Using Manual and Automated Annotations to Search Images by Semantic Similarity

João Magalhães
Faculdade de Ciências e Tecnologia
Universidade Nova de Lisboa
Portugal
jmag@di.fct.unl.pt

Stefan Rüger
Knowledge Media Institute
The Open University
Milton Keynes, UK
s.rueger@open.ac.uk

Abstract

Finding semantically similar images is a problem that relies on image annotations manually assigned by amateurs or professionals, or automatically computed by some algorithm using low-level image features. These image annotations create a keyword space where a dissimilarity function quantifies the semantic relationship among images. In this setting, the objective of this paper is two-fold. First, we compare amateur to professional user annotations and propose a model of manual annotation errors, more specifically, an asymmetric binary model. Second, we examine different aspects of search by semantic similarity. More specifically, we study the accuracy of manual annotations versus automatic annotations, the influence of manual annotations with different accuracies as a result of incorrect annotations, and revisit the influence of the keyword space dimensionality. To assess these aspects we conducted experiments on a professional image dataset (Corel) and two amateur image datasets (one with 25,000 Flickr images and a second with 269,648 Flickr images) with a large number of keywords, with different similarity functions and with both manual and automatic annotation methods. We find that Amateur-level manual annotations offers better performance for top ranked results in all datasets (MP@20). However, for full rank measures (MAP) in the real datasets (Flickr) retrieval by semantic similarity with automatic annotations is similar or better than amateur-level manual annotations.

1. Introduction

Multimedia retrieval systems are best at processing user queries represented by Boolean expressions, and not everyone has the same skills at expressing ideas, emotions and feelings in such a formal way. While in text retrieval we express our query in the format of the document (text), in multimedia retrieval systems this is more difficult due to semantic ambiguities. The user is not aware of the low-level representation of multimedia, e.g., colour, texture, shape features, pitch, volume or tones. Instead the user is often more interested in the semantic richness of multimedia information. This demands a search system that relies on a high-level representation of multimedia, thus, providing a semantic layer to multimedia documents. The usefulness of such a semantic space ranges from search-by-example to tag-suggestion systems and recommender-systems.

Figure 1. Images with *baby* annotation.

Figure 2. Images with *baby*, *elephant, and mum* annotations.

In this paper, we address the problem of search-by-semantic-example. This paradigm allows the user to submit a single example image of a yellow flower and retrieve images of flowers of all colours, textures and backgrounds. This is possible because the search space does not represent images by their low-level features but by their high-level concepts (e.g., flowers, mountains, river, or sky). Figure 1 illustrates an example where the user searched for “baby” to find all images annotated with the word baby. Note that the ambiguity of the search query is the reason for such diverse results. If the user then clicks on an image containing a baby elephant and its mother, a search by semantic example

uses the annotations of that image to search for semantically similar images, i.e., images sharing the same set of annotations as is illustrated in Figure 2. The important fact to retain from this example is that images are compared in terms of their annotations and not in terms of their colours, textures or shapes.

The first decision that we face in this framework is the representation of images. Formally, for any given image d, we capture its annotations by the keyword vector

$$d_W = \left(d_{W,1}, \ldots, d_{W,L}\right) \in \left[0,1\right]^L, \quad (1)$$

for L keywords from the vocabulary $\mathcal{W} = \left\{w_1, \ldots, w_L\right\}$, where each component d_i corresponds to the likelihood that keyword w_i is present in document d. These likelihood values indicate the confidence that a keyword is present in the document and are manually assigned by users or automatically computed by an algorithm.

We divide users who annotate images into professional or amateur users. Amateur users annotate images with keywords as a form of entertainment – they do it as personal annotations, as annotations for other people, to gain popularity, or simply as spam. Professional users annotate images in a responsible manner for companies that, for example, want their content to be found by Arts and Design experts who will use their images in products. Thus, given the nature of amateur and professional annotations, one can compare the two types of annotations to assess annotation errors and devise a model of how amateur users annotate. In our view, the first most important contribution of this paper is the model of user manual annotations, more specifically an asymmetric binary annotation error model estimated from the difference between professional and amateur user annotations.

Automated algorithms rely on low-level image features and image understanding algorithms to infer the keywords present on images. Thus, if for any given image d we represent its low-level image features as d_V and its keyword annotations as d_W, an automated algorithm implements the transformation $p : d_V \rightarrow d_W$.

Note that while d_W contains probabilities indicating the presence of certain keywords, the vector d_V represents an image by its texture and colour features. For convenience, we shall represent an image document as $d = \left(d_V, d_W\right)$.

Now that images are represented by their keywords, the distance $\mathrm{dist}_w\left(d_W^a, d_W^b\right)$ between vectors d_W^a and d_W^b is equivalent to the semantic dissimilarity between documents d^a and d^b, i.e., $\mathrm{dist}_w\left(d_W^a, d_W^b\right) \approx 1 \,/\, \mathrm{SemSim}\left(d_W^a, d_W^b\right)$. Formally, we define the semantic dissimilarity between two documents as

$$\mathrm{dist}_w : \left[0,1\right]^L \times \left[0,1\right]^L \rightarrow \mathbb{R}_0^+, \tag{2}$$

the function in the L dimensional space that returns the distance between two keyword vectors. It is important for L to be as large as possible to accommodate as many keywords as possible in the search space to preserve the user's idea without losing any meaning. In this high-dimensional feature space, images are represented by their semantic content and semantic similarity is easily computed because semantically similar images are placed in the same neighbourhood.

Search by semantic example is a young search paradigm, e.g., [16, 24], with some variables affecting the advantages and disadvantages this search space. Thus, the second most important contribution is the careful evaluation assessing the main aspects of this novel search paradigm:

1. The influence of the accuracy of manual annotations on the computation of semantic similarity functions; this is a direct application of the model of user manual annotations.
2. Manual versus automatic methods of transforming a multimedia document into the keyword space, i.e., the $p : d \rightarrow d_W$ transformation.
3. The influence of the keyword space dimensionality on the distance functions dist_w. We include this experiment included for the completeness of our evaluation, which is related a study by Rasiwasia and Vasconcelos [26].

The organization of this paper is as follows: Section 3 proposes the model of user manual annotations and describes how to simulate different levels of user annotations accuracy. The baseline automated annotation algorithm is described in Section 4 (naïve Bayes). Section 5 describes the main steps of the search by semantic similarity framework: the querying composition and the ranking computation based on some dissimilarity function. Section 7 presents the experiments on Corel and Flickr images.

2. Related Work

Querying and ranking multimedia by semantic similarity has been a problem in computer science for many years and has been tackled with different types of paradigms: some approaches have processed data (user query and multimedia information) at feature level or at the concept level, others have exploited user interaction to refine the user query, while some have explored a combination of these paradigms.

2.1 Content based Queries

Early research in multimedia retrieval produced several systems that allowed users to search multimedia information by its content. The user would provide an example image (or an audio file) or a sketch image (or a melody humming) containing what they wanted to search for. QBIC [7] is by far the best known of such systems but several other systems appeared at the same time: VisualSeek [29]; Informedia [38]; PicHunter [3]; Virage [1]; MARS [22]; SIMPlicity [39]. This multitude of systems explored new techniques and introduced others into the area of multimedia retrieval. Many of these techniques are present in systems produced nowadays. For example, VisualSeek was one of the pioneers in Web image crawling and search, and MARS introduced a new relevance feedback method that became highly popular [28]. All these systems implement a content based search paradigm where query processing methods are based on the principle that information needs can be expressed by example images or sketch images provided by the user.

