Inf Retrieval (2012) 15:1–32
DOI 10.1007/s10791-011-9170-z

Multimodal indexing based on semantic cohesion for image retrieval

Hugo Jair Escalante · Manuel Montes · Enrique Sucar

Received: 17 March 2010 / Accepted: 21 May 2011 / Published online: 5 June 2011

Abstract This paper introduces two novel strategies for representing multimodal images with application to multimedia image retrieval. We consider images that are composed of both text and labels: while text describes the image content at a very high semantic level (e.g., making reference to places, dates or events), labels provide a mid-level description of the image (i.e., in terms of the objects that can be seen in the image). Accordingly, the main assumption of this work is that by combining information from text and labels we can develop very effective retrieval methods. We study standard information fusion techniques for combining both sources of information. However, whereas the performance of such techniques is highly competitive, they cannot capture effectively the content of images. Therefore, we propose two novel representations for multimodal images that attempt to exploit the semantic cohesion among terms from different modalities. Such representations are based on distributional term representations widely used in computational linguistics. Under the considered representations the content of an image is modeled by a distribution of co-occurrences over terms or of occurrences over other images, in such a way that the representation can be considered an expansion of the multimodal terms in the image. We report experimental results using the SAIAPR TC12 benchmark on two sets of topics used in ImageCLEF competitions with manually and automatically generated labels. Experimental results show that the proposed representations outperform significantly both, standard multimodal techniques and unimodal methods. Results on manually assigned labels provide an upper bound in the retrieval performance that can be obtained, whereas results with automatically generated labels are encouraging. The novel representations are able to capture more effectively the content of multimodal images. We emphasize that although we

H. J. Escalante (✉) · M. Montes · E. Sucar
Computer Science Department, National Institute of Astrophysics, Optics and Electronics,
Luis Enrique Erro # 1, 72840 Puebla, Mexico
e-mail: hugojair@inaoep.mx

M. Montes
e-mail: mmontesg@inaoep.mx

E. Sucar
e-mail: esucar@inaoep.mx

have applied our representations to multimedia image retrieval the same formulation can be adopted for modeling other multimodal documents (e.g., videos).

Keywords Multimedia image retrieval · Image annotation · Distributional term representations · Semantic cohesion modeling

1 Introduction

Nowadays images are one of the main sources of information available preceded only by 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.

Image retrieval has been an active research area for over two decades (Smeulders et al. 2000; Goodrum 2000; Datta et al. 2008; Liu et al. 2007; Lew et al. 2006; Rui et al. 1999). However, despite the substantial advances that have been achieved so far, most of the reported work focuses on methods that consider a single information modality (i.e., either image or text), limiting the effectiveness and applicability of such methods. On the one hand, text-based methods are unable to retrieve images that are visually similar to a query image. On the other hand, image-based techniques cannot retrieve relevant images to queries that involve non-visual information (e.g., about places, events or dates). Further, visual methods present additional complications; for example, the need of specifying query images, providing relevance feedback and, more importantly, the ambiguity on determining the underlying user information need from a sample image.

Because of the above limitations, in the last few years there has been an increasing interest from the scientific community in the study and development of retrieval techniques that incorporate both visual and textual information (Liu et al. 2007; Clough et al. 2007; Grubinger et al. 2008; Arni et al. 2009; Westerveld 2004; Aly et al. 2007). Most researchers that adopt the latter approach attempt to exploit the complementariness and diversity of information from different modalities available in multimodal image collections. Despite the fact that such approach seems logical and intuitive, it is not easy to develop multimodal methods that can yield satisfactory retrieval results. Moreover, current techniques fail at exploiting the availability of multimodal information for effectively representing the content of images.

Another current research trend is on the development and usage of automatic image annotation (AIA) techniques for overcoming the limitations of content-based image retrieval (CBIR). The goal of AIA techniques is to provide images with labels so that users can search for images by using keywords (Datta et al. 2008; Barnard et al. 2003, 2008; Jeon et al. 2003; Escalante et al. 2011). Usually, a set of labeled images are used as training set for learning algorithms that attempt to learn a mapping between visual features, extracted from the images, and labels. AIA labels can be considered an intermediate modality between visual and textual information (Hare et al. 2006) and, therefore, it is worthwhile studying the benefits offered by this modality into the multimedia image retrieval task. However, AIA labels have been barely used in combination with other modality for image retrieval (Ah-Pine et al. 2008; Escalante et al. 2008a, 2009). Thus, the usefulness of AIA labels in multimedia image retrieval has not been studied; in this work

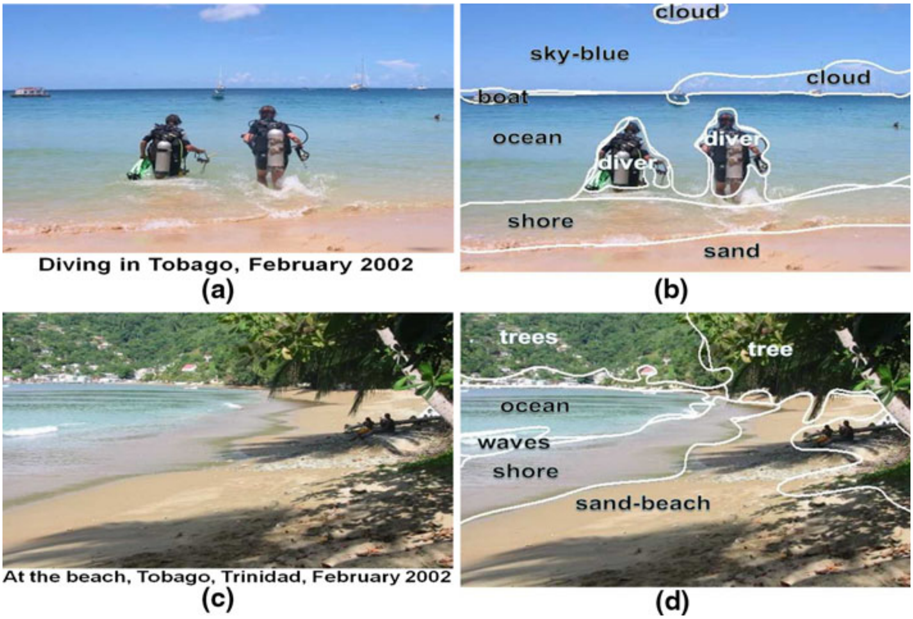

Fig. 1 Illustration of the complementariness between text (**a** and **c**) and labels from AIA (**b** and **d**). Images taken from the SAIAPR TC12 collection (Escalante et al. 2010)

we consider images that are composed of text and AIA labels and study the benefits offered by AIA labels into multimedia image retrieval.

1.1 Our approach

This paper introduces two novel techniques for representing multimodal images with application to multimedia image retrieval. In the considered setting, an image is described by both, textual information assigned to the image as a whole, and labels that are associated with regions in the image (see Fig. 1). The proposed strategies attempt to exploit the semantic cohesion as provided by labels and text for obtaining image representations that can effectively capture their semantic content. The representations we propose are based on distributional term representations (DTRs) in which terms are represented by a distribution of either co-occurrences over terms or occurrences over other images. Accordingly, in our proposal a multimodal document (an image in our case) is represented by a combination of the DTRs corresponding to the terms appearing in the document (i.e., the representation for a document is a distribution itself). In this way the representation for a document captures second order dependencies between terms and documents, which results in better retrieval performance. Also, the proposed representations allow us to search for images by using either labels, text, or combinations of both modalities.

We adopted the vector space retrieval model (VSM) with the proposed representations, although several other retrieval techniques can be used as well. We report experimental results on the SAIAPR TC12[1] collection, a benchmark database for the evaluation of multimedia image retrieval techniques. For comparison we also developed several baseline

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

techniques for combining labels and text. Experimental results show that the baselines are very competitive, outperforming unimodal techniques; however, the novel representations obtain better performance than the standard information-fusion techniques. Furthermore, the proposed strategies are able to capture the content of images more effectively, motivating further research in several aspects.

1.2 Considered scenario

We consider images that are described by text and labels as shown in Fig. 1. Text is manually assigned by users and usually refers to the high-level semantic content of the image; for example, making reference to places, events, dates, etcetera (see Fig. 1a, c). Labels or concepts, on the other hand, are derived from the image itself; they can be assigned manually or generated by AIA techniques and commonly make reference to visual aspects of the image (see Fig. 1b, d); specifically, we consider as labels the names of the objects[2] that can be seen in images.

Since we are interested on studying whether or not the use of AIA labels can offer benefits to multimedia image retrieval, we consider manually assigned labels for most of our experiments. In this way, our results represent an upper bound in the performance that can be obtained with AIA labels. Nevertheless, we also perform experiments with labels generated with a state-of-the art AIA technique.

One should note that up to now the use of labels generated with AIA methods has been restricted to collections that lack any textual description. However, the main hypothesis of this work is that labels can also be helpful for collections that already have an associated textual description. Our hypothesis is based on the fact that both modalities offer information at different semantic levels that is complementary[3] and redundant[4] at the same time (see Fig. 1). Thus, it is reasonable to assume that we can exploit this multimodal information to develop effective multimedia image retrieval methods.

In the following we focus on methods for representing multimodal images that contain terms from two information modalities, namely, text and labels. Whereas this is a very specific scenario, the methods described herein are applicable to multimodal documents described by other types of terms and using more than two modalities. For example, documents described by text and phonetic codes (Reyes et al. 2011), multimodal images described by visual terms (i.e., textons (Winn et al. 2005)), labels and text; or video described by terms from automated speech recognition transcripts, visual terms and concept detectors (Snoek et al. 2006; Larson et al. 2009; Kraaij et al. 2006). Therefore, despite the fact we face a particular multimodal problem, our methods can be used similarly in several domains and applications. In future work we will explore the application of our methods into other multimodal tasks. Note that whereas text and labels belong to the same information media, we consider them as belonging to different modalities because text and labels are obtained from two different sources of information.

[2] Note that we do not adopt the definition of concept from Snoek et al. (2006), which makes reference to concepts of a higher level of semantics (e.g., *sports, people walking, violence, racing*); instead, we adopt the notions from the image retrieval community, see for example the work by Hanbury (2006, 2008) and Grubinger (2007).

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

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

1.3 Contributions

The contributions of this work are summarized as follows:

- We introduce two novel representations for multimodal images based on DTRs and use them for the multimedia image retrieval task. The proposed representations outperform unimodal and standard information-fusion techniques; also, our proposals are able to represent the content of images effectively. Even when we use one of the simplest retrieval models (i.e., the VSM) without any improvement (e.g., query expansion, relevance feedback, etcetera), the performance obtained with our representations is competitive with the best reported results using the same collection we considered. We emphasize that the proposed representations can be used with other types of multimodal documents (e.g., videos) and can be used to combine information from more than two modalities (e.g., text, labels and visual features).
- We propose the combination of labels obtained via AIA with textual descriptions, and show evidence that this approach is a promising solution to the multimedia image retrieval problem. Our results motivate further research on the development of multimedia image retrieval methods based on labels and text.

1.4 Structure of the paper

The rest of this paper is organized as follows. The next section presents related work on multimedia image retrieval. Section 3 describes the multimedia image retrieval setting we consider and presents baseline techniques for combining information from labels and text. Section 4 presents the DTRs we adopt for representing images. Section 5 introduces the novel representations for multimodal documents. Section 6 reports experimental results that show the effectiveness of our methods. Section 7 presents the conclusions derived from this work and outlines future work directions.

2 Related work

The multimedia image retrieval task has been mainly faced with information fusion techniques that attempt to combine visual and textual information (Goodrum 2000; Datta et al. 2008; Rui et al. 1999; Clough et al. 2007; Grubinger et al. 2008; Arni et al. 2009; Westerveld 2004; Escalante et al. 2008a, 2009; Westerveld et al. 2000; Raicu 2006; Escalante et al. 2008b; Chua et al. 1994; Ishikawa et al. 1998; La et al. 1998; Sclaroff et al. 1999; Elworthy 2000; Inoue and Ueda 2005; Hoi et al. 2005; Peinado et al. 2005; Jones and McDonald 2005; Martínez-Fernández et al. 2005, 2006; Adriani and Framadhan 2005; Izquierdo-Beviá et al. 2005; Chang et al. 2005; Martín-Valdivia et al. 2005; Besancon and Millet 2006; Chang and Chen 2006; Curtoni 2006; Maillot et al. 2006; Rahman et al. 2006; Lestari-Paramita et al. 2009). Visual information is incorporated through vectors of visual features that are extracted from images, whereas text is incorporated through vectors of occurrence statistics.

