhang, J. Y. Chai, and R. Jin, "User term feedback in interactive text-based image retrieval," in *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 51–58, ACM Press, Salvador, Brazil, 2005.

[30] Z. H. Zhou, K. E. J. Chen, and H. B. Dai, "Enhancing relevance feedback in image retrieval using unlabeled data," *ACM Transactions on Information Systems*, vol. 24, no. 2, pp. 219–244, 2006.

[31] S. Tong and E. Chang, "Support vector machine active learning for image retrieval," in *Proceedings of the ninth ACM international conference on Multimedia*, pp. 107–118, ACM Press, Ottawa, Canada, 2001.

[32] T. Deselaers, R. Paredes, E. Vidal, and H. Ney, "Learning weighted distances for relevance feedback in image retrieval," in *Proceedings of the 19th International Conference on Pattern Recognition (ICPR '08)*, pp. 1–4, Tampa, Fla, USA, December 2008.

[33] R. Yan, A. G. Hauptmann, and R. Jin, "Negative pseudo-relevance feedback in content-based video retrieval," in *Proceedings of the 11th ACM International Conference on Multimedia*, pp. 343–346, ACM Press, Berkeley, Calif, USA, 2003.

[34] R. Yan, A. G. Hauptmann, and R. Jin, "Multimedia search with pseudo-relevance feedback," in *Proceedings of the International Conference on Image and Video Retrieval*, ACM Press, Urbana, Ill, USA, 2003.

[35] H. Ma, J. Zhu, M. R. Lyu, and I. King, "Bridging the semantic gap between image contents and tags," *IEEE Transactions on Multimedia*, vol. 12, no. 5, pp. 462–473, 2010.

[36] H. Tong, J. He, M. Li, W. Y. Ma, H. J. Zhang, and C. Zhang, "Manifoldranking-based keyword propagation for image retrieval," *EURASIP Journal on Applied Signal Processing*, vol. 2006, Article ID 079412, 2006.

[37] J. Ah-Pine, M. Bressan, S. Clinchant, G. Csurka, Y. Hoppenot, and J. M. Renders, "Crossing textual and visual content in different application scenarios," *Multimedia Tools and Applications*, vol. 42, no. 1, pp. 31–56, 2009.

[38] K. Porkaew and K. Chakrabarti, "Query refinement for multimedia similarity retrieval in MARS," in *Proceedings of the 7th ACM International Conference on Multimedia*, pp. 235–238, ACM Press, 1999.

[39] K. Porkaew, M. Ortega, and S. Mehrotra, "Query reformulation for content based multimedia retrieval in MARS," in *Proceedings of the 6th International Conference on Multimedia Computing and Systems (IEEE ICMCS '99)*, pp. 747–751, June 1999.

[40] G. Giacinto and F. Roli, "Nearest-prototype relevance feedback for content-based image retrieval," in *Proceedings of the 17th International Conference on Patternt Recognition*, vol. 2, pp. 989–992, Washington, DC, USA, 2004.

[41] G. Giacinto and F. Roli, "Instance-based relevance feedback for image retrieval," in *Advances in Neural Information Processing Systems*, vol. 17, pp. 489–496, MIT Press, 2005.

[42] G. Giacinto and F. Roli, "Instance-based relevance feedback in image retrieval using dissimilarity spaces," in *Case-Based Reason-Ing for Signals and Images*, pp. 419–430, Springer, 2007.

[43] P. H. Gosselin and M. Cord, "Active learing techniques for user interactive systems: application to image retrieval," in *Proceedings of the Workshop Machine Learning Techniques for Processing Multimedia Content*, Bonn, Germany, 2005.

[44] L. Setia, J. Ick, and H. Burkhardt, "SVM-based relevance feedback in image retrieval using invariant feature histograms," in *Proceedings of the IAPR Workshop on Machine Vision Applications*, Tsukuba Science City, Japan, 2005.

[45] Y. Chen, X. Zhou, and T. Huang, "One-class SVM for learning in image retrieval," in *Proceedings of the International Conference on Image Processing*, pp. 34–37, Thessaloniki, Greece, 2001.

[46] Y. Freund, R. Iyer, R. E. Schapire, and Y. Singer, "An efficient boosting algorithm for combining preferences," *Journal of Machine Learning Research*, vol. 4, no. 6, pp. 933–969, 2004.