This was a good starting point and when users can provide relevant examples then it is much easier for the system to find relevant documents. Query processing algorithms start by analysing the provided examples and extract low-level features from them. Once user examples are represented by low-level features (colour, texture, regions, motion, pitch, tones or volume features), the next step is to rank the database documents by similarity. In this process, two aspects are fundamental to query processing in content-based search. The first one is the reduction of a user example to a set of low-level features. This implies that the extracted low-level features capture the user understanding of the provided example. The second aspect is the subjective notion of similarity. There is always some ambiguity as to what exactly the provided example illustrates. The problem of visual similarity was studied by Ortega et al. [22] and by many others, e.g., [11, 31, 36]. Low-level features capture part of

the knowledge represented in a multimedia document, and there are situations where search by colour, texture or shape is an excellent solution. However, low-level features might not be the ideal representation when the search is semantic and the goal is to find examples of cars, dogs, etc. This is the so called semantic gap. To overcome this problem two types of methods have been proposed: manual methods that rely on manual annotations (e.g., librarians and cataloguers) and automatic methods that rely on high-level feature representations of information.

2.2 Semantic based Queries

Systems that are aware of multimedia semantics have already flourished in the multimedia information retrieval community allowing different search paradigms. These search paradigms work on a high-level feature space that can be obtained through a manual method, an automatic method or a semi-automatic method.

Automatic algorithms are attractive as they involve a low analysis cost when compared to manual alternatives. Automatic methods are based on heuristics or on some pattern recognition algorithm. Heuristic techniques rely on metadata attached to the multimedia: for example, Lu et al. [15] analyse HTML text surrounding an image and assign the most relevant keywords to an image. Pattern recognition algorithms exploit low-level features extracted from the multimedia itself and create a model for each keyword that needs to be detected. Several techniques have been proposed in the literature: Feng, Lavrenko and Manmatha [6] proposed a Bernoulli model with a vocabulary of visual terms for each keyword, Magalhães and Rüger [17] developed a maximum entropy framework to detect multi-modal concepts, while Snoek et al. [30] proposed an SVM based multi-modal feature fusion framework. All these techniques exploit visual information to annotate visual content.

Another similar family of techniques address the problem of detecting higher level semantics such as events and places, see [14] and [27]. The extraction of this information is a task that directly addresses users' needs in social media Web sites like Flickr. These type of approaches are outside the scope of this paper.

Keyword based Queries. The direct application of keyword annotations, i.e. high-level features, allows the user to specify a set of keywords that are used to search for multimedia content containing these concepts. This is already a large step towards more semantic search engines. Although quite useful in some cases this still might be too limiting: semantic multimedia content captures knowledge

that goes beyond the simple listing of keywords. The interaction between concepts, the semantic structure and the context are aspects that humans rely on to express some information need. Natural language based queries and semantic-example based queries explore these aspects.

Natural Language based Queries. In text IR systems the user can create text based queries by combining keywords with simple Boolean expressions as in inference networks [35] or by writing a natural language query expression [4]. These types of query expressions are now possible in multimedia information retrieval owing to algorithms that can detect multimedia concepts. Recently, Town and Sinclair [34] proposed an ontology based search paradigm for visual information that allows users to express their query as a sentence, e.g., “red flower with sky background”. It relied not only on the detection of concepts but also on the information stored in the ontology regarding concept relations.

Natsev et al. [20] explored the idea of using concept-based query expansion to re-rank multimedia documents. They discuss several types of methods to expand the query with visual concepts. Another approach to query expansion in multimedia retrieval by Haubold et al. [9] uses lexical expansions of the queries. This approach exploits linguistic knowledge to increase the breadth of the query. Linguistic knowledge is a specific case of ontology-based methods that captures the semantic structure of the problem in an efficient representation. Wei and Ngo [40] proposed an ontology-enriched semantic space (OSS) for modeling and reasoning with concepts in a linear space. OSS enlightens the possibility of mapping query-to-concept and incorporates ontological knowledge from WordNet. Our framework contrasts with these approaches as we do not make use of external ontologies.

Semantic Example based Queries. In cases where users can formulate a query with a semantic example of what they want to retrieve, the system will infer the semantics of the query example and use it to search the image database. Rasiwasia et al. proposed a framework to compute semantic similarity to rank images according to the submitted query example [24, 26] and show that the semantic space offers a better retrieval precision than visual spaces based on the DCT coefficients. They start by extracting semantics with an algorithm based on a hierarchy of mixtures and compute the semantic similarity as the Kullback-Leibler divergence. More recently, Rasiwasia et al. [25] extended their work to a new technique called query-by-context-example. The semantic space is built

in the same manner but a third step smoothes the semantic space by modelling each concept as a mixture of Dirichlets. Note that they have not compared manual with automatic tags as we have done in this paper. Tesic et al. [33] address the same problem but replace the Kullback-Leibler divergence by an SVM. The SVM uses the provided examples as positive examples, while negative examples are randomly sampled from clusters in the database where the positive examples have low probability. Their results show good improvements over text-only search. Hauptman et al. [10] present an estimation of the number of concepts that is required to fill the semantic gap. They employ a topic search experiment to assess the number of required concepts to achieve a high precision retrieval system – their study suggests 3,000 concepts. This approach associates the success of semantic-multimedia IR to a single factor (number of concepts) and leaves several aspects of the problem, e.g., similarity functions, out of the analysis. Note that none of these works examine the performance of retrieval by semantic example with automatic keywords and manual keywords as we do in this paper.

3. A Model of User Manual Annotations

Several media applications, such as Flickr (www.flickr.com) or YouTube (www.youtube.com), allow users to annotate images with keywords corresponding to concepts depicted in that image. The quality of these manual annotations is dependent on the type of user – amateur or professional. With amateur users, annotations are sometimes random, incomplete or incorrect for several reasons: the user might not be rigorous, users have different understanding of the same keyword, users might have different criteria to decide the presence of concept, or it might be the result of spam annotations. Professional annotations are done by experts that received some training on how to identify concepts in multimedia content, clarified all ambiguities regarding the meaning of keywords, and have no hidden intention of incorrectly annotating content. Moreover, in most cases, professional annotations are obtained by a redundant voting scheme intended to remove disagreement between professional annotators. It constitutes an extra method of cleaning data annotations, see [37].

While professional users annotate images in a responsible manner, amateur users do it as form of entertainment resulting in some annotation errors. Thus, given the nature of amateur and professional annotations, one can compare the two types of annotations to assess annotation errors and devise a

model of how amateur users annotate. We inferred the model described in the following section from two sets of amateur and professional annotations based on a Flickr dataset [13].

3.1 An Asymmetric Binary Annotation Error Model

Most commercial stock images and photo collections such as GettyImages (www.gettyimage.com) or Corbis (www.corbis.com), have annotations with 100% accuracy produced by professional annotators. In contrast, annotations of non-commercial image collections are done by amateur users. In these scenarios, where non-professional users annotated images, one would expect to have keyword annotations with accuracies below 100%. To verify and quantify this assumption we examined a sample of 25,000 Flickr images [13] and measured the accuracy of 24 keywords. We verified that errors are not uniformly distributed: on average, users annotate 18.36% of all true annotations (true positive) and annotate 3.71% of false annotations (false positive). This obviously implies that 81.64% of true annotations are missing. Figure 3 shows the average true positive and false positive annotations. The ground-truth to compute these figures is provided by [13].