Feature vectors have been usually combined through late fusion (LF) or early fusion (EF) techniques. The LF approach to multimedia image retrieval consists of running several unimodal retrieval methods (either textual or visual) and combining their outputs to obtain a single list of ranked images (documents) per query (Westerveld 2004; Escalante et al. 2008b; Peinado et al. 2005; Izquierdo-Beviá et al. 2005; Besancon and Millet 2006;

Chang and Chen 2006; Rautiainen and Seppdnen 2005; Rautiainen et al. 2004; Snoek et al. 2005). Because of its simplicity and its effectiveness, LF is one of the most used techniques for information fusion in general (Shu and Taska 2005), although a disadvantage of this method is that it may be inefficient as several retrieval methods must be run for each query. Usually a single method is used per modality (Peinado et al. 2005; Izquierdo-Beviá et al. 2005; Besancon and Millet 2006; Chang and Chen 2006; Rautiainen et al. 2004), although the use of multiple and heterogeneous techniques has been also studied (Escalante et al. 2008b).

The EF formulation, on the other hand, consists of merging the vectors corresponding to textual and visual information beforehand and then using a straight retrieval technique (Rautiainen et al. 2004; Rautiainen and Seppdnen 2005; Snoek et al. 2005; Westerveld 2000; van Gemert 2003; Escalante et al. 2009). In its basic form, EF consists of concatenating the vectors of textual and visual features. For example, Cascia et al. and S. Sclaroff et al. concatenated color and texture attributes with the representation for text as obtained with latent semantic indexing (La et al. 1998; Sclaroff et al. 1999). Westerveld and van Gemert adopted a similar approach, although they applied dimensionality reduction techniques to the heterogeneous vectors for compressing the multimodal information (Westerveld 2000; van Gemert 2003). Similar approaches are reported in (Rautiainen et al. 2004; Rautiainen and Seppdnen 2005; Snoek et al. 2005). Compared to LF, EF can be more efficient as a single retrieval stage is performed, however, the dimensionality in which EF methods may work can be huge.

The most popular method for multimedia image retrieval in the last few years is the so called inter-media (pseudo) relevance feedback (IRF) technique (Clough et al. 2007; Grubinger et al. 2008; Ah-Pine et al. 2008; Escalante et al. 2008a, 2009; Chang and Chen 2006; Chang et al. 2005; Lestari-Paramita et al. 2009; Clinchant et al. 2007; Ah-Pine et al. 2009a, b). IRF consists of two retrieval stages using two information modalities, say X and Y, where it is possible that $X = Y$. In the first stage, documents and query are represented under the modality X; the top k_0 retrieved documents are considered to build a new query, which is created by processing the information in modality Y from the k_0 documents. Next, the just created query is used for the second retrieval stage considering documents represented under modality Y. Intuitively, this method switches between modalities with the goal of improving the retrieval performance of unimodal methods. Satisfactory results have been reported with weighted versions of this technique (Clough et al. 2007; Grubinger et al. 2008; Ah-Pine et al. 2008; Chang et al. 2005; Chang and Chen 2006; Clinchant et al. 2007; Ah-Pine et al. 2009a); although it is not clear how to select the initial and final modalities.

Other formulations that have been adopted are based on query expansions (Ishikawa et al. 1998) or incorporate user relevance feedback (Rui et al. 1998; Zhang et al. 2005); other methods have faced the retrieval problem as a supervised (Bradshaw 2000; Grangier et al. 2006; Grangier and Bengio 2006) or semi-supervised learning task (Cox et al. 2000; Zhou et al. 2006). Westerveld proposed a combination of independent Gaussian mixture models (one for each modality) for representing multimodal images; image retrieval was performed by estimating the (combined) probability for a query to be generated by the model of each document (Westerveld 2004). Also, traditional retrieval methods (e.g., VSM or latent semantic indexing) have been modified to incorporate visual information (Westerveld et al. 2000; Raicu 2006; Inoue and Ueda 2005). Satisfactory results have been obtained with the latter methods; however, these approaches are either based on specific assumptions or the data they consider are difficult to obtain, hence the applicability of such methods is limited.

2.1 Discussion

The above described methods present a common limitation: they do not exploit the association among multimodal terms to obtain better representations of images, which also limits the retrieval performance of these methods. Current methods work with the individual modalities separately and in some stage of their processes merge the unimodal information. Thus, in traditional approaches the representation of a multimodal image is always a combination of two unimodal representations. Even though acceptable performance has been obtained with such strategies, we believe that the interactions between terms from different modalities can reveal useful information about the content of images. Therefore, the possible dependencies between terms from different modalities must be considered for representing multimodal images. The methods we propose are based on this idea.

In this work we propose two novel representations that attempt to exploit the multimodal term association with the goal of obtaining better representations for multimodal images. The proposed techniques are based on distributional term representations (DTRs) that have not been previously used for image retrieval. Carrillo et al. 2009 have previously studied the use of DTRs (in combination with other techniques) for textual information retrieval; results reported by Carrillo et al. by using the DTRs alone are rather poor. In this work we were able to obtain acceptable results with DTRs for multimedia image retrieval.

One should note that all of the above referred methods have been designed to combine visual and textual features. However, in this work we develop methods that combine labels and text. Despite the fact that we are implicitly fusing visual (i.e., labels are derived from processed visual data) and textual information, in our formulation images are described with information from a higher semantic level, (see Hare et al. 2006) for a categorization of the semantic levels by which an image can be described). We believe that this approach is advantageous because the information to fuse lies in modalities that are *closer* in their level of semantics (Hare et al. 2006), thus facilitating its combination.

The use of labels for image retrieval has been studied for a while (Barnard et al. 2003, 2008; Jeon et al. 2003; Hare et al. 2006; Allan and Verbeek 2009; Carneiro et al. 2007), however, labels have been mostly used for content-based image retrieval (i.e., a unimodal formulation). There are a few works that have considered the combination of labels and text for image retrieval, see for example (Ah-Pine et al. 2008; Escalante et al. 2008a). In Escalante et al. (2008a) labels automatically assigned to images were used to expand their textual descriptions. To the best of our knowledge this was the first reported evidence on the usefulness of labels for improving the image retrieval performance of text-based methods. Instead of expanding documents, in this work we evaluate the advantages of using labels and text to represent images. More importantly, we propose two novel representations that resulted very effective for multimedia image retrieval. In Ah-Pine et al. (2008), Escalante et al. (2008b) the authors also studied the combination of labels, generated with visual concept detectors, with other information sources, however, results were not encouraging; the latter can be due to the limited content-coverage of the 17 concepts that were considered.

In the video retrieval domain, the use of visual concept detectors has been widely studied in the framework of the TRECVID forum (Kraaij et al. 2006). In that forum visual concept detectors are used to associate video shots with a probability distribution over all of the considered concepts, which indicates how likely is each concept to be present in the shot. Such information has been combined with text obtained by automatic speech recognizers over the audio extracted from the video (Kraaij et al. 2006; Aly et al. 2009). The latter approaches are highly relevant to our work, although, there are several differences to take into account. Firstly, concepts considered in video retrieval are far more general (e.g., *sports, military, explosion*) than those used in image retrieval, even some concepts refer to information that is not visually present in the image (e.g., *monologue, entertainment*) (Snoek et al. 2006). Secondly, concepts in video retrieval are associated to the video frame as a whole (Aly et al. 2007, 2009; Snoek et al. 2006; Kraaij et al. 2006). In contrast, in this work we consider labels attached to regions in images. Thirdly, we consider a scenario in which labels either occur or not in documents, whereas in video retrieval, it is considered a probability distribution over all concepts. In future work we would like to explore the use of our methods for video retrieval tasks.

A highly related work from the video retrieval community is that by Boldareva and Hiemstra, where a probabilistic interactive retrieval approach is proposed (Boldareva and Hiemstra 2004). As one of the DTRs we consider (see Sect. 4), the authors of the latter paper consider an indexing model where documents are represented by its (unimodal) associations with other documents. In the work of Boldareva and Hiemstra associations are probabilities derived from visual similarity measurements, while in this work the associations are given by term co-occurrence statistics. Using co-occurrence statistics makes possible to estimate associations among documents according to terms from different modalities.

3 Baseline methods: combining text and labels

We face the problem of representing images that are composed of terms that belong to one of two different vocabularies. Specifically, we consider multimodal images that can be composed of terms taken from either a vocabulary of labels (L) or a vocabulary of text (T).

As baseline unimodal representation we consider the widely used *tf-idf* indexing scheme where a document is represented by a vector of term occurrence statistics (Salton et al. 1975; Salton and Buckley 1987). Specifically, the ith element of the vector representing document j according to modality X is given by:

$$\mathbf{d}^X_{i,j} = tf^X(i,j) \times \log\left(\frac{N_D}{N^X_i}\right) \tag{1}$$

where $tf^X(i, j)$ is the number of occurrences of term i from the modality X in document j; N_D is the number of documents in the collection and N^X_i denote the number of documents in which term i appears. Vectors are normalized using cosine normalization: $\mathbf{d}^X_{i,j} = \frac{\mathbf{d}^X_{i,j}}{\sum_{k=1}^{|V|}(\mathbf{d}^X_{i,k})^2}, i = 1, \ldots, |V|$, see Salton and Buckley (1987). Queries are represented in a similar way. Thus, $\mathbf{q}^X$ denotes the corresponding *tf-idf* representation for a query in modality X. For retrieving documents the query vector and every document in the collection are compared via the cosine similarity:

$$S^X(\mathbf{q}^X, \mathbf{d}^X_j) = \frac{\mathbf{q}^X \cdot \mathbf{d}^X_j}{||\mathbf{q}^X||||\mathbf{d}^X_j||} \tag{2}$$

documents are sorted in descending order of $S^X(\mathbf{q}^X, \mathbf{d}^X_j)$ and the top $-k$ are shown to the user.

The above representation can be used to search images by using a single modality: either text ($X = T$) or labels ($X = L$). For combining both sources of information we consider the three most frequently used techniques for information fusion in information retrieval, namely: late fusion, early fusion and inter-media relevance feedback. These techniques are not new, however, this is the first work in which such techniques are used to combine information from labels and text. In the rest of this section we describe the standard information fusion techniques used.

3.1 Late fusion (LF)

Under this approach the outputs of multiple unimodal retrieval models (usually from different modalities) are combined to obtain a new ranking of the documents for each query. LF has reported satisfactory performance in this task by combining the outputs of visual and textual methods; however, it can be inefficient because it requires running multiple retrieval models for each query.

For combining labels and text we consider the following LF formulation. Given queries in both modalities, we retrieve documents separately by using the unimodal VSM approach. The result of this process is two lists of ranked documents, l^L and l^T, corresponding to documents retrieved with labels and text, respectively. Each document $\mathbf{d}_j$ that appears in the union of lists l^T and l^L, is assigned a score that depends on the position in the lists in which the document appears and on the number of lists in which the document appears. The weight for each document is thus assigned as follows:

$$\Psi(\mathbf{d}_j) = m \times \left(\alpha_L \times h(l^L, \mathbf{d}_j) + \alpha_T \times h(l^T, \mathbf{d}_j)\right) \tag{3}$$

where $\Psi(\mathbf{d}_j)$ is the relevance weight for document j under LF; $h(l^X, \mathbf{d}_y)$ is the inverse of the position of document $\mathbf{d}_y$ in list l^X, by definition $h(l^X, \mathbf{d}_y) = 0$, if $\mathbf{d}_y$ does not appear in $l^X; m \in \{1,2\}$ is the number of lists in which document $\mathbf{d}_j$ appears; α_L and α_T are scalars that weight the contribution of each of the considered modalities. The weight in Eq. (3) is computed for every document and the ranked documents (in descending order) are shown to the user. This form of fusion resembles the CombMNZ method (Fox and Shaw 1994), one of the most widely used in information retrieval (Shu and Taska 2005; Farah and Vanderpooten 2007). Where we assume that the inverse of the position of a document in a list is proportional to the similarity of the document to the corresponding query. We consider this specific formulation because in previous work we have obtained satisfactory results in multimedia image retrieval with it (Escalante et al. 2008b, 2009).

3.2 Early fusion (EF)

Under this scheme, the unimodal representations in both modalities as obtained with Eq. (1) are concatenated as follows:

$$\mathbf{d}_j^{LT} = [\alpha_L \times \mathbf{d}_j^L, \alpha_T \times \mathbf{d}_j^T] \tag{4}$$

where $\mathbf{d}_j^{LT}$ is the merged representation of document $\mathbf{d}_j$, α_L and α_T are scalars that weight the contribution of each modality. Documents and queries are represented as described in Eq. (4). Then the cosine distance is used to rank documents. One should note that in order to obtain satisfactory results with EF under the VSM, the representations for the document in each modality should be *comparable* (i.e., they must account for similar statistics and must be properly normalized); otherwise, Eq. (2) may be uninformative for computing similarity.