[47] V. Lavrenko and W. B. Croft, "Relevance-based language models," in *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 120–127, ACM Press, 2001.

[48] G. Winkler, *Image Analysis, Random Fields and Markov Chain Monte Carlo Methods*, Springer Series on Applications of Mathematics, Springer, 2006.

[49] S. Z. Li, *Markov Random Field Modeling in Image Analysis*, Springer, 2nd edition, 2001.

[50] S. Z. Li, "Markov random field models in computer vision," in *Proceedings of the European Conference on Computer Vision*, vol. 801 of *Lecture Notes in Computer Science*, pp. 361–370, Springer, Stockholm, Sweden, 1994.

[51] K. Held, E. Kops, B. Krause, W. Wells III, R. Kikinis, and H. Mueller, "Markov random field segmentation of brain MR images," *IEEE Transactions on Medical Imaging*, vol. 16, no. 6, pp. 878–886, 1997.

[52] S. Geman and D. Geman, "Stochastic relaxation, gibbs distributions, and the bayesian restoration of images," in *Readings in Computer Vision: Issues, Problems, Principles, and Paradigms*, pp. 564–584, 1987.

[53] P. Carbonetto, N. de Freitas, and K. Barnard, "A statistical model for general context object recognition," in *Proceedings of the 8th European Conference on Computer Vision*, vol. 3021 of *Lecture Notes in Computer Science*, pp. 350–362, Springer, Prague, Czech Republic, 2004.

[54] C. Hernandez and L. E. Sucar, "Markov random fields and spatial information to improve automatic image annotation," in *Proceedings of the Pacic-Rim Symposium on Image and Video Technology*, vol. 4872 of *Lecture Notes in Computer Science*, pp. 879–892, Springer, Santiago, Chile, 2007.

[55] 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*, vol. 2, pp. 600–609, Warwick, UK, 2007.

[56] D. Metzler and B. Croft, "A markov random field model for term dependencies," in *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 472–479, ACM Press, 2005.

[57] D. Metzler and W. B. Croft, "Latent concept expansion using Markov random fields," in *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '07)*, pp. 311–318, ACM Press, July 2007.

[58] M. Lease, "An improved markov random field model for supporting verbose queries," in *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '09)*, pp. 476–483, ACM Press, July 2009.

[59] J. Besag, "On the statistical analysis of dirty pictures," *Jounal of the Royal Statistical Society B*, vol. 48, pp. 259–302, 1986.

[60] S. Kirkpatrick, C. Gelatt, and M. Vecchi, "Optimization by simulated annealing," *Science*, vol. 220, no. 4598, pp. 671–680, 1983.

[61] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 11, pp. 1222–1239, 2001.

[62] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[63] E. A. Fox and J. A. Shaw, "Combination of multiple searches," in *Proceedings of The 3rd Text REtrieval Conference (TREC '04)*, NIST Publication, 1994.

[64] H. J. Escalante, J. A. Gonzalez, C. Hernandez et al., "Annotation-based expansion and late fusion of mixed methods for multimedia image retrieval," in *Evaluating Systems for Multilingual and Multimodal Information Access*, vol. 5706 of *Lecture Notes in Computer Science*, pp. 669–676, Springer, 2009.

[65] I. Mani, *Automatic Summarization (Natural Language Processing)*, John Benjamins Publishing Co, 2001.

[66] M. D. Smucker, J. Allan, and B. Carterette, "Agreement among statistical significance tests for information retrieval evaluation at varying sample sizes," in *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '09)*, pp. 630–631, ACM Press, July 2009.

[67] K. G. Kanji, *100 Statistical Tests / Gopal K. Kanji*, Sage, London, UK, 1993.

[68] H. J. Escalante, C. A. Hernández, J. A. Gonzalez et al., "The segmented and annotated IAPR TC-12 benchmark," *Computer Vision and Image Understanding*, vol. 114, no. 4, pp. 419–428, 2010.

[69] C. Snoek, M. Worring, A. Smeulders, and W. M. Arnold, "Early versus late fusion in semantic video analysis," in *Proceedings of the 13th Annual ACM International Conference on Multimedia (MULTIMEDIA '05)*, pp. 399–402, ACM Press, New York, NY, USA, 2005.