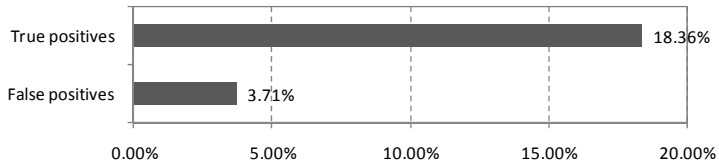

Figure 3. Accuracy of real user manual annotations.

Formally, this corresponds to the asymmetric binary channel depicted in Figure 4. This diagram illustrates that a user adds an annotation when it is present with a probability of $p = \alpha$ and adds an annotation when it is not present with a probability of $p = \beta$.

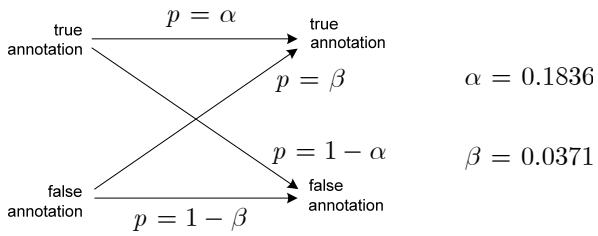

Figure 4. The manual annotation error model as an asymmetric binary channel.

From the above statistics, we verify that annotations made by amateur users follow a non-uniform error distribution: amateur users are less complete than professional users when inserting true annotations and amateur users insert some false annotations. Thus, true positives and true negatives have different probability distributions.

It should be noted that due to linguistic ambiguities and subjective interpretations, these values are an average indicator. In fact, we are dealing with an AI-complete problem because a fully automated system would have to deal with both computer vision and natural language techniques to understand images and linguistically express their content.

3.2 Manual Keyword Simulation

In this section we describe the process of simulating amateur annotations from professional annotations. We start with professional annotations and progressively add errors according to the asymmetric binary annotation error model:

- **Obtain manual annotations**: load the manual keywords from the professional annotations of the collection of N multimedia documents. This corresponds to the annotations ground-truth.
- **Add errors to annotations**: given the professional annotations, we insert errors in the annotations ground-truth according to the asymmetric binary annotation error model described in the previous section. To simulate different levels of accuracy we consider values of α and β in the range

$$\alpha \in \left[0.1836, 1.0\right], \qquad \beta \in \left[0.0, 0.0371\right]. \tag{3}$$

This corresponds to manual annotations with accuracies varying from professional level annotations $\left(\alpha = 1.0,\, \beta = 0.0\right)$ to amateur level annotations $\left(\alpha = 0.1836,\, \beta = 0.0371\right)$.

4. Automatic Annotations

An automatic annotation algorithm supports a large number of keywords so that the keyword space can wrap the semantic understanding that the user gives to a document. In this section we describe how to estimate a probability function p that automatically computes the vector

$$d_W = p_A\left(d\right) = \left(p\left(w_1 \mid d_V\right),\ \ldots\ , p\left(w_L \mid d_V\right)\right) \tag{4}$$

of L keyword probabilities from document low-level features d_V. Following the approach proposed in [17], the Bernoulli random variable w_i, represented by a naïve Bayes model, indicates the probability of observing the keyword w_i on document d_V. The model allows expressing multimodal information as described in the following sections.

Keywords are modelled as text and visual data with a naïve Bayes classifier. In our approach we look at each document as a concatenated feature vector $d_V = \left(d_{V,1},...,d_{V,M} \right)$ of visual features and a bag-of-words. The naïve Bayes classifier results from the direct application of Bayes law and independence assumptions between dimensions of a feature vector:

$$p\left(w_l \mid d_V \right) = \frac{p\left(w_l \right) \prod_{i=1}^{M} p\left(d_{V,i} \mid w_l \right)}{p\left(w_l = 0 \right) \prod_{i=1}^{M} p\left(d_{V,i} \mid w_l = 0 \right) + p\left(w_l = 1 \right) \prod_{i=1}^{M} p\left(d_{V,i} \mid w_l = 1 \right)} \tag{5}$$

Formulating naïve Bayes in the log-odds space results in

$$\log \frac{p\left(w_l = 1 \mid d \right)}{p\left(w_l = 0 \mid d \right)} = \log \frac{p\left(w_l = 1 \right)}{p\left(w_l = 0 \right)} + M \sum_{i=1}^{M} p\left(d_{V,i}^{k} \right) \log \frac{p\left(d_{V,i} \mid w_j = 1 \right)}{p\left(d_{V,i} \mid w_j = 0 \right)}, \tag{6}$$

which casts it as a linear model that avoids decision thresholds in annotation problems. Three different low-level visual features are used in our implementation: marginal HSV distribution moments, a 12 dimensional colour feature that captures the histogram of 4 central moments of each colour component distribution [23]; Gabor texture, a 16 dimensional texture feature that captures the frequency response (mean and variance) of a bank of filters at different scales and orientations [12]; and Tamura texture, a 3 dimensional texture feature composed by measures of image coarseness, contrast and directionality [12]. The images are tiled in 3 by 3 parts before extracting the low-level features, which are concatenated for each feature type.

5. Searching Images in a Keyword Space

Our goal is to devise a search space capable of representing documents according to their semantics. A keyword space is similar to other feature spaces like colour or texture feature spaces where the space structure replicates a human notion of colour or texture similarity (assuming image documents). The distinction is clear: while in the first case images are organized by their texture or colour similarity, in keyword spaces images are organized by their semantic similarity.

5.1 Querying the Keyword Space

The algorithm that parses the user request produces query vectors in the keyword space with the same characteristics as the indexed images. For each query, the system analyses the submitted example and infers a keyword vector with an automatic algorithm or a user provides the keywords present in the

example. Once we have the query keyword vector q_W, the semantic similarity between the query and a candidate document d_W is computed as the inverse of the dissimilarity $\mathrm{dist}_w\left(q_W, d_W\right)$ between the corresponding keyword vectors.

Note, that the query analysis algorithm must generate the query description in a fixed amount of time and with a low computational cost. It is commonly recognised that the system needs to answer user requests in less than one second because "this is the limit for the user's flow of thought to stay uninterrupted" [19], and it should also be able to support several users simultaneously.

5.2 Upper and Lower Bounds

Automatic annotation algorithms are not completely accurate and we do not foresee that a new algorithm will achieve a high accuracy in the near future. Thus, professional user annotations define the upper bound of the retrieval effectiveness that can be obtained in a search by semantic example scenario. Correspondingly, we deliberately chose the naïve Bayes algorithm as the automatic keyword annotation algorithm defining the lower bound of the retrieval effectiveness that can be obtained in a search by semantic example scenario.