3.3 Inter-media relevance feedback (IRF)

While LF and EF can be considered traditional approaches to multimedia image retrieval, IRF is a recently proposed strategy that has proved to be very effective (Chang et al. 2005; Ah-Pine et al. 2009b). IRF is based on the widely adopted technique of relevance feedback (Rui et al. 1998), where users mark relevant images after an initial retrieval stage and using such information a second retrieval stage attempts to refine the initial results. The main difference in IRF is that it uses different modalities for the initial and final retrieval stages (e.g., images and text). Commonly, a pseudo-relevance feedback formulation is adopted (Chang et al. 2005; Ah-Pine et al. 2009b), where instead of requiring user interaction, the top k_0 documents are considered relevant to the query. Accordingly, in this work we consider the pseudo-relevance feedback version of IRF, which will be referred to as IRF in the rest of the paper.

The very first step in this formulation is to define the *initial* and *final* modalities, each of which can be either labels or text. Then, using the documents and the query under the initial modality, documents are ranked according to the unimodal VSM; where the top k_0 documents are kept (usually $k_0 \leq 20$). Next, we extract all of the terms from the final modality that occur in the k_0 documents and create a query using such information; we denote such query with $\mathbf{q}^C$; next, the just created query is combined with the query in the final modality (as provided by the user), we denote the query in the final modality with $\mathbf{q}^F$. Then we create an extended query by combining the queries as follows:

$$\mathbf{q}^E = \alpha_C \times \mathbf{q}^C + \alpha_F \times \mathbf{q}^F \tag{5}$$

where α_C and α_F are scalars weighting the contribution of queries $\mathbf{q}^C$ and $\mathbf{q}^F$, respectively. Finally, we use the unimodal VSM, using $\mathbf{q}^E$ as query and documents represented under the final modality, to rank the documents in the collection; see Chang et al. (2005), Ah-Pine et al. (2009b) for further details on the IRF approach. One should note that under the IRF formulation the same modality can be used for both retrieval stages (e.g., text-text or labels-labels), in which case the IRF is just a straight pseudo-relevance feedback technique. In Sect. 6 we report experimental results with this unimodal formulation as well. We used $k_0 = 5$ in our experiments based on preliminary experimentation with several values for k_0 in Smeulders et al. (2000), Snoek et al. (2006).

We have just described standard information-fusion techniques for combining labels and text. It is clear that although those methods take into account information from both modalities, they do not explicitly take advantage of the association among terms from the different modalities, which we believe can be helpful for better modeling the content of images. The next section describes two new ways of representing multimodal images that indeed exploit such multimodal term association.

4 Distributional term representations

Distributional term representations (DTRs) are tools for term representation that rely on term occurrence and co-occurrence statistics (Lavelli et al. 2005). The intuition behind DTRs is that the meaning of a term can be deduced by its context; where the context for a term is determined by the other terms it co-occurs with frequently or by the documents in which the term occurs more frequently.

DTRs have been mostly used in computational linguistics for tasks that include term processing; for example, for term clustering (Lewis and Croft 1990), automatic thesaurus construction (Chen et al. 1995) and word-sense disambiguation (Gale et al. 1993). Little work has been reported on DTRs for (unimodal) information retrieval (Carrillo et al. 2009; Lavelli et al. 2005). In the latter field, DTRs have been only used for processing unimodal information (e.g., text) where the representation of a term is determined by its context in unimodal information. The novelty of our approach is that we study the use of DTRs for representing terms that belong to different modalities. Under our formulation the meaning of a term depends on statistics derived from both unimodal and multimodal information, which result in richer term representations and, in consequence, in better document representations.

We explore two different DTRs for term representation that are subsequently used for representing multimodal images. The rest of this section presents the specific DTRs we consider and the next section describes how multimodal terms can be represented under these DTRs. For our description, we consider documents that are composed by terms from a particular vocabulary M with $|M|$ terms, for now we will not assume that M is associated to any specific modality.

4.1 Document occurrence representation (DOR)

The document occurrence representation (DOR) is considered the dual of the *tf-idf* representation, see Eq. (1), that is used for representing documents: since documents can be represented by a distribution over the terms, terms can be represented by a distribution over documents. More formally, each term $t_j \in M$ is represented by a vector[5] of weights $\mathbf{w}_j^{dor} = <w_{j,1}^{dor}, \ldots, w_{j,N}^{dor}>$, where N is the number of documents in the collection and $0 \leq w_{j,k}^{dor} \leq 1$ represents the contribution of document $\mathbf{d}_k$ to the representation of t_j. Specifically, we consider the following weighting scheme (Lavelli et al. 2005):

$$\mathbf{w}^{dor}(t_j, \mathbf{d}_k) = df(t_j, \mathbf{d}_k) \times \log\left(\frac{|M|}{N_k}\right) \tag{6}$$

where N_k is the number of different terms that appear in document $\mathbf{d}_k$ and $df(t_j, \mathbf{d}_k)$ is given by:

$$df(t_j, \mathbf{d}_k) = \begin{cases} 1 + log\big(\#(t_j, \mathbf{d}_k)\big) & \text{if } \#(t_j, \mathbf{d}_k) > 0 \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

where $\#(t_j, \mathbf{d}_k)$ denotes the number of times term t_j occurs in document $\mathbf{d}_k$. The weights are normalized using cosine normalization. Intuitively, the more frequent the term t_j occurs in document $\mathbf{d}_k$, the more important $\mathbf{d}_k$ is to characterize the semantics of t_j; on the other hand, the more different terms occur in $\mathbf{d}_k$, the less it contributes to characterize the semantics of t_j.

4.2 Term co-occurrence representation (TCOR)

The term co-occurrence representation (TCOR) is similar to DOR, although it is based on the idea of representing a term t_j according to a distribution over the terms it co-occurs with. Under TCOR each term $t_j \in M$ is represented by a vector of weights $\mathbf{w}_j^{tcor} = <w_{j,1}^{tcor}, \ldots, w_{j,|M|}^{tcor}>$, where $0 \leq w_{j,k}^{tcor} \leq 1$ represents the contribution of term t_k in the

[5] With abuse of notation we will use both $w_{i,j}^{x}$ and $\mathbf{w}^{x}$ $(t_i, \mathbf{d}_j)$ to make reference to the jth element of vector $\mathbf{w}_i^{x}$.

TCOR representation of t_j. Specifically, we consider the following weighting scheme (Lavelli et al. 2005):

$$\mathbf{w}^{tcor}(t_j, t_k) = ttf(t_j, t_k) \times \log\left(\frac{|M|}{N_k}\right) \tag{8}$$

where N_k is the number of terms in M that co-occur with t_j in at least one document and $ttf(t_j, t_k)$ is given by:

$$ttf(t_j, t_k) = \begin{cases} 1 + log\big(\#(t_j, t_k)\big) & \text{if}\#(t_j, t_k) > 0 \\ 0 & \text{Otherwise} \end{cases} \tag{9}$$

where $\#(t_j, t_k)$ indicates the number of documents in which t_k and t_j co-occur. The intuition of TCOR is that the more t_k and t_j co-occur, the more important t_k is for describing term t_j; on the other hand, the more the number of terms that co-occur with t_k, the less it will contribute to characterize the semantics of t_j.

5 Semantic cohesion modeling

This section introduces two strategies for representing multimodal documents that attempt to capture the semantic cohesion between terms from different modalities. For both strategies we first represent multimodal terms according to DOR or TCOR. Next, for representing a document we combine the representations of the terms that occur in the document. Accordingly, we first describe how to represent multimodal terms with the DTRs and then we introduce the strategies for representing documents. For all of the illustrations and descriptions of this section we consider the multimodal documents (i.e., annotated images) from the SAIAPR TC12 collection that we used for the evaluation of our methods; see Sect. 6 for details.

5.1 Multimodal distributional term representations

Building DTRs for multimodal terms is rather simple because in the above description of DTRs we have made no assumption on the modality of the vocabulary M by which documents are composed. Hence, if we consider multimodal documents that contain terms from k different modalities with vocabularies $X_1, \ldots, X_k$, it is enough to define the vocabulary M as the union of the available vocabularies, that is $M = \{X_1, \ldots, X_k\}$; next, we just apply the corresponding weighting schemes described in Eqs. (6) and (8) by equally accounting for both unimodal and multimodal statistics. In this way the representation for a term will be determined by its semantic cohesion with other terms, which can belong to the same modality or not. In the rest of this section we describe the benefits of representing labels and textual terms according to DTRs and show how multimodal images can be represented by combining DTRs. Despite we have considered images, all of the methods we propose are applicable to other types of multimodal documents.

5.1.1 Multimodal DOR

Recall that in the multimedia image retrieval scenario we consider the multimodal terms are words (text) and concepts (labels). Hence, the multimodal DOR representations (MDOR) for labels and words are distributions of occurrences over the documents in the

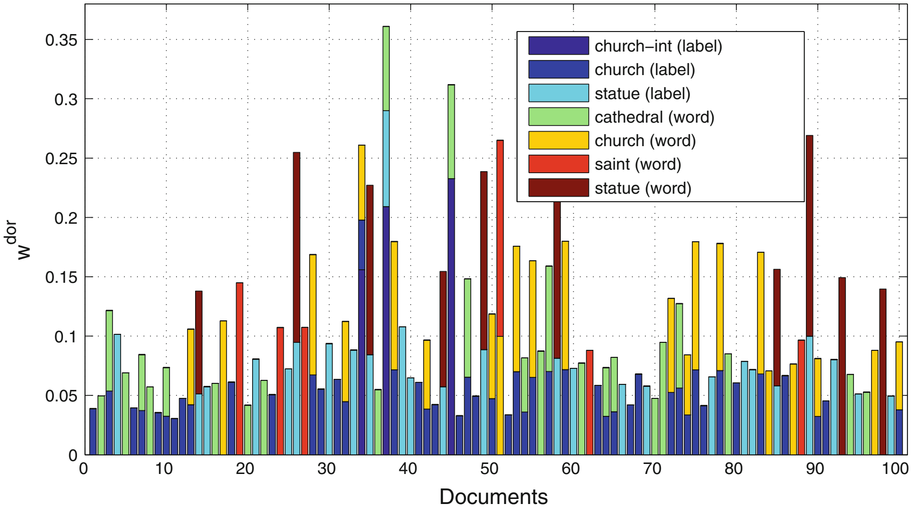

Fig. 2 MDOR representations for the labels *church-interior* (*dark-blue*), *church* (*blue*) and *statue* (*light-blue*) and the words *cathedral* (*green*), *church* (*orange*), *saint* (*red*) and *statue* (*dark-red*). For clarity, we show the value of w^{dor} for 100 documents randomly selected. The *x*-axis shows the documents in which the terms occur at least once, while the *y-axis* indicates the value of the MDOR representation for the terms at the corresponding documents (Color figure online)

collection in which such terms occur. Since multimodal images are composed of very few words and labels (Goodrum 2000; Grubinger 2007; Escalante et al. 2010), the MDOR representations for multimodal terms can be sparse distributions. For example, Fig. 2 shows (part of) the MDOR representations for the labels *church-interior, church*, and *statue*, and the words *cathedral, church, saint*, and *statue*, seven multimodal terms that are semantically related. The MDOR representation for those terms are based on 2,586 documents out of the 20,000 that compose the SAIAPR TC12 collection; that is, 12.93% of the documents contribute to the MDOR representation for the seven terms.

From Fig. 2 we can see that MDOR provides reliable information for characterizing the semantics of a term: as the considered terms are semantically related they co-occur in similar documents and hence they have similar MDOR representations. One should note that the words *cathedral* and *church* do not co-occur in similar documents, despite they hold synonymy and hypernym relations (Miller et al. 1990). This lack of co-occurrence can be due to the short length of the annotations, which is a rather common issue in image retrieval. Hence a user searching for a religious building that uses as query the word *church* would not retrieve images that are associated with the word *cathedral*. However, when representing documents with MDOR (e.g., by adding the MDOR representations of terms that occur in the document), documents containing either word *church* or *cathedral* will have similar representations because of the fact that both words co-occur with the label *church*; hence the label *church* acts as a link between two words that never occur in common documents.