5.3 Dissimilarity Functions

Manhattan Distance. Corresponds to the human notion of distance between two points placed over a squared grid. The Manhattan distance is the accumulated sum of the distances in each dimension,

$$D_{\mathrm{Manhattan}}\left(q_W, d_W\right) = L_1\left(q_W, d_W\right) = \sum_{i=0}^{L}\left|q_{W,i} - d_{W,i}\right|. \tag{7}$$

This distance is identical to the length of all shortest paths connecting q_w and d_w along lines parallel to the coordinate system.

Euclidean Distance. Corresponds to the human notion of distance between two points in a real coordinate space, expressed as

$$D_{\mathrm{Euclidean}}\left(q_W, d_W\right) = L_2\left(q_W, d_W\right) = \sqrt{\sum_{i=0}^{L}\left(q_{W,i} - d_{W,i}\right)^2} \tag{8}$$

Cosine Distance. Since we work in high-dimensional spaces, in geometric terms one can define the independence between two vectors as the angle between them. This gives an indication as to whether

two vectors point to a similar direction or not. This is the well known cosine similarity which becomes a dissimilarity by taking the difference to 1:

$$D_{\text{Cosine}}\left(q_W, d_W\right) = 1 - \cos\left(q_W \measuredangle d_W\right) = 1 - \frac{q_W \cdot d_W}{\left\|q_W\right\| \cdot \left\|d_W\right\|} \tag{9}$$

Geometric correlation is one of the several possible ways to measure the independence of two variables. Also, the cosine distance is a special case of Pearson correlation Coefficient when data are normalized with mean zero.

Kullback-Leibler (KL) Divergence. In statistics and information theory the KL divergence is a measure of the difference of two probability distributions. It is the distance between a "true" distribution (the query vector) to a "target" distribution (the document vector). The KL divergence is defined as

$$D_{\text{KL}}\left(q_W \;||\; d_W\right) = \sum_{i=1}^{L} p\left(q_{W,i}\right) \log \frac{p\left(q_{W,i}\right)}{p\left(d_{W,i}\right)}. \tag{10}$$

In information theory it can be interpreted as the expected extra message length needed by using a code based on the candidate distribution (the document vector) compared to using a code based on the true distribution (the query vector). Note that the KL divergence is not a true metric as it is not symmetric.

6. Evaluation

We carried out experiments on similarity ranking of semantic multimedia using three image collections. Collections were split into training and test set, and they have two levels of annotations: one used to build the keyword models corresponding to the lexicon of keywords of the keyword space; and a second level of categories corresponding to a particular query category:

- **Keywords**: multimedia annotations representing meaningful concepts in that multimedia content.
- **Categories** are groups of multimedia documents whose content concern a common meaningful theme, i.e., documents in the same category are semantically similar.

The above definitions create two types of content annotations – at the document level (keywords) and at the group of documents level (categories). Because both keywords and categories describe the content of multimedia one would assume that categories can be inferred from keywords. In our experimental framework, keywords and categories of multimedia documents are defined by each collection ground truth: keywords are used to compute semantic similarity, and categories are used to evaluate semantic similarity.

6.1 Collections

Corel Images. This dataset was compiled by Duygulu et al. [5] from a set of COREL Stock Photo CDs. The dataset has some visually similar concepts (*jet, plane, Boeing*), and some concepts have a limited number examples (10 or less). The collection is split into a training set of 4,500 images and a test set of 500 images. Each image is annotated with one to five keywords from a vocabulary of 371 keywords. Only keywords with at least one image both in the test and training set were used, which reduces the size of the vocabulary to 260 keywords. The collection is already organized into 50 image categories, such as *rural France*, *Galapagos wildlife* and *nesting birds*, Despite the critics that this dataset has received [32] many others (e.g., [6, 25]) have used this dataset as it constitutes a good reference and permits the comparison of different algorithms.

	Training Examples	Test Examples	Query Examples	Keywords	Categories
Corel Images	4,500	500	All test images	260	50
mirFlickr08	9,973	15,027	1,842 test images	126	11
NUS-WIDE	161,789	107,859	708 test images	81	6

Table 1. Summary of collections used on the experiments.

mirFlickr08 Images. To test semantic similarity on an amateur image collection we used the Flickr data provided by Huiskes and Lew [13]. It contains 25,000 images annotated by users with a folksonomy and annotated by professionals with 24 hierarchical-keywords. From the folksonomy we selected 126 tags to build the keyword space and from the 24 professional hierarchical-keywords we selected the 11 top-level keywords as query categories (*people, sky, water, architecture plant, food, transport, night, indoor, sunset* and *animals*). The 25,000 images were randomly split into 9,973

training images and 15,027 test images. From the test set only images with one category were used as query examples to avoid ambiguities in the evaluation.

NUS-WIDE. The final set of experiments was conducted on a large-scale dataset with more than a quarter million Flickr images provided by Chua et. al [2]. The dataset is composed by 269,648 images annotated with 1000 user keywords. A controlled annotation effort was conducted by the authors to annotate images with 81 concepts belonging to six categories (events, program, scene, people, objects and graphics). We use this manual annotation effort as ground truth, or professional annotations. The low-level features employed in the experiments were the 128-D wavelet texture and 225-D block-wise color moments.

6.2 Experiments Design

Before proceeding to the semantic dissimilarity evaluation experiments, we first learned the naïve Bayes keyword models on the training set of each collection. Dissimilarity evaluation is done on the collections test set and with the corresponding keyword models. Note that some images belong to multiple categories. For this reason we only used query images belonging to only one category: 708 images on the NUS-WIDE dataset and 1,842 on the mirFlickr08 dataset. These single-category images were used as query examples to rank the remaining test images by semantic-similarity, 107,859 on the NUS-WIDE dataset and 15,027 on the mirFlickr08. Formally, the evaluation protocol was the following:

1. Learn the naïve-Bayes model for each keyword on the training set of each collection (260 models for Corel, 126 for mirFlickr08 and 81 for NUS-WIDE). Note that we do not reuse the training set as the search database in contrast to Rasiwasia et al. [24].
2. Submit a test document as a query example to rank the remaining test examples by semantic similarity.
3. Compute keyword annotations for both documents and query:
    a. Automatic keywords with the naïve-Bayes algorithm (260 keywords for Corel, 126 for mirFlickr08 and 81 for NUS-WIDE).
    b. Manual keywords with varying accuracy.
4. Rank documents by their semantic similarity to the query example according to a given dissimilarity function.

5. Use the category of the query example as relevance judgment.
6. Repeat steps 2 to 5 for all test examples.

The above methodology is repeated for each dissimilarity function, dataset, and keyword vector computation algorithm. This way we isolate the variables of the problem that we are interested in studying: semantic-similarity functions, influence of manual annotations accuracy, influence of the number of keywords.

Average precision and precision-recall curves are the measures used for comparing the different ranks to a particular query. Mean average precision (MAP), by taking the mean over all queries allows comparing different systems. Conceptually average precision is the area under the precision recall curve, the later is calculated by averaging the precision found at every relevant document. Average precision as a performance measure has the advantage that it gives a greater weight to results retrieved early.

6.3 Results and Discussion

6.3.1 Automated Annotations

These results are obtained with the output of the naïve Bayes classifier and for the keyword space with the maximum number of keywords – it evaluates the dissimilarity functions in a fully automated scenario. The MAP obtained with Cosine was consistently better than the others as we can see from the precision-recall graphs in Figure 5 and the summary of MAP values in Table 2. The differences on the precision-recall curves from one dataset to the other is justified by the fact that in the Corel dataset there aren't many examples for each category (10 per category), while in the mirFlickr08 dataset there are several examples for the same category and each image can have more than one category. This disparity in the number of relevant examples justifies the observed differences from one dataset to the other. Note that the cosine similarity function performs quite well in both datasets.