The latter *expansion* cannot be exploited in unimodal DOR, because usually terms that are semantically related (e.g., synonyms and terms under an is-a relationship) are not used together in a document; that is why *church* and *cathedral* do not co-occur in common images. Nevertheless, related words may co-occur with common terms from other

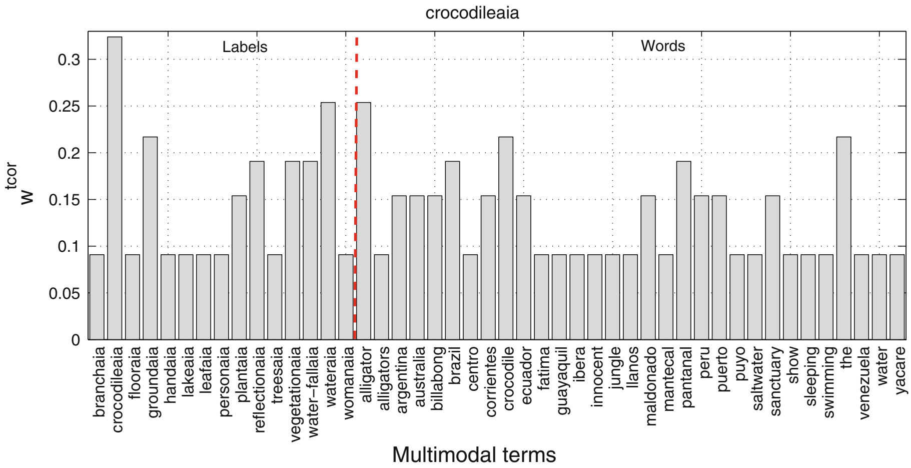

Fig. 3 MTCOR representation for the label term *crocodile*. The *vertical line* separates label terms from textual terms. We only show the distribution over terms that co-occur at least once with *crocodile*

modalities, thus the incorporation of more than one modality is indeed providing additional useful information that is effectively exploited by MDOR.

5.1.2 Multimodal TCOR

The multimodal TCOR (MTCOR) representations of words and labels are distributions of co-occurrences over the terms in the multimodal vocabulary; hence a term is represented through its multimodal context. For example, Fig. 3 shows the MTCOR representation for the label *crocodile*, whereas Fig. 4 shows the MTCOR representation for the word *igloo*; we have added the postfix *"aia"* to labels to differentiate them from words in the textual vocabulary.[6]

From Figs. 3 and 4 we can see that the MTCOR distributions of *crocodile* and *igloo* are determined by those terms that are more semantically related to them (i.e., their context). For example, for *igloo* the more associated labels are *ice, sky, snow, trees* and *wall*, whereas the most related words are *angara, frozen, igloo, irkutsk, river* and *Russia*. On the other hand, for the label *crocodile*, the most associated labels are *crocodile, water, ground, vegetation*, and *plant*, and the most related words are *alligator, crocodile, pantanal*, and *Brazil*. Therefore, MTCOR can be considered a multimodal expansion of the original term and, of course, a multimodal expansion is more informative than a unimodal one (as straight TCOR).

When representing multimodal images with MTCOR, documents are represented by the multimodal expansion of the terms in the document and not by the terms themselves. Thus, second order (multimodal) relationships are implicitly taken into account for document representation, which result in richer document representations. Also, as queries are represented similarly, a multimodal query expansion is performed implicitly under MTCOR.

[6] Note that we have not removed stop-words (e.g., *the*) in order to evaluate the robustness of our method to such useless words (see Sect. 5.2). Also, we did not apply word stemming in order to facilitate the estimation of co-occurrences. However, the same setting was used for every method we evaluated in this work.

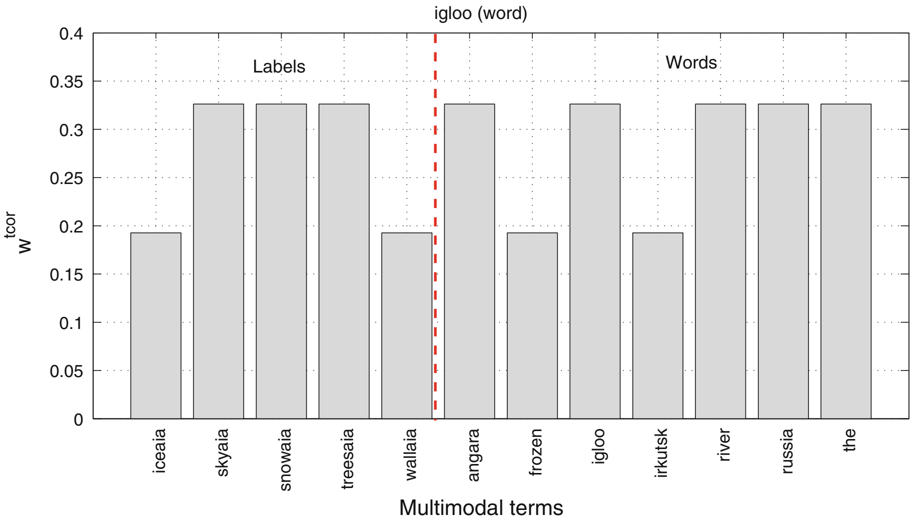

Fig. 4 MTCOR representation for the textual term *igloo*. The *vertical line* separates label terms from textual terms. We only show the distribution over the terms that occur at least once with *igloo*

5.2 Representing images through MDOR and MTCOR

In this work we attempt to exploit both MDOR and MTCOR to represent documents for multimedia image retrieval. A way of representing documents using MDOR and MTCOR is by adding the representations of the terms that occur in the document; we denote this strategy MTCOR-B and MDOR-B, for MTCOR and MDOR, respectively, where the "B" is for *basic* representation. Thus, if document $\mathbf{d}_i$ contains the following N_i terms: $\{t_1^i, \ldots, t_{N_i}^i\}$, with representations MTCOR and MDOR: $\{\mathbf{w}_i^{tcor}, \ldots, \mathbf{w}_{N_i}^{tcor}\}$ and $\{\mathbf{w}_i^{dor}, \ldots, \mathbf{w}_{N_i}^{dor}\}$, then, the corresponding MTCOR-B and MDOR-B representations for $\mathbf{d}_i$ are given by:

$$\mathbf{d}_i^{mtcor-b} = \sum_{k=1}^{N_i} \mathbf{w}_k^{tcor} \tag{10}$$

$$\mathbf{d}_i^{mdor-b} = \sum_{k=1}^{N_i} \mathbf{w}_k^{dor} \tag{11}$$

where the dimensionality of $\mathbf{d}_i^{mtcor-b}$ is $|M|$, the number of terms in the vocabulary and that of $\mathbf{d}_i^{mdor-b}$ is N, the number of documents in the collection. Both vectors $\mathbf{d}_i^{mtcor-b}$ and $\mathbf{d}_i^{mdor-b}$ are normalized using cosine normalization.

Despite the fact that MDOR-B and MTCOR-B can reflect the content of documents they present some limitations. For example, they equally account for the occurrence of both common and uncommon terms (e.g., see the influence of the article *the* in Figs. 3 and 4), which is related to the importance of terms across the collection; also MDOR-B and MTCOR-B do not account for multiple occurrences of particular terms in documents, which reflects the importance of a term in a particular document. For alleviating the latter issues, we consider an alternative representation that is equivalent to the *tf-idf* representation used in information retrieval (Salton et al. 1975). The corresponding representations are as follows:

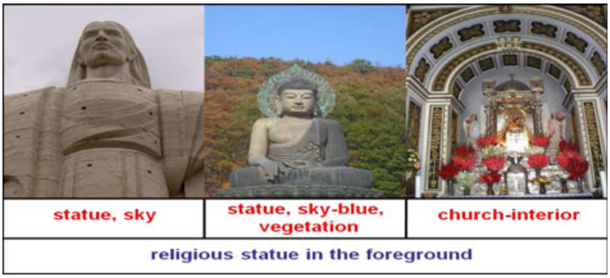
statue, sky
statue, sky-blue, vegetation
church-interior
religious statue in the foreground

Fig. 5 Sample topic as considered in our experiments. Each topic is composed of a free text description and three sample images. For our experiments we manually annotated the query images and used such labels as query for the labels modality

$$\mathbf{d}_i^{mdor-tfidf} = \sum_{k=1}^{N_i} (tf_m(i, t_i^k) \times \mathbf{w}_k^{dor}) \times \left(\log \left(\frac{N}{N_k^m} \right) \right) \tag{12}$$

$$\mathbf{d}_i^{mtcor-tfidf} = \sum_{k=1}^{N_i} (tf_m(i, t_i^k) \times \mathbf{w}_k^{tcor}) \times \left(\log \left(\frac{N}{N_k^m} \right) \right) \tag{13}$$

where $tf_m(i, t_i^k)$ is the number of times that term t_i^k appears in document i and N_k^m is the number of documents in the collection in which term t_i^k appears. Thus, the representations MDOR-TF-IDF (Eq. (12)) and MTCOR-TF-IDF (Eq. (13)) take into account the frequency of occurrence of terms and their usage across the collection.

For retrieving documents under the above representations we represent multimodal queries in a similar way. Then we compute the cosine similarity between query and documents and sort documents according to their similarity to the query.

Besides providing richer representations, an advantage of representations based on MDOR and MTCOR is that queries can be formulated by using terms from either single modality; that is, a query can be a set of words (i.e., textual-only query), a set of labels (i.e., labels-only query) or a combination of words and labels (multimodal query). This is in contrast to common image retrieval techniques (e.g., LF, EF and IRF) where no multimodal information is considered for image retrieval when using unimodal queries.

Under MDOR and MTCOR the representation of a unimodal query is a multimodal expansion of the unimodal terms; therefore, independently of whether queries are multimodal or not, the representation will always be based on multimodal information. For example, consider the query shown in Fig. 5, taken from the ImageCLEF[7] forum, which is composed of the textual terms: *"religious statue in the foreground"* and of the label terms *"church-interior, statue, sky, vegetation"*.

Figures 6 and 7 show the MTCOR-TF-IDF representations for such query; Fig. 6 shows the MTCOR-TF-IDF representation for the textual part of the query, whereas Fig. 7 shows the labels part of the query. From these figures we can see that, even though the queries are unimodal, the corresponding MTCOR-TF-IDF representations capture important multimodal terms that can be useful for retrieving documents.

One should note that documents that have information in only one modality can still be represented with MDOR and MTCOR based representations. This is advantageous on

[7] http://imageclef.org/.

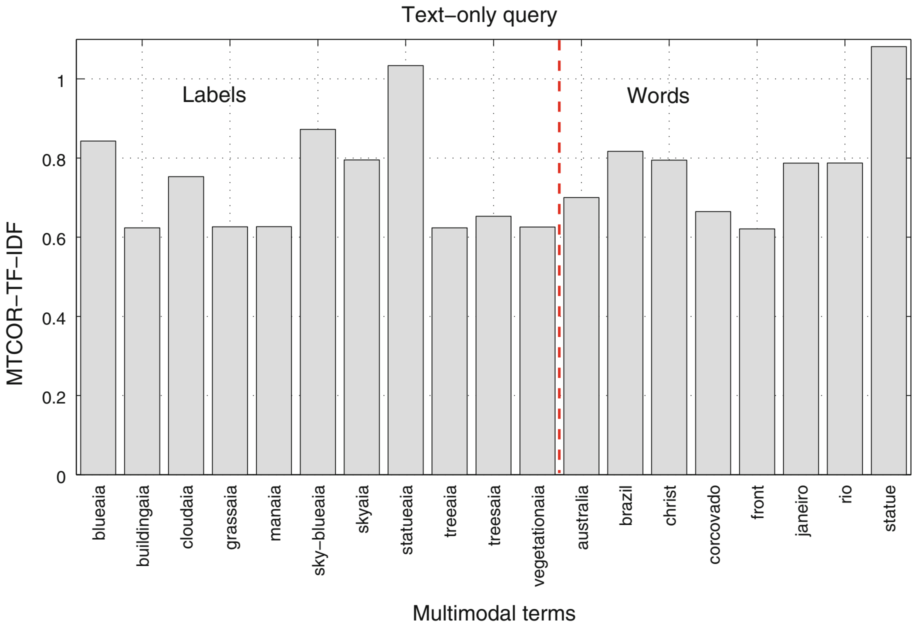

Fig. 6 MTCOR-TF-IDF representation for the query text *"religious statue in the foreground"*. We only show the distribution over the top 20 multimodal terms

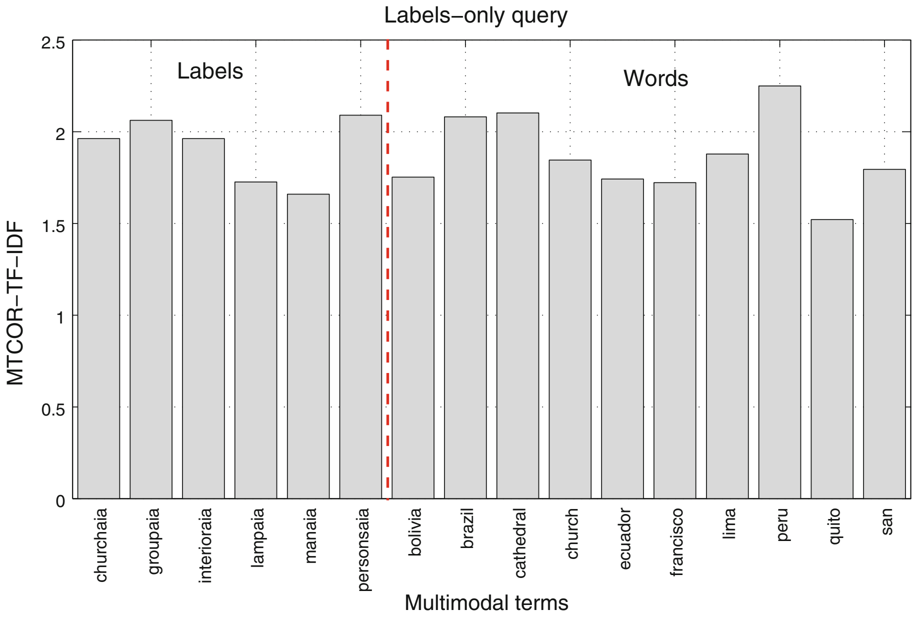

Fig. 7 MTCOR-TF-IDF representation for the query labels *"church, interior, statue"*. We only show the distribution over the top 18 multimodal terms

realistic settings because it allows us to index multimodal documents with mixed sources of information.

6 Experimental results

In this section we evaluate the multimodal document representations described in Sect. 5. The goal is to show the benefits offered by multimodal distribution term representations (DTRs) over standard methods. For all of our experiments we used the SAIAPR TC12 benchmark (Escalante et al. 2010). Sample images are shown in Fig. 1. This collection is composed of 20,000 images that have been manually annotated at image level with free text descriptions (Grubinger 2007); additionally, images have been segmented and manually annotated with a predefined vocabulary of 255 labels (Escalante et al. 2010). Thus, multimodal documents in the SAIAPR TC12 collection are composed of (1) text accompanying images, and (2) labels assigned to regions in the corresponding images.

Since we are interested in evaluating the benefits of using labels into multimedia image retrieval, we considered manually assigned labels for most of our experiments. In this way, our results represent an upper bound in the performance that can be obtained with automatic image annotation (AIA) labels. However, we also performed experiments with labels generated with a state-of-the art technique for AIA. The latter is a more realistic setting as most image collections lack labels. There our goal is to show that automatic labels can be helpful for image retrieval despite the limited performance of current AIA methods.

For evaluating the retrieval effectiveness of our methods we considered two sets of topics as used in the photographic retrieval task at the ImageCLEF2007 (60 topics) and ImageCLEF2008 (39 topics) forums (Grubinger et al. 2008; Arni et al. 2009). Each topic is composed of a fragment of text and three query images. Figure 5 shows a sample topic. For all of our experiments we manually assigned labels to query images, which were considered as query terms for the labels modality. The labels that we used for each topic are available from the website of our research group.[8]

The acronyms for considered techniques are summarized in Table 1.

The considered scenario is as follows. For each retrieval technique (i.e., LF, EF, IRF, MDOR-B, MDOR-TF-IDF, MTCOR-B, MTCOR-TF-IDF) and for each set of topics (i.e., ImageCLEF2007 or ImageCLEF2008) we retrieve relevant documents to each topic. Next, using the corresponding relevance judgements, we evaluate the retrieval effectiveness for the considered methods for each topic. We report average results over the topics for each of the following evaluation measures: mean-average precision (**MAP**, our leading evaluation measure), precision at 20 documents (**P20**), recall at 20 documents (**R20**) and the number of relevant retrieved documents (**RR**). In the following we will refer to the paired t-student test with a confidence level of 95% percent when mentioning statistical significance between retrieval results. We consider this test due to its widely adopted use in the evaluation of information retrieval methods (Smucker et al. 2009).

We have divided the experimental results into four sections that aim to evaluate different aspects of our methods: (1) first we evaluate the performance of baseline methods; (2) next we assess the effectiveness of representations based on multimodal DTRs; (3) then we show how such representations can give support to unimodal queries; (4) finally, we evaluate the effectiveness of the developed techniques when using automatically generated labels.

[8] http://ccc.inaoep.mx/~tia/

6.1 Results with baseline methods

Table 2 shows the retrieval results obtained with the baseline techniques (LF, EF and IRF), we also show the performance obtained by unimodal retrieval models (VSM with *tf-idf* weighting). For LF and EF we report the performance obtained when a similar weight is

Table 1 Summary of acronyms for the retrieval techniques that are compared

Acronym	Description	Section
Text	VSM with textual information	3
Labels	VSM with labels information	3
LF	Late fusion	3.1
EF	Early fusion	3.2
IRF	Intermedia relevance feedback	3.3
MDOR-B	Multimodal document occurrence (basic) representation	5.2
MDOR-TF-IDF	Multimodal document occurrence (term-frequency, inverse document frequency) representation	5.2
MTCOR-B	Multimodal term co-occurrence (basic) representation	5.2
MTCOR-TF-IDF	Multimodal term co-occurrence (term-frequency, inverse document frequency) representation	5.2

The postfix **'-W'** indicates that different weights have been used for different modalities. Whereas the postfix **'-U'** indicates that both modalities were assigned equal weights

Table 2 Retrieval performance obtained with unimodal methods, using LF, EF and IRF

Desc.	Weights		ImageCLEF2007				ImageCLEF2008			
	α_L	α_T	MAP	P20	R20	RR	MAP	P20	R20	RR
Individual performance										
Labels	1	0	0.0587	0.1417	0.1066	1,201	0.053	0.141	0.1133	727
Text	0	1	0.1241	0.1767	0.1694	1,424	0.1033	0.1795	0.1534	1,014
Late fusion										
LF-U	1	1	0.1273	0.185	0.1817	1,731	0.106	0.191	0.1683	1,138
LF-W	1	2	0.1348	0.1858	0.1879	1,703	0.1126	0.1936	0.1759	1,139
Early fusion										
EF-U	1	1	0.1276	0.2475	0.2498	1,722	0.1167	**0.2551**	0.2342	1,044
EF-W	**1**	**8**	**0.189**	**0.2508**	**0.2996**	**2,226**	**0.1565**	0.2372	**0.2695**	**1416**
Inter-media relevance feedback										
IRF-1	l	l	0.0841	0.165	0.136	1,273	0.0714	0.1577	0.123	706
IRF-2	l	t	0.1659	0.2142	0.262	1,952	0.1326	0.1987	0.2205	1,302
IRF-3	t	t	0.1435	0.1867	0.2029	1,717	0.1253	0.1974	0.2009	1,255
IRF-4	t	l	0.055	0.1008	0.0884	1,227	0.0484	0.1077	0.0847	774

For individual methods, LF and EF the columns 2 and 3 show the weights assigned to each modality; whereas for IRF they show the initial (column 2) and final (column 3) modalities, *t* is for the textual modality and *l* is for the labels modality. For EF and LF we show the result when assigning equal weights to each modality (U) and the best obtained result when we modified the corresponding weights (W); for IRF we only show the best result. The best result of each column is shown in bold used for both modalities (U) and when such weights are optimized (W, we performed several trial and error evaluations under a bijective search approach, at most 10 parameter settings were tried); for IRF we only show results obtained with the best configuration ($\alpha_f = 5$ and $\alpha_c = 1$) as results with equal weights were very poor.

The best individual results in all measures were obtained when only text was used (see Text in Table 2), which is not surprising as the considered topics include many queries with high semantic level. Nevertheless, the combination of labels and text outperformed the textual techniques under all of the considered baseline methods; the differences in **MAP** between the best result of each baseline technique and the textual method were statistically significant for both sets of topics.

The best results overall the evaluation measures was obtained with EF-W/ This method obtained a relative improvement of 53.59% in **MAP** over the textual method on average over both sets of topics, although it uses a weight for the textual modality that is 8-times superior than that assigned to the labels modality. LF-W obtained a relative gain of 10.02%; thus, labels were not very helpful for this strategy.

For LF-W, a weighting ratio of 1:2 yielded an improvement of 7.69% over the equal-weighting scheme (LF-U), in terms of MAP over the two sets of topics; while for EF-W the ratio 1:8 resulted in a corresponding improvement of 41.10%. The ratios are not comparable because weights α_T and α_L have different meanings for LF-W and EF-W: in LF-W, α_T and α_L weight the scores of individual VSMs based on a single modality, while in EF-W the weights have an impact in the features (terms) from the different modalities when estimating the similarity among documents. The difference in improvement can be explained by the fact that the VSM based on labels obtained low retrieval performance. Hence, combining its outputs with those of the textual VSM resulted in a limited improvement for LF-W. On the other hand, information from label occurrence resulted very helpful for EF-W: by adjusting the weights for the different features we can obtain significant improvements. Thus evidencing the fact that labels can be very helpful for image retrieval, although it is not clear what is the best way of incorporating them.

The IRF-2 setting improved by 32.51% the **MAP** obtained by the best unimodal technique. It is interesting that the best result was obtained when the initial modality was labels and the final was text (IRF-2 in Table 2). This result is counterintuitive as the individual results of the unimodal technique based on labels are rather poor. This can be due to the fact that despite such individual method obtained low performance in **MAP, R20** and **RR**, the precision (**P20**) obtained by such technique was much more competitive. Thus, the first documents retrieved by using only labels can be relevant with high probability. The latter fact combined with a high weighting to the final query ($\alpha_f = 5$) make that this configuration obtained the best results for IRF. Results obtained with IRF when using the same modality as initial and final modalities (IRF-1 and IRF3 in Table 2) improve the corresponding unimodal techniques, confirming the effectiveness of the straight pseudo-relevance feedback formulation.

Summarizing, the results in Table 2 show that the use of labels can be very helpful for improving the results of unimodal techniques based on either text or labels. The improvement offered by the multimodal baselines was statistically significant and the corresponding results are very competitive.

6.2 Results with multimodal distributional representations

In this section we report the performance obtained with representations based on MDOR and MTCOR as described in Sect. 5; we compare those representations with the best result

Table 3 Retrieval results when using representations based on MDOR and MTCOR

Topics	ImageCLEF2007				ImageCLEF2008			
Method	MAP	P20	R20	RR	MAP	P20	R20	RR
MDOR-B	0.1751	**0.2667**	0.2754	1,975	0.1358	0.241	0.231	1,218
MDOR-TF-IDF	**0.2102**	0.2467	**0.3231**	**2,521**	**0.1954**	**0.2705**	**0.3217**	**1,686**
MTCOR-B	0.1518	0.2492	0.2484	1,699	0.1328	0.2359	0.2216	1,109
MTCOR-TF-IDF	0.1844	0.2383	0.2807	2,256	0.1647	0.2397	0.2688	1,473
LF-W	0.1348	0.1858	0.1879	1,703	0.1126	0.1936	0.1759	1,139
EF-W	0.189	0.2508	0.2996	2,226	0.1565	0.2372	0.2695	1,416
IRF-2	0.1659	0.2142	0.262	1,952	0.1326	0.1987	0.2205	1,302
Labels	0.0587	0.1417	0.1066	1,201	0.053	0.141	0.1133	727
Text	0.1241	0.1767	0.1694	1,424	0.1033	0.1795	0.1534	1,014

We also show the best results obtained with baseline techniques and results obtained with individual methods that only use labels (IL) and text (IT). The best result of each column is shown in bold

obtained with baseline techniques (see Table 2). Results obtained with multimodal distributional representations on both versions (i.e., B and TF-IDF) are shown in Table 3.

This table shows interesting results. First, both types of representations, MDOR and MTCOR, in both versions, B and TF-IDF, obtain highly competitive results; even tough we have not applied any strategy for optimizing the retrieval process just as we did with the baseline techniques. On both representations the best results were obtained with the TF-IDF version, which is not surprising as this version takes into account the frequency of appearance of terms as well as their usage.

MTCOR-TF-IDF obtained better results than EF-W (the best baseline method) on ImageCLEF2008 topics, the relative improvement in **MAP** is of 5.24%; although it obtained slightly lower results on ImageCLEF2007 topics. The decreasing in performance is by 2.43%. On the other hand, MDOR-TF-IDF outperforms significantly (in all measures but **P20** in ImageCLEF2007 topics) the results obtained with any other technique on both sets of topics. The relative improvement of MDOR-TF-IDF over EF-W in **MAP** is of 18.01% (averaged over the two sets of topics), while the corresponding improvement over the textual method (Text) is of 81.15%. The latter results show the benefits of using the proposed representations over the baseline methods (which require a parameter optimization stage for obtaining satisfactory results) and over unimodal retrieval methods.