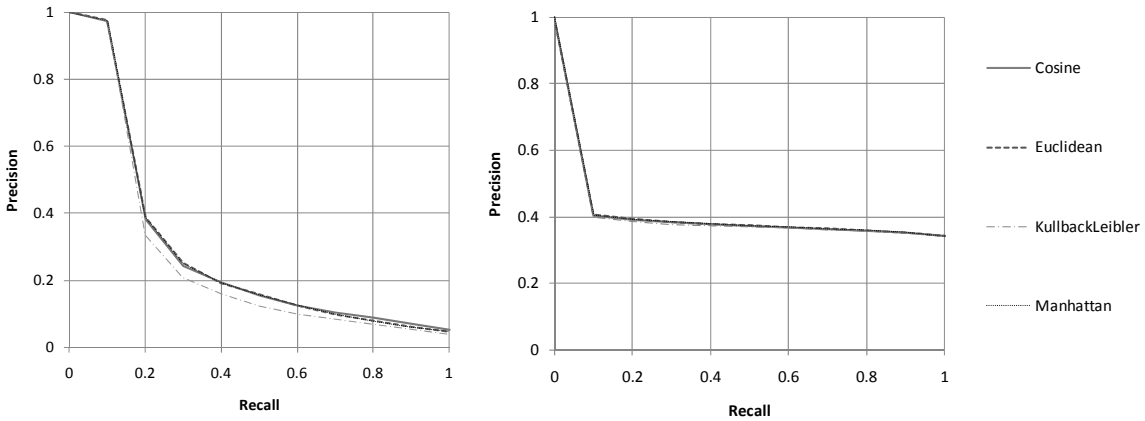

Figure 5. Dissimilarity functions retrieval evaluation for Corel (left) and mirFlickr08 (right).

6.3.2 Manual Annotations

The evaluation presented in this section creates a keyword space with the manual annotations. An important distinction concerning the manual annotations must be noted: in the Corel dataset manual annotations were done by experts and in the mirFlickr08 dataset manual annotations were done by amateur users. This experiment allows assessing how different dissimilarity functions behave in the presence of user generated annotations.

In the Corel dataset, the user-keyword results provide us with a good approximation to the retrieval effectiveness upper bound. The upper bound is obviously dependent on the similarity function: Figure 6 illustrates the precision-recall graphs, and Table 2 and Table 3 summarize the MAP values. The most noticeable fact is that even with completely accurate annotations we cannot pass a value of 50% of MAP. Most metrics have a similar pattern because most functions are based on some linear combination of individual keywords.

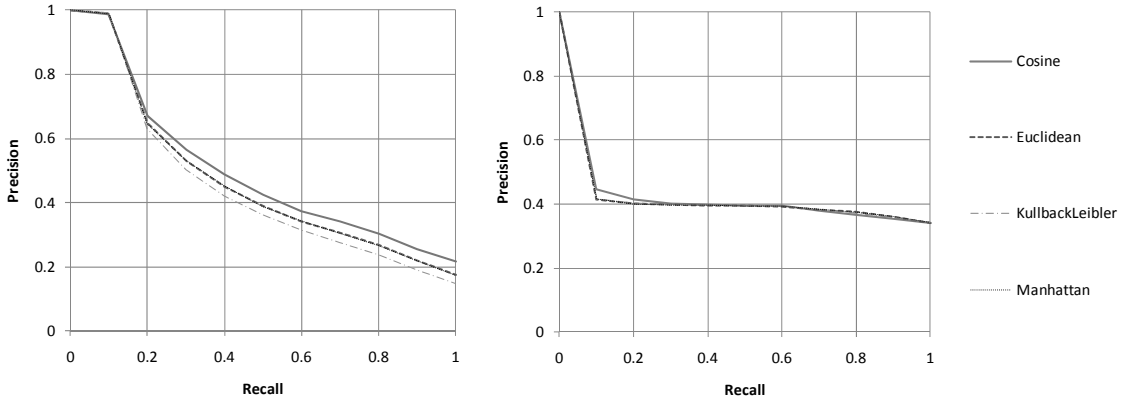

Figure 6. Evaluation with user keywords for Corel (left) and mirFlickr08 (right).

In the mirFlickr08 dataset, the user-keyword results show that precision-recall is not much better than automatic-keywords. Note that in this case, we are using amateur level annotations which have ~18% of true positives and ~3% of false positives. These results allow us to draw some conclusions and are a good source of many new research questions. There is an obvious gap between the annotated keywords and the unknown query categories. Note that this is different from the notion of semantic gap between low-level features and keywords. It is actually a gap among concepts, in this case between the annotated keywords and the user information need. This points to two possible solutions: increase the number of keywords or investigate keyword spaces to represent multimedia information and possible similarity metrics. The first solution is the simple application of brute force, hoping to have comprehensive annotations with better high-level keyword extractors. The second solution suggests investigating similarity functions that incorporate keyword interdependencies and are robust to noisy document descriptions.

6.3.3 Manual Keywords Accuracy

In this experiment we study the influence of the accuracy of user annotations on the retrieval MAP. We start with professional level manual annotations and insert errors (false positives and false negatives) according to the model described in Section 3 (the asymmetric binary annotation error model). The model inserts false positives and false negatives at asymmetric rates – these error rates are inferred from the mirFlickr08 dataset as described in Section 3. Note that the mirFlickr08 annotations already have errors, so, only the Corel dataset can be used in this experiment. Thus, we vary the annotation accuracy from a professional level (100% true positives and 0% false positives) to an amateur level (18.8% true positives and 3.7% false positives). This procedure simulates different user annotations with a wide range of accuracies, i.e., from amateur level to expert level. This variation is actually real as incorrect annotations might occur for different reasons, e.g., interpretation of a keyword, spam, or incomplete annotations. Figure 7 displays the results of searches that rely on user keywords with a varying degree of annotation accuracy. It is noticeable that the presence of incorrect annotations affects the performance of the retrieval by semantic example.

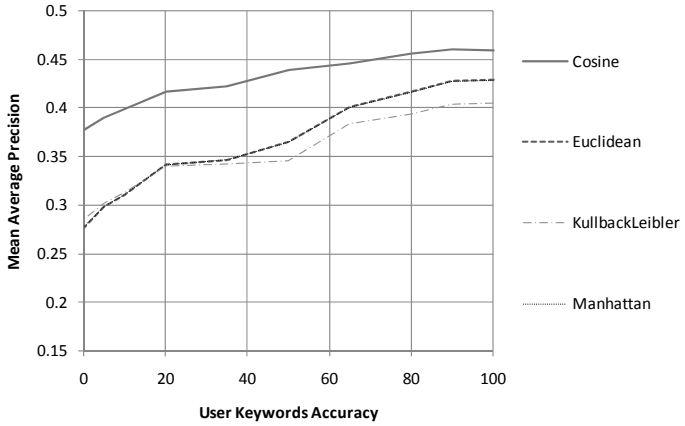

Figure 7. Effect of user keywords accuracy on Corel.

6.3.4 Manual Annotations versus Automated Annotations

The retrieval upper bound of search by semantic similarity is computed with completely accurate user keywords. This bound is specific for the set of keywords and categories. In the Corel collection the upper bound with the cosine distance is 0.464, and 0.395 in the mirFlickr08 collection. Table 2 shows the MAP of ranking by similarity that uses the naïve Bayes classifier and various dissimilarity functions for the Corel collection. Note that there is a considerable difference between automatic keywords and professional-level manual keywords. This fact is also observable on the precision-recall curves behaviour for professional-level manual-keywords (Figure 6) and automatic keywords (Figure 5).

Dissimilarity	Corel Images	
	Automatic keywords	(Pro) Manual keywords
Manhattan	0.230	0.435
Euclidean	0.226	0.435
Cosine	**0.235**	**0.464**
Kullback-Leibler	0.210	0.415

Table 2. MAP of automatic keywords and user keywords on the Corel dataset.

In the mirFlickr08 collection, we note that the difference in terms of MAP between amateur level manual-keywords and automatic keywords is not significant. The precision-recall behaviour of search by semantic example confirms this fact, see Figure 6 and Figure 5. However, further investigation reveals that when ranks are evaluated on the top 20 retrieved documents, the manual annotations are actually much better than automatic annotations. The encouraging news here is that we are comparing

a simple automatic annotation algorithm trained on noisy data with manual annotations, and one would expect there to be scope for improvement.

Dissimilarity	mirFlickr08 Images			
	MAP for automatic keywords	MAP for manual keywords	MP@20 for automatic keywords	MP@20 for manual keywords
Manhattan	**0.372**	0.388	**0.444**	0.557
Euclidean	0.372	0.388	0.444	0.557
Cosine	0.368	**0.395**	0.440	**0.570**
Kullback-Leibler	0.364	0.388	0.434	0.562

Table 3. MAP and MP@20 of automatic and manual keywords on the mirFlickr08 dataset.

6.3.5 Keyword Space Dimensionality

The previous evaluation used the space with the full range of keywords, independently of their value to the ranking process. This affects accuracy as some of the keywords are either noise or are irrelevant to most searches. In this section we study the effect of removing noisy keywords from the keyword space in the ranking process. The keyword space is built by progressively adding keywords according to the retrieval precision of the corresponding classifier. Thus, keywords with higher average precision are added first. This is similar to unsupervised feature selection that is exclusively based on the accuracy of the keywords. Thus, we do not use the query categories to select the keywords (e.g., use the final objective to select dimensions like in normal feature selection) because in this experiment one should not know the query category beforehand.

In the Corel collection, Figure 8, we can observe that the first keywords carry more information value – as lower precision keywords are added to the keyword space the MAP increases. It is important to note the robustness to noise that this experiment illustrates: the Cosine measure continues to show a good robustness to noise.

The same general conclusions can be drawn from the MAP curves on the mirFlickr08 collection, Figure 9. However, in this dataset we see that the MAP curves are more unstable while the MP@20 is relatively stable. This is probably because in the mirFlickr08 dataset keywords are actually very noisy and the addition of a new keyword has a greater influence in the retrieval performance.

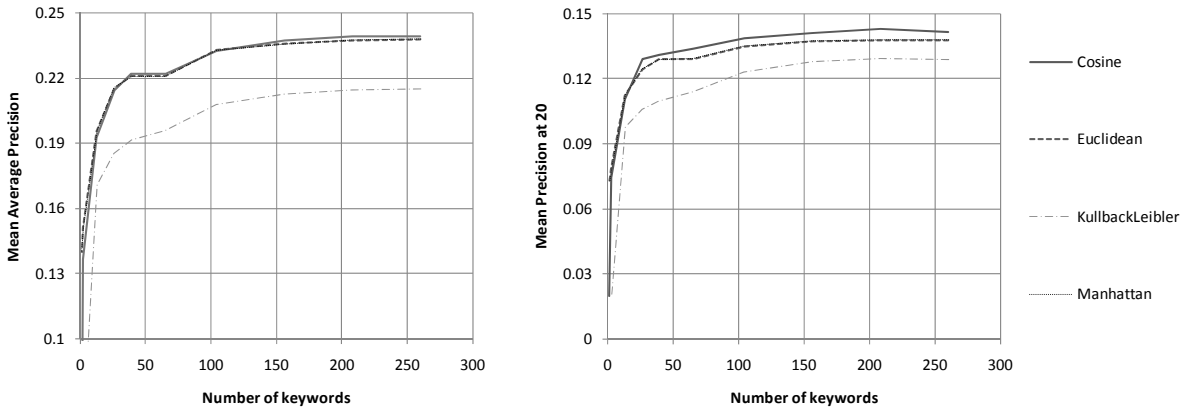

Figure 8. Dimensionality of the keyword space on Corel: MAP and MP@20.

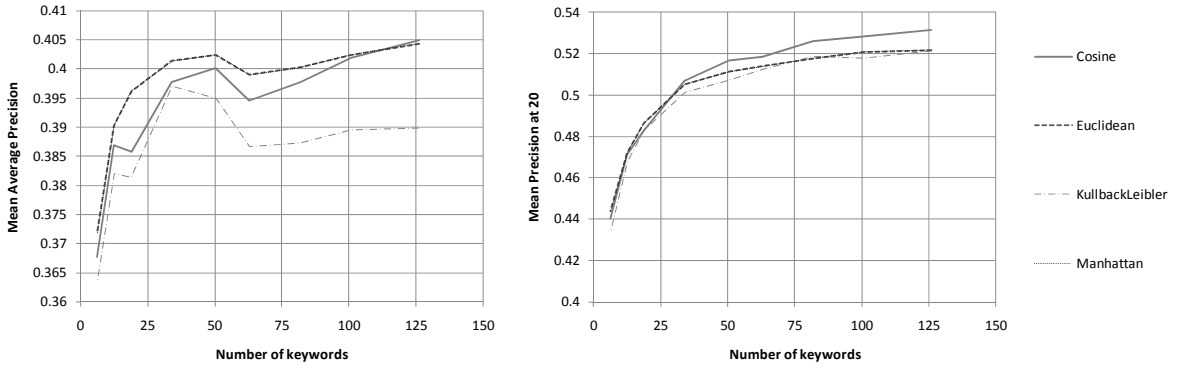

Figure 9. Dimensionality of the keyword space on mirFlickr08 Images: MAP and MP@20.

6.3.6 Large-scale experiments

To verify the previous results on a real world large-scale dataset we conducted experiments on the NUS-WIDE dataset containing 269,648 Flickr images. First we trained the classifiers on the 81 concepts using 161,789 training images. After training the classifiers, the training set was no longer used, from this point further we only used the test set. Those 81 concepts belong to 6 categories that we used to evaluate the retrieval by semantic similarity. From the test set we chose 708 query images to search the remaining 107,859 images for images of the same category. As noted previously query images belong to a single category because images with more than one category would cause an ambiguous query with multiple possible ranks.

Results are summarized on Table 4 and Figure 10. The most evident fact from this data is that amateur-level manual annotations offered better performance for top the 20 ranked results (MP@20). It almost reached 100% performance while automated methods did not reach 50% performance. In

our view, this is related to the fact that images from one category have consistent annotations among each other.