In Sect. 6.1 we showed that varying the weights associated to each modality (i.e., α_T and α_L) for the baseline methods increases their retrieval effectiveness. Therefore, we performed an experiment in which we tried to control the contribution of each modality into the representations MDOR-TF-IDF and MTCOR-TF-IDF. It should be noted that under the proposed representations we cannot assign weights to the involved modalities in a natural way. Nevertheless, we decided to weight the representations of the terms that belong to different modalities with a different weight; for example, the weighted version of MTCOR-TF-IDF is given by:

$$\mathbf{d}_i^{mtcor-tfidf-w} = \alpha_L \times \mathbf{d}_{i,l}^{mtcor-tfidf} + \alpha_T \times \mathbf{d}_{i,t}^{mtcor-tfidf} \tag{14}$$

with:

$$\mathbf{d}_{i,l}^{mtcor-tfidf} = \sum_{kl=1}^{N_i^l} (tf_m(i, t_i^{kl}) \times \mathbf{w}_{kl}^{tcor}) \times (\log(\frac{N}{N_k^m})) \tag{15}$$

$$\mathbf{d}_{i,t}^{mtcor-tfidf} = \sum_{kt=1}^{N_i^t} (tf_m(i, t_i^{kt}) \times \mathbf{w}_{kt}^{tcor}) \times (\log(\frac{N}{N_k^m})) \tag{16}$$

where α_T and α_L are scalars weighting the contribution of the textual and labels modalities, respectively. $\mathbf{d}_{i,l}^{mtcor\text{-}tfidf}$ and $\mathbf{d}_{i,t}^{mtcor\text{-}tfidf}$ are the MTCOR-TF-IDF representations for document $\mathbf{d}_i$ using only terms from the labels and textual modalities, respectively. N_i^l is the number of terms from the labels vocabulary that appear in the document i and N_i^t is the number of terms from the vocabulary of text that occur in document i. The weighted MDOR-TF-IDF representation can obtained similarly. Intuitively, the weights α_L and α_T weight the importance of the labels (α_L) and words (α_T) that occur in document $\mathbf{d}_i$. We call the weighted versions of the representations MDOR-TF-IDF-W and MTCOR-TF-IDF-W.

We performed experiments with different values for α_T and α_L with both MDOR-TF-IDF-W and MTCOR-TF-IDF-W, the best results are shown in Table 4. The improvement in **MAP** obtained with the weighted version of MDOR-TF-IDF over the un-weighted one is of only 1.03%. On the other hand, the weighted version of MTCOR-TF-IDF improved by 5.98% to the un-weighted version. Note that while MTCOR-TF-IDF-W improved the performance of MTCOR-TF-IDF in all measures, MDOR-TF-IDF-W did not improve the performance of MDOR-TF-IDF in **P20** and **RR**. The latter result indicates that MDOR-TF-IDF-W modified negatively the order of documents in the first positions and that less relevant documents were retrieved with MDOR-TF-IDF-W.

The results from Table 4 show that the way we have weighted the contribution of each modality may not be the best, as when we modified the weights for the standard techniques we obtained improvements of over 40% in terms of **MAP** (see for example the results obtained with EF-W in Table 2). However, those results give evidence that the proposed representations are more robust to the optimization of weights than standard techniques. Thus, even without optimizing the weights we can obtain highly competitive performance.

Table 4 Retrieval performance obtained with MDOR-TF-IDF-W and MTCOR-TF-IDF-W; we also show the results obtained by the un-weighted versions and baseline techniques

Weights α		ImageCLEF2007				ImageCLEF2008			
α_L	α_T	MAP	P20	R20	RR	MAP	P20	R20	RR
MDOR-TF-IDF-W									
1	1.825	**0.2141**	0.2425	**0.3295**	2,507	**0.1958**	0.25	**0.3333**	1,679
1	1	0.2102	**0.2467**	0.3231	**2,521**	0.1954	**0.2705**	0.3217	**1,686**
MTCOR-TF-IDF-W									
1	2	0.1935	0.2392	0.2951	2,379	0.1763	0.2564	0.2893	1,632
1	1	0.1844	0.2383	0.2807	2,256	0.1647	0.2397	0.2688	1,473
Baseline methods									
LF-W		0.1348	0.1858	0.1879	1,703	0.1126	0.1936	0.1759	1,139
EF-W		0.189	0.2508	0.2996	2,226	0.1565	0.2372	0.2695	1,416
IRF-2		0.1659	0.2142	0.262	1,952	0.1326	0.1987	0.2205	1,302

The best result of each column is shown in bold

It is interesting that the best configurations were obtained by assigning similar weights to both modalities (i.e., the weight assigned to the textual modality is at most the double of that assigned to the labels modality). The latter suggest that both modalities contribute similarly to obtain an effective representation for documents. This is opposed to what happened with EF-W where the best configuration was obtained by assigning a weight 8 times larger to the textual modality.

If we compare the results obtained by MDOR-TF-IDF-W (row 4 in Table 4) with the results reported in the ImageCLEF2007 (Grubinger et al. 2008) and ImageCLEF2008 (Arni et al. 2009) forums, we could see that our method would be ranked in positions 36 and 143 of a total of 474 and 1,047 entries submitted to ImageCLEF2007 and ImageCLEF2008, respectively; that is, within the 7.6 and 13.7% of the total of submitted runs, respectively. This is a very positive result, since we should note that the retrieval systems that were evaluated in those forums were composed of a considerable number of components besides the retrieval model. For example, some researchers analyzed syntactically the queries (Maillot et al. 2006), applied named entity recognizers (Martínez-Fernández et al. 2006; Maillot et al. 2006), performed query expansion with external resources (Chang and Chen 2006), included processes that required user interaction (Clinchant et al. 2007) or even combined multiple retrieval models (Grubinger et al. 2008; Arni et al. 2009). In this paper, however, we have obtained satisfactory results by using a simple VSM based on the representations we proposed. In consequence, we expect to obtain better results when other components will be combined with our methods.

Table 5 shows results of statistical significance tests among the different results in terms of **MAP**. From this table we can verify that the configuration MDOR-TF-IDF-W outperforms any other method and that these differences are statistically significant; MDOR-TF-IDF outperforms to all but the EF-W method, which proved to be very effective. On the other hand, results obtained with MTCOR representations are not statistically better than the methods EF-W and LF-W. Thus, the best option for representing documents is MDOR-TF-IDF-W or even MDOR-TF-IDF. Representations based on MTCOR are less effective, although, in average, better performance than LF, EF and IRF can be obtained with such techniques.

We have seen that MDOR-based representations outperformed consistently those based on MTCOR. This can be due to the size of the collection and the number of terms per document. Since documents are composed of 11.43 terms in average, the co-occurrence statistics used by MTCOR are not enough for effectively capturing the context of terms; which directly affects the corresponding document representations. On the other hand, each term from the multimodal vocabulary occurs (in average) in 33.13 documents of the

Table 5 Results of the tests of statistical significance among different results

Configuration	Labels	Text	LF-W	EF-W	IRF-2
MDOR-B	+/+	+/−	+/−	−/−	−/−
MDOR-TF-IDF	+/+	+/+	+/+	−/−	+/+
MDOR-TF-IDF-W	+/+	+/+	+/+	+/+	+/+
MTCOR-B	+/+	+/−	−/−	−/−	−/−
MTCOR-TF-IDF	+/+	+/+	+/−	−/−	−/−
MTCOR-TF-IDF-W	+/+	+/+	+/+	−/−	−/−

Each row indicates whether (+) or not (−) the difference with the methods of the corresponding column was statistical significant. The slash / separates the results for the different sets of topics (i.e., ImageCLEF2007/ImageCLEF2008)

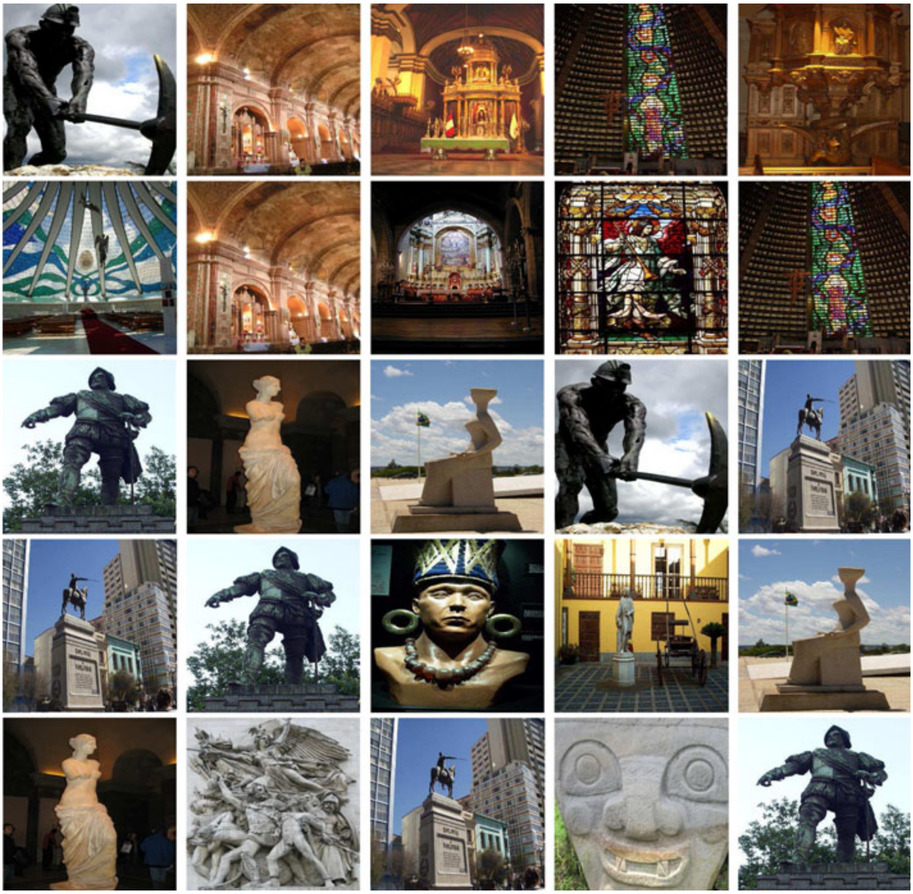

Fig. 8 Top-5 non-relevant documents retrieved with MDOR-TF-IDF (*row 1*), MTCOR-TF-IDF (*row 2*), EF-W (*row 3*), LF-W (*row 4*) and IRF-2 (*row 5*) for the topic *"religious statue in the foreground"*, depicted in Fig. 5

collection; therefore, the document occurrence information is less sparse and hence MDOR based representations are richer than MTCOR ones.

Besides retrieval performance, document representations based on MDOR and MTCOR offer an additional advantage, namely: they can capture better the content of documents and queries; for illustrating the latter aspect, in Fig. 8 we show the top 5 non-relevant documents retrieved with the proposed representations and with baseline methods for the topic shown in Fig. 5.

We can see that even non-relevant documents that were retrieved with MDOR-TF-IDF and MTCOR-TF-IDF (rows 1 and 2) are semantically related to the topic (i.e., religious statues), giving evidence that the multimodal DTRs effectively capture the semantic content of the queries. On the other hand, the non-relevant documents as retrieved with EF-W, LF-W and IRF-2 (rows 3-6) are not related at all to the topic of interest. Thus, besides that multimodal DTRs are helpful for improving the multimedia image retrieval performance of the VSM, they are able to capture effectively the content of queries and documents.

Table 6 Retrieval results when using unimodal queries (the modality of the query is indicated in columns 1 and 2, *L* is for labels and *T* is for text) and the multimodal representations for documents

Query		ImageCLEF2007				ImageCLEF2008			
L	T	MAP	P20	R20	RR	MAP	P20	R20	RR
MDOR-TF-IDF									
1	0	0.1357	0.1708	0.2135	2,000	0.1466	0.2013	0.2348	1,282
0	1	0.1497	0.1658	0.1969	1,980	0.1292	0.159	0.1833	1,347
avg.		**0.1427**	0.1683	**0.2052**	**1,990**	**0.1379**	0.1801	**0.2091**	**1,314.5**
MTCOR-TF-IDF									
1	0	0.1347	0.175	0.2182	1,876	0.1413	0.1974	0.2453	1,194
0	1	0.123	0.1442	0.1709	1,696	0.1067	0.1282	0.1712	1,191
avg.		0.1288	0.1596	0.1945	1,786	0.124	0.1628	0.2082	1,192.5
Late fusion									
1	0	0.0587	0.1417	0.1066	1,201	0.053	0.141	0.1133	727
0	1	0.1241	0.1767	0.1694	1,424	0.1033	0.1795	0.1534	1,014
avg.		0.0914	0.1592	0.138	1,312.5	0.0781	0.1602	0.1333	870.5
Early fusion									
1	0	0.1004	0.1908	0.1752	1,561	0.0947	0.2	0.1734	947
0	1	0.1094	0.1700	0.1601	1,416	0.0993	0.1718	0.1452	970
avg.		0.1049	**0.1804**	0.676	1,488.5	0.0970	**0.1859**	0.1593	958.5
Inter-media relevance feedback									
1	0	0.0841	0.165	0.136	1,273	0.0714	0.1577	0.123	706
0	1	0.1435	0.1867	0.2029	1,717	0.1253	0.1974	0.2009	1,255
avg.		0.1138	0.1758	0.1694	1,495	0.0983	0.1775	0.1619	980.5

For each technique we show the average obtained under both modalities. We show in bold the best (average) result of each column

6.3 Unimodal queries with multimodal representations

Although the proposed representations are based on information resulting from the interaction between multimodal terms, it may be possible that users of image collections are not willing to formulate queries in both modalities. That is, for some users it may be more comfortable to specify queries by means of images, or through the labels associated to the images; while for other users it may be more easy to provide textual queries. With the goal of evaluating the usefulness of the proposed representations under the latter retrieval scenarios we performed the following experiment.