Dissimilarity	NUS-WIDE Images			
	MAP for automatic keywords	MAP for manual keywords	MP@20 for automatic keywords	MP@20 for manual keywords
Manhattan	0.365	0.268	0.378	0.881
Euclidean	0.367	0.271	0.387	0.900
Cosine	0.361	0.353	0.370	0.950
Kullback-Leibler	0.364	0.282	0.370	0.881

Table 4. MAP and MP@20 of automatic and manual keywords on the Flickr dataset.

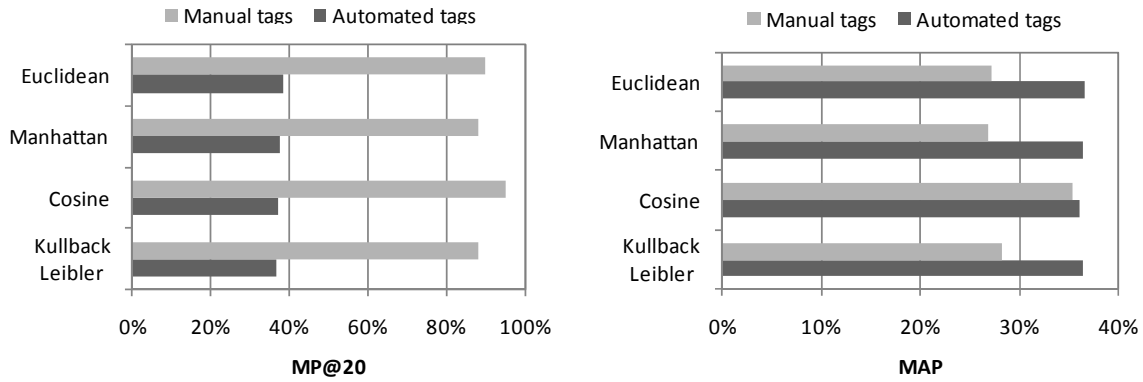

Figure 10. Large-scale retrieval performance comparison.

For full rank measures (MAP) retrieval by semantic similarity with automatic annotations is similar or better than amateur-level manual annotations. Noisy tags are more critical for lower positions in the rank. This is also the reason why social-tagging has been so successful: noisy tags do not affect top ranked results but they greatly affect longer ranks. This is not a critical aspect for social-media applications where only the top 20 or 50 results are actually important for the user.

6.3.7 Uncontrolled Vocabularies

Non-professional users annotate content with every keyword that they wish. This generates uncontrolled vocabularies, called folksonomies. Their advantages are obvious from the multitude of social-media Web applications that apply it successfully. Marlow et al. [18] proposed a taxonomy to help in the analysis, design and evaluation of these applications, hence, confirming the variety of Web 2.0 applications. However, the uncontrolled nature of folksonomies causes many problems in the

computation of semantic dissimilarities between two multimedia documents. First, it is never possible to know the correct meaning that a user gives to a keyword, e.g., the keyword football means different sports for different cultures. Second, the user might dishonestly annotate a document with a popular keyword to attract other users. Third, users might have different criteria to annotate documents, e.g., some users might rigorously annotate all keywords while others might skip the obvious ones. The direct application of uncontrolled vocabularies offer a good solution to the problem of multimedia annotation but it is not a solution that delivers 100% accuracy. Thus, understanding how users annotate as a whole becomes a critical task to exploit the full potential of uncontrolled vocabularies, [21].

With automatic methods these problems do not exist: algorithmic errors are always consistent for the same type of content, e.g., similar content suffer the same type of annotations noise. Thus, we believe that the results of the proposed framework show that automatic methods have an important role in the semantic exploration of multimedia content.

6.3.8 Semantic Relevance

Assessing the user information needs from an example is always a difficult task. We assumed information needs can be represented by a set of keywords extracted from the example and evaluated with categories. The measures of precision and recall use a binary relevance model to identify relevant and non relevant documents. However, in the current scenario the relevance of a document is difficult to measure because semantic relevance is gradual and contextual. The problem is even more complex for several reasons, e.g., for a particular query an image with one matching keyword might be more meaningful than an image with two matching keywords; an image might belong to different categories but only one category is the required one. This ambiguity in the interpretation of the meaning of a keywords is actually what humans explore as way to formalize their abstract idea, see [8].

This is a consequence of the two problems of semantic relevance judgments: incompleteness and type of relevance judgment. Incompleteness of relevance judgments derives from the fact that not all labels present in a document are marked as present. The second problem concerns the type of relevance judgments (keywords and query categories) used in these experiments. Thus, multi-level relevance

model would be more adequate to learn the keyword models and ranked relevance is more adequate to investigate functions for semantic similarity. Note that, although binary relevance judgments are an approximation to this ideal situation, they still provide a good research setup.

7. Conclusions

This paper addressed the problem of searching images by semantic similarity in a keyword space. Automatically managing multimedia by their keyword annotations is a complex task involving a long chain of information processing algorithms. We presented experiments to analyze different aspects of the process: (1) comparison of amateur to professional annotations, (2) accuracy of manual annotations versus automatic annotations, (3) the dimensionality of the keyword space, and (4) manual annotations with different accuracies resulting from incorrect annotations. Our evaluation allows us to draw the following observations:

1. We verified that annotations made by amateur users follow a non-uniform error distribution: on average, users annotate 18.36% of all true keywords and annotated 3.71% of false keywords. Based on these facts, we propose an asymmetric binary annotation error model;
2. The mean average precision (MAP) metric indicates that in a real dataset (Flickr), retrieval by semantic similarity with amateur-level manual annotations is comparable to automatic annotations (the lower bound);
3. The mean precision at 20 retrieved documents (MP@20) metric showed that in a real dataset (the NUS-WIDE Flickr images dataset), retrieval by semantic similarity with amateur-level manual annotations performs much better than automatic annotations;
4. The increase of the keyword space dimensionality, results in a corresponding increase in retrieval effectiveness – the increase is stable in terms of MP@20 and less stable in terms of MAP.

The results of the proposed framework show that automatic methods have an important role in the semantic exploration of multimedia content. Finally, we outline some recommendations inferred from our experiments:

- In the presence of manual annotations, the Cosine dissimilarity function is the best choice;
- In the presence of automatic annotations, the Manhattan dissimilarity function is the best choice;

- Automated annotations offer an inexpensive solution to discover relevant images with no annotations, i.e., to increase recall.

These conclusions together with the experiments results shed some light on the problem of semantically comparing two multimedia documents.

8. References

[1] J. R. Bach, C. Fuller, A. Gupta, A. Hampapur, B. Horowitz, R. Humphrey, R. C. Jain, and C.-F. Shu, "Virage image search engine: an open framework for image management," Proc. SPIE Int. Soc. Opt. Eng, San Jose, 1996.

[2] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng, "NUS-WIDE: a real-world web image database from National University of Singapore," Proceeding of the ACM International Conference on Image and Video Retrieval, Santorini, Fira, Greece, 2009.

[3] I. J. Cox, M. L. Miller, S. M. Omohundro, and P. N. Yianilos, "PicHunter: Bayesian relevance feedback for image retrieval," Proceedings of the International Conference on Pattern Recognition, 1996.