For each topic we considered separately the modalities that compose the query. That is, we search for images by using as query either labels or words, but using the multimodal representations for documents/queries as described in Sect. 5. Results of this experiment are shown in Table 6, for comparison we also show the results obtained with LF, EF and IRF.[9]

It is clear that while the effectiveness of all methods decreased considerably, the representations MDOR-TF-IDF and MTCOR-TF-IDF were still able to obtain acceptable performance. The baseline methods obtained satisfactory results when the query was

[9] For IRF we report the best results obtained when the initial and final queries belong to the same modality.

Table 7 Retrieval results using automatically generated labels

Weights α		ImageCLEF2007				ImageCLEF2008			
α_L	α_T	MAP	P20	R20	RR	MAP	P2	R20	RR
MDOR-TF-IDF-W									
1	10	**0.1388**	0.1592	0.1922	**1,938**	**0.1205**	0.159	**0.1817**	**1,332**
1	1	0.0903	0.1542	0.1487	1,321	0.0565	0.1308	0.0967	826
MTCOR-TF-IDF-W									
0.5	1	0.0932	0.1283	0.1389	1,476	0.0803	0.1282	0.1378	1,045
1	1	0.0815	0.1492	0.1313	1,155	0.0518	0.1244	0.0871	744
Late fusion									
1	1	0.1198	0.1758	0.1634	1,451	0.0934	0.1744	0.1436	972
1	10	0.1236	0.1792	0.1697	1,543	0.101	0.1885	0.1534	1,003
Early fusion									
1	8	0.0567	0.1458	0.0991	1,072	0.0386	0.1308	566	691
1	8	0.1375	**0.205**	**0.2115**	1,798	0.1139	**0.2013**	0.178	1,223
Inter-media relevance feedback									
1	t	0.1196	0.1617	0.1642	1,583	0.1048	0.1769	0.1688	1,094
Individual methods									
1	0	0.0158	0.0525	0.0224	656	0.0119	0.05	0.0188	440
0	1	0.1241	0.1767	0.1694	1,424	0.1033	0.1795	0.1534	1,014

The best result of each column is shown in bold

formulated in text, however, when the query was formulated by means of labels, the standard techniques obtained very poor results. On the other hand, the methods MDOR-TF-IDF and MTCOR-TF-IDF obtained acceptable results for queries in both modalities, giving evidence that documents are effectively represented with the proposed representations and illustrating another benefit of using multimodal DTRs for representing documents.

6.4 Results with automatically generated labels

For all of the experiments described so far we have considered labels that were manually assigned to images. However, in many databases images have not associated any label, thus the use of automatic image annotation (AIA) methods is imperative in such collections. In this section we report experimental results obtained with the proposed representations when using the AIA technique introduced in Escalante et al. (2011). Such method uses multiclass classifiers and a labeling refinement method that attempts to maximize the cohesion of labels through the minimization of an energy function. For our experiments we performed 10-fold cross-validation for automatically assigning labels to each region in the SAIAPR TC12 collection.

In Table 7 we show the retrieval results when using labels automatically generated. As expected, the retrieval effectiveness of all of the methods decreased significantly. Nevertheless, MDOR-TF-IDF-W still outperforms the results obtained with unimodal techniques. The best performance in **MAP** was obtained with MDOR-TF-IDF-W, with a weight 10 times higher for text than for labels. While the best result of baseline methods was obtained by EF with a weight 8 times higher for the textual modality. In this scenario a high weight to the textual part compensates the uncertainty inherent in assigning labels automatically.

One should note that the use of automatic labels affected significantly the representation MTCOR-TF-IDF-W. This can be due to the fact that co-occurrence statistics among labels and text are not reliable any more (i.e., if many images in the collection are incorrectly labeled, the co-occurrence statistics among labels and words will be erroneous) and hence the MTCOR representations of terms do not reflect the semantics of terms.

MDOR-TF-IDF-W is more robust than MTCOR-TF-IDF-W because this representation is based on term occurrence in documents, thus even when labels assigned to documents are incorrect, the mistakes made by the AIA method will be (more or less) consistent across most documents in the collection (e.g., if the AIA method confuses the label *"sky"* with the label *"cloud"*, this confusion will prevail in most documents of the collection) and therefore the automatic labels still will be helpful to represent terms and hence documents as well.

An additional issue that affected the performance of the evaluated techniques is that for assigning automatic labels we only considered 90 concepts (those for which there exist at least 200 regions across the collection) out of the 255 used for the other experiments. Hence, it is very likely that the considered labels did not include labels that were used for building queries (e.g the label *"statue"* was not considered, hence, labels cannot improve the retrieval performance for the topic shown in Fig. 5). Also, the accuracy of the AIA method that we used is below 40%. Nevertheless, we expect that the retrieval performance of the proposed techniques will improve when better AIA methods are used.

7 Conclusions

We have introduced two novel representations for multimodal documents that are based on the idea of modeling the semantic cohesion among multimodal terms. The proposed representations were successfully applied to the multimedia image retrieval task. On the one hand, under the MDOR representation (for *multimodal document occurrence representation*) each term of the vocabulary is represented by a distribution of occurrences over the documents in the collection. This way, the documents where mostly occur a term determine its representation; terms that occur in similar documents will have similar representations. On the other hand, under the MTCOR representation (for *multimodal term co-ocurrence representation*), the terms are represented by a distribution of co-occurrences over the terms in the vocabulary. Terms that frequently co-occur with related terms will have similar representations. We represent multimodal documents by combining the representations of the terms that occur in the document. Thus, documents are represented by distributions over either documents or terms. Both forms of document representation can be considered an expansion of the terms that occur in the document, such expansion will be determined by the context in which terms occur.

We compare the proposed representations with highly effective baseline methods that combine text and labels. Our results show that the proposed representations outperform the baseline techniques significantly. Furthermore, the proposed representations offer additional benefits. Specifically, they are able to capture the semantics of documents better and allow us to search for images by using unimodal queries. The results reported in this paper give evidence of the effectiveness and usefulness of the proposed representations.

The main contributions of this paper are as follows:

- We introduced two novel ways of representing multimodal documents that are based on distributional term representations. Both sort of representations combine effectively information from text and labels for representing documents. The proposed representations capture effectively the semantics of multimodal documents and hence they can be used for other aspects of the multimedia retrieval task, for example, for query/document expansion, for the organization of documents into semantic categories and for the generation of visual resources. Despite that in this work we have only applied the multimodal distributional representations to multimedia image retrieval, our techniques can also be used to model other types of multimedia documents (e.g., video).
- We reported experimental evidence that shows how the proposed representations can improve significantly the performance of strong multimedia image retrieval baselines. Furthermore, the retrieval performance of the proposed representations was significantly better than baseline techniques when using unimodal queries.
- We presented retrieval results with the proposed methods using labels generated manually and automatically. On the one hand, results obtained with manually assigned labels motivate the development of methods for combining labels and text as the potential of retrieval performance that can be obtained by using labels is worthwhile. On the other hand, results with automatically generated labels show that even when the performance of current annotation techniques is limited they still can be helpful for multimedia image retrieval; however, better annotation techniques are required.

Future work directions include using the MDOR and MTCOR based representations for modeling other multimodal documents and with other retrieval models; using the MDOR and MTCOR representations to obtain richer visual vocabularies for the tasks of object recognition and image categorization; taking advantage of MDOR and MTCOR for the development of methods for query expansion and relevance feedback; extending the MDOR and MTCOR document representations; studying the complementariness and redundancy of MDOR and MTCOR representations; and developing an effective way to weight the contribution of both modalities (text and labels) under the MDOR and MTCOR representations.

Acknowledgments The authors are grateful with the reviewers for their comments that have helped to improve the content of the article. We also thank the valuable support of Thamar Solorio in the preparation of the final manuscript. This project was supported by CONACyT under project grant 61335 and scholarship 205834.

References

Adriani, M., & Framadhan, R. (2005). University of Indonesia participation at IMAGE-CLEF 2005. *Working notes of the CLEF workshop*, Vienna, Austria, 2005.

Ah-Pine, J., Clinchant, S., Csurka, G., & Liu, Y. (2008). XRCE's participation in ImageCLEF 2008. Working notes of the 2008 CLEF Workshop, Aarhus, Denmark.

Ah-Pine, J., Clinchant, S., Csurka, G., & Liu, Y. (2009). XRCE's participation in ImageCLEF 2009a. *Working notes of the CLEF workshop*, Corfu, Greece.

Ah-Pine, J., Bressan, M., Clinchant, S., Csurka, G., Hoppenot, Y., & Renders, J. M. (2009b). Crossing textual and visual content in different application scenarios. *Multimedia Tools and Applications, 42*:31–56.

Allan, M., & Verbeek, J. (2009). Ranking user annotated images for multiple query terms. *Proceedings of the 20th British Machine Vision Conference*, London, UK.

Aly, R., Hiemstra, D., & Ordelman, R. (2007). Building detectors to support searches on combined semantic concepts. *In Proceedings of the SIGIR multimedia information retrieval workshop*, pp. 40–45, Amsterdam, The Netherlands.

Aly, R., Hiemstra, D., & de Vries, A. (2009). Reusing annotation labor for concept selection. *Proceedings of the international conference on content-based image and video retrieval*, pp. 44, ACM Press.

Arni, T., Sanderson, M., Clough, P., & Grubinger, M. (2009). Overview of the ImageCLEFphoto 2008 photographic retrieval task. *Evaluating systems for multilingual and multimodal information access, lecture notes in computer science*, Vol. 5706, pp. 500–511, Springer.

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

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

Besancon, R., & Millet, C. (2006). Using text and image retrieval systems: Lic2m experiments at ImageCLEF 2006. *Working notes of the CLEF workshop*, Alicante, Spain.

Boldareva, L., & Hiemstra, D. (2004). Interactive content-based retrieval using pre-computed object–object similarities. *Proceedings of the international conference on image and video retrieval, lecture notes in computer science*, Vol. 3115, pp. 308–316, Springer.

Bradshaw, B. (2000). Semantic based image retrieval: A probabilistic approach. *Proceedings of the 8th ACM international conference on Multimedia*, pp. 167–176, ACM Press, Los Angeles, CA, USA.

Carneiro, G., Chan, A. B., Moreno, P. J., & Vasconcelos, N. (2007). Supervised learning of semantic classes for image annotation and retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 29*(3):394–410.

Carrillo, M., Eliasmith, C., & López-López A. (2009). Combining text vector representations for information retrieval. *Proceedings of the 12th international conference on text, speech and dialogue, lecture notes in computer science*, Vol. 5729, pp. 24–31, Springer, Czech Republic.

Chang, Y., & Chen, H. (2006). Approaches of using a word-image ontology and an annotated image corpus as intermedia for cross-language image retrieval. *Working Notes of the CLEF Workshop*, Alicante, Spain.

Chang, Y., Lin, W., & Chen, H. H. (2005). Combining text and image queries at ImageCLEF 2005. *Working notes of the CLEF workshop*, Vienna, Austria.

Chen, H., Yim, T., Fye, D., & Schatz, B. R. (1995). Automatic thesaurus generation for an electronic community system. *Journal of the American Society for Information Science, 46*(3), 175–193.

Chua, T., Pung, H., Lu, G., & Jong, H. (1994). A concept-based image retrieval system. *Proceedings of the 27th annual Hawaii international conference on system sciences*, pp. 590–598, IEEE, Wailea, HI, USA.

Clinchant, S., Renders, J., & Csurka, G. (2007). XRCE's participation to ImageclefPhoto 2007. *Working notes of the 2007 CLEF workshop*, Budapest, Hungary.

Clough, P., Grubinger, M., Deselaers, T., Hanbury, A., & Müller, H. (2007). Overview of imageCLEF 2006 photographic retrieval and object annotation tasks. *7th Workshop of the cross-language evaluation forum, CLEF 2006, revised selected papers, lecture notes in computer science*, Vol. 4730, pp. 579–594, Springer.

Cox, J., Miller, M., Minka, P., Papathomas, V., & Yianillos, N. (2000). The bayesian image retrieval system, PicHunter: Theory, implementation, and psychophysical experiments. *IEEE Transactions on Image Processing, 9*(1):20–37.

Curtoni, P. (2006). CELI participation at ImageCLEF 2006: Comparison with the Ad-hoc track. *Working Notes of the CLEF Workshop*, Alicante, Spain.

Datta, R., Joshi, D., Li, J., & Wang, J. Z. (2008). Image retrieval: Ideas, influences, and trends of the New Age. *ACM Computing Surveys, 40*(2):1, 60.

Elworthy, D. (2000).. Retrieval from captioned image databases using natural language processing. *Proceedings of the 9th international conference on information and knowledge management*, pp. 430–437, ACM Press, McLean, VA, USA.

Escalante, H. J., Montes, M., & Sucar, E. (2011). An energy-based model for region-labeling. *Computer vision and image understanding*, In press, http://dx.doi.org/10.1016/j.cviu.2011.02.00.

Escalante, H. J., González, J. A., Hernández, C. A., López, A., Montes, M., Morales, E., et al. (2009). Annotation-based expansion and late fusion of mixed methods for multimedia image retrieval. *Evaluating systems for multilingual and multimodal information access, lecture notes in computer science*, Vol. 5706, pp. 669–676, Springer.

Escalante, H. J., Hernández, C., López, A., Marin, H., Montes, M., Morales, E., et al. (2008a). Towards annotation-based query and document expansion for image retrieval. *Advances in multilingual and multimodal information retrieval, lecture notes in computer science*, Vol. 5152, pp. 546–553, Springer.

Escalante, H. J., Hernández, C., Sucar, E., & Montes. M. (2008b). Late fusion of heterogeneous methods for multimedia image retrieval. *Proceedings of the 2008 ACM multimedia information retrieval conference*, pp. 172–179, ACM Press, Vancouver, BC, Canada.

Escalante, H. J., Montes, M., & Sucar, E. (2009). On multimedia image retrieval baselines. *Proceedings of the CIMAT-PI'09 workshop*, Guanajuato, Mexico.

Escalante, H. J., Grubinger, M., Hernández, C. A., González, J. A., López, A., Montes, M., et al. (2010). The segmented and annotated IAPR TC-12 benchmark. *Computer Vision and Image Understanding, 114*(4):419–428.

Farah, M., & Vanderpooten, D. (2007). An outranking approach for rank aggregation in information retrieval. *Proceedings of the 30th international ACM SIGIR conference on research and development in information retrieval*, pp. 591–598, ACM Press, Amsterdam, The Netherlands.

Fox, E. A., & Shaw, J. A. (1994). Combination of multiple searches. *Proceedings of TREC'3, NIST Publication.*

Gale, W. A., Church, K. W., & Yarowsky, D. (1993). A method for disambiguating word senses in a large corpus. *Computers and the Humanities, 26*(5), 415–439.

Goodrum, A. (2000). Image information retrieval: An overview of current research. *Journal of Informing Science, 3*(2):63, 66.

Grangier, D., Monay, F., & Bengio, S. (2006). A discriminative approach for the retrieval of images from text queries. *Proceedings of the 17th European conference on machine learning, lecture notes in artificial intelligence*, Vol. 4212, pp. 162–173, Springer, Berlin, Germany.

Grangier, D., & Bengio, S. (2006). A neural network to retrieve images from text queries. *Proceedings of international conference on artificial neural networks, lecture notes in computer science*, Vol. 4132, pp. 24–34, Springer, Athens, Greece.

Grubinger, M. (2007). Analysis and evaluation of visual information systems performance. *PhD Thesis. School of Computer Science and Mathematics, Faculty of Health, Engineering and Science, Victoria University*, Melbourne, Australia.

Grubinger, M., Clough, P., Hanbury, A., & Müller, H. (2008). Overview of the imageCLEF 2007 photographic retrieval task. *Advances in multilingual and multimodal information retrieval, lecture notes in computer science*, Vol. 5152, pp. 433–444, Springer

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

Hanbury, A. (2006). Review of image annotation for the evaluation of computer vision algorithms. *Tech. rep., PRIP, Vienna University of Technology*, 102, Vienna, Austria.

Hare, J. S., Lewis, P. H., Enser, P. G. B., & Sandom, C. J. (2006). Mind the gap: Another look at the problem of the semantic gap in image retrieval. *Proceedings of multimedia content analysis, management and retrieval: Trends and challenges*, Vol. 6073(1), pp. 1–12, SPIE, San Jose, CA, USA.

Hoi, S. C. H., Zhu, J., & Lyu, M. R. (2005). CUHK experiments with ImageCLEF 2005. Working notes of the CLEF workshop, Vienna, Austria.

Inoue, M., & Ueda, N. (2005). Retrieving lightly annotated images using image similarities. *Symposium on applied computing*, pp. 1031–1037, ACM Press, Santa Fe, New Mexico.

Ishikawa, Y., Subramanya, R., & Faloutsos, C. (1998). MindReader: Querying databases through multiple examples. *Proceedings of the 24th international conference on very large data bases*, pp. 218–227, IEEE, New York, NY, USA.

Izquierdo-Beviá, R., Tomás, D., Saiz-Noeda, M., & Luis Vicedo, J. (2005). University of Alicante in ImageCLEF2005, *Working Notes of the CLEF Workshop*, Vienna, Austria.

Jeon, J., Lavrenko, V., & Manmatha, R. (2003). Automatic image annotation and retrieval using cross-media relevance models. *SIGIR'03: Proceedings of the 26th international ACM-SIGIR conference on research and development on information retrieval*, pp. 119–126, Toronto, Canada.

Jones, G. J. F., & McDonald, K. (2005). Dublin city university at CLEF 2005: Experiments with the ImageCLEF St Andrews collection. *Working Notes of the CLEF Workshop*, Vienna, Austria.

Kraaij, W., Smeaton, A. F., Over, P., & Ianeva, T. (2006). TrecVID 2005 an overview. In TREC video retrieval evaluation online proceedings.

La Cascia, M., Sethi, S., & Sclaroff, S. (1998). Combining textual and visual cues for content-based image retrieval on the world wide web. *Proceedings of the IEEE workshop on content-based access of image and video libraries*, pp. 24–28, Santa Barbara, CA, USA.

Larson, M., Newman, E., & Jones, G. J. F. (2009). Overview of VideoCLEF 2009: New perspectives on speech-based multimedia content enrichment. *Working notes of the CLEF workshop*, Corfu, Greece.

Lavelli, A., Sebastiani, F., & Zanoli, R. (2005). Distributional term representations: An experimental comparison. *Proceedings of the international conference of information and knowledge management*, pp. 615–624, ACM Press, Kuala Lumpur, Malaysia.

Lestari-Paramita, M., Sanderson, M., & Clough, P. (2009). Diversity in photo retrieval: Overview of the ImageCLEFPhoto task 2009. *Working Notes of the CLEF Workshop*, Corfu, Greece.

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

Lewis, D. D., & Croft, W. B. (1990). Term clustering of syntactic phrases. *Proceedings of the 13th international ACM SIGIR conference on research and development in information retrieval*, pp. 385–404, ACM Press, Bruxelles, Belgium.

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

Maillot, N., Chevallet, J., Valea, V., & Lim, J. H. (2006). IPAL inter-media pseudo-relevance feedback approach to ImageCLEF 2006 photo retrieval. *Working Notes of the CLEF Workshop*, Alicante, Spain.

Martínez-Fernández, J., Villena, J., García-Serrano, A., Martínez, P. (2006). MIRACLE team report for ImageCLEF IR in CLEF 2006. *Working Notes of the CLEF Workshop*, Alicante, Spain.

Martínez-Fernández, J., Villena, J., García-Serrano, A., González-Tortosa, S., Carbone, F., & Castagnone, D. (2005). Exploiting semantic features for image retrieval at CLEF 2005. *Working notes of the CLEF workshop*, Vienna, Austria.

Martín-Valdivia, M. T, García-Cumbreras, M. A., Díaz-Galiano, M. C., Urea-López, L. A., & Montejo-Raez, A. (2005). SINAI at ImageCLEF 2005. *Working Notes of the CLEF Workshop*, Vienna, Austria.

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

Peinado, V., López-Ostenero, F., & Gonzalo, J. (2005). UNED at ImageCLEF 2005: Automatically structured queries with named entities over metadata. Working Notes of the CLEF Workshop, Vienna, Austria.

Rahman, M. M., Sood, V., Desai, B. C., & Bhattacharya, P. (2006). CINDI at ImageCLEF 2006: Image retrieval and annotation tasks for the general photographic and medical image collections. *Working Notes of the CLEF Workshop*, Alicante, Spain.

Raicu, D. S, & Sethi, I. K. (2006). Synobins: An intermediate level towards annotation and semantic retrieval. *EURASIP Journal on Applied Signal Processing, 2006*, Article ID 63124, pp. 1–19.

Rautiainen, M., Ojala, T., & Tapio, S. (2004). Analysing the performance of visual, concept and text features in content-based video retrieval. *Proceedings of the 6th ACM international workshop on multimedia information retrieval*, pp. 197–204, ACM Press, New York, NY, USA.

Rautiainen, M., & Seppdnen, T. (2005). Comparison of visual features and fusion techniques in automatic detection of concepts from news video. *Proceedings of the international conference on multimedia and expo*, pp. 932–935, IEEE, Amsterdam, The Netherlands.

Reyes, A., Montes, M., & Villaseñor, L. (2011). Combining word and phonetic-code representations for spoken document retrieval. *Proceedings of the 12th international conference on intelligent text processing and computational linguistics*, LNCS, Forthcoming, Tokio, Japan, Springer.

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

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

Salton, G., Yang, C. S., & Wong, A. (1975). A vector space model for automatic indexing. *Communications of the ACM, 18*(11), 613–620.

Salton, G., & Buckley, C. (1987). Term weighting approaches in automatic text retrieval. *Technical Report, Cornell University*, TR87–881, Ithaca, NY, USA.

Sclaroff, S., La Cascia, M., & Sethi, S. (1999). Unifying textual and visual cues for content-based image retrieval on the world wide web. *International Journal of Computer Vision, 75*(1–2), 86–98.

Shu, D. F., & Taska, I. (2005). Comparing rank and score combination methods for data fusion in information retrieval. *Information retrieval, 8*, 449–480.

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

Smucker, D., Allan, J., & Carterette, B. (2009). Agreement among statistical significance tests for information retrieval evaluation at varying sample sizes. *Proceedings of the 32th international ACM SIGIR conference on research and development in informaion retrieval*, ACM Press, Boston, MA, USA.

Snoek, C. G. M., Worring, M., van Gemert, J. C., Geusebroek, J. M., & Smeulders, A. W. M. (2006). The challenge problem for automated detection of 101 semantic concepts in multimedia. *Proceedings of the 14th annual ACM conference on multimedia.* ACM Press, New York.

Snoek, C., Worring, M., & Smeulders, A. (2005). Early versus late fusion in semantic video analysis. *Proceedings of the 13th Annual ACM Conference on Multimedia*, pp. 399–402, ACM Press, Singapore.

van Gemert, J. (2003). Retrieving images as text, *MS Thesis*, Intelligent Sensory Information Systems, University of Amsterdam, The Netherlands.

Westerveld, T. (2000). Image retrieval: Content versus context. *Proceedings of the RIAO Conference*, pp. 276–284, Paris, France.

Westerveld, T. (2004). Using generative probabilistic models for multimedia retrieval. *PhD Thesis*, Twente University, The Netherlands.

Westerveld, T., Hiemstra, D., & de Jong, F. M. G. (2000). Extracting bimodal representations for language-based image and text retrieval. *Proceedings of the Eurographics w*, pp. 33–42, Milan, Italy.

Winn, J., Criminisi, A., & Minka, T. (2005). Object categorization by learned universal visual dictionary. *Proceedings of the international conference on computer vision*, pp. 1800–1807, IEEE, Beijing, China.

Zhang, C., Chai, J. Y., & Jin, R. (2005). User term feedback in interactive text-based image retrieval. *Proceedings of the 28th international ACM SIGIR conference on research and development in information retrieval*, pp. 51–58, ACMPress, Salvador, Brazil.

Zhou, Z., Chen, K., & Dai, H. (2006). Enhancing relevance feedback in image retrieval using unlabeled data. *ACM Transactions on Information Systems, 24*(2), 219–244.