[4] W. B. Croft, H. R. Turtle, and D. D. Lewis, "The use of phrases and structured queries in information retrieval," ACM SIGIR Conf. on research and development in information retrieval, Chicago, Illinois, United States, 1991.

[5] P. Duygulu, K. Barnard, N. de Freitas, and D. Forsyth, "Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary," European Conf. on Computer Vision, Copenhagen, Denmark, 2002.

[6] S. L. Feng, V. Lavrenko, and R. Manmatha, "Multiple Bernoulli relevance models for image and video annotation," IEEE Conf. on Computer Vision and Pattern Recognition, Cambridge, UK, 2004.

[7] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yanker, "Query by image and video content: the QBIC system," IEEE Computer, vol. 28, pp. 23-32, 1995.

[8] J. S. Hare, P. H. Lewis, P. G. B. Enser, and C. J. Sandom, "Mind the Gap: Another look at the problem of the semantic gap in image retrieval," SPIE Conference on Multimedia Content Analysis, Management and Retrieval, San Jose, California, USA, 2006.

[9] A. Haubold, A. Natsev, and M. Naphade, "Semantic multimedia retrieval using lexical query expansion and model-based re-ranking," IEEE Int'l Conference on Multimedia and Expo, Toronto, Canada, 2006.

[10] A. Hauptmann, R. Yan, and W.-H. Lin, "How many high-level conecpts will fill the semantic gap in news video retrieval?," ACM Conf. on image and video retrieval, Amsterdam, The Netherlands, 2007.

[11] D. Heesch and S. Rüger, "Three interfaces for content-based access to image collections," Int'l Conf. on Image and Video Retrieval, Dublin, Ireland, 2004.

[12] P. Howarth and S. Rüger, "Evaluation of texture features for content-based image retrieval," Int'l Conf. on Image and Video Retrieval, Dublin, Ireland, 2004.

[13] M. J. Huiskes and M. S. Lew, "The MIR Flickr Retrieval Evaluation," ACM International Conference on Multimedia Information Retrieval Vancouver, Canada, 2008.

[14] L. Kennedy, M. Naaman, S. Ahern, R. Nair, and T. Rattenbury, "How Flickr helps us make sense of the world: context and content in community-contributed media collections," ACM Multimedia, Augsburg, Germany, 2007.

[15] Y. Lu, C. Hu, X. Zhu, H. Zhang, and Q. Yang, "A unified framework for semantics and feature based relevance feedback in image retrieval systems," ACM Conf. on Multimedia, Los Angeles, CA, USA, 2000.

[16] J. Magalhães, F. Ciravegna, and S. Rüger, "Exploring multimedia in a keyword space," ACM Multimedia, Vancouver, Canada, 2008.

[17] J. Magalhães and S. Rüger, "Information-theoretic semantic multimedia indexing," ACM Conf. on Image and Video Retrieval, Amsterdam, The Netherlands, 2007.

[18] C. Marlow, M. Naaman, d. boyd, and M. Davis, "HT06, Tagging Paper, Taxonomy, Flickr, Academic Article, To Read," Conference on Hypertext and Hypermedia, Odense, Denmark, 2006.

[19] R. B. Miller, "Response time in man-computer conversational transactions," AFIPS Fall Joint Computer Conference 1968.

[20] A. Natsev, A. Haubold, J. Tesic, L. Xie, and R. Yan, "Semantic concept-based query expansion and re-ranking for multimedia retrieval," ACM Conf. on Multimedia, Augsburg, Germany, 2007.

[21] R.-A. Negoescu and D. Gatica-Perez, "Analyzing Flickr groups," ACM Conference on Image and Video Retrieval, Niagara Falls, Ontario, Canada, 2008.

[22] M. Ortega, Y. Rui, K. Chakrabarti, S. Mehrotra, and T. S. Huang, "Supporting similarity queries in MARS," ACM Conf. on Multimedia, Seattle, Washington, United States, 1997.

[23] M. J. Pickering, D. Heesch, R. O'Callaghan, S. Rüger, and D. Bull, "Video retrieval using global features in keyframes," TREC Text Retrieval Conf. , Gaithersburg, USA, 2002.

[24] N. Rasiwasia, P. Moreno, and N. Vasconcelos, "Bridiging the gap: Query by semantic example," IEEE Transactions on Multimedia, vol. 9, pp. 923-938, 2007.

[25] N. Rasiwasia and N. Vasconcelos, "Image Retrieval using Query by Contextual Example," ACM Multimedia Information Retrieval, Vancouver, Canada, 2008.

[26] N. Rasiwasia and N. Vasconcelos, "A study of query by semantic example," Workshop SLAM at CVPR, 2008.

[27] T. Rattenbury, N. Good, and M. Naaman, "Towards automatic extraction of event and place semantics from Flickr tags," ACM SIGIR, Amsterdam, The Netherlands, 2007.

[28] Y. Rui, T. Huang, M. Ortega, and S. Mehrota, "Relevance feedback: a power toll for interactive content-based image retrieval," IEEE Transactions on Circuits Systems and Video Technology, vol. 8, pp. 644-655, 1998.

[29] J. R. Smith and S.-F. Chang, "VisualSEEk: a fully automated content-based image query system," ACM Conf. on Multimedia, Boston, MA, USA, 1996.

[30] C. G. M. Snoek, M. Worring, J.-M. Geusebroek, D. C. Koelma, F. J. Seinstra, and A. W. M. Smeulders, "The semantic pathfinder: using an authoring metaphor for generic multimedia indexing," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 28, pp. 1678-1689, 2006.

[31] M. J. Swain and D. H. Ballard, "Color indexing," International Journal of Computer Vision, vol. 7, pp. 11-32, 1991.

[32] J. Tang and P. H. Lewis, "A study of quality issues for image auto-annotation with the Corel data-set," IEEE Trans. on Circuits and Systems for Video Technology, vol. 17, 2007.

[33] J. Tesic, A. Natsev, and J. R. Smith, "Cluster-based data modelling for semantic video search," ACM Conf. on Image and Video Retrieval, Amsterdam, The Netherlands, 2007.

[34] C. P. Town and D. A. Sinclair, "Language-based querying of image collections on the basis of an extensible ontology," International Journal of Image and Vision Computing, vol. 22, pp. 251-267, 2004.

[35] H. Turtle and W. B. Croft, "Evaluation of an inference network-based retrieval model," ACM Trans. Inf. Syst, vol. 9, pp. 187-222, 1991.

[36] N. Vasconcelos, "On the efficient evaluation of probabilistic similarity functions for image retrieval," IEEE Transactions on Information Theory, vol. 50, pp. 1482-1496, 2004.

[37] T. Volkmer, J. A. Thom, and S. M. M. Tahaghoghi, "Modeling human judgment of digital imagery for multimedia retrieval," IEEE Transactions on Multimedia, vol. 9, pp. 967-974, 2007.

[38] H. D. Wactlar, T. Kanade, M. A. Smith, and S. M. Stevens, "Intelligent access to digital video: Informedia project," IEEE Computer vol. 29, pp. 46-52, 1996.

[39] J. Z. Wang, J. Li, and G. Wiederhold, "SIMPLIcity: semantics-sensitive integrated matching for picture libraries," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, pp. 947-963, 2001.

[40] X.-Y. Wei and C.-W. Ngo, "Ontology-enriched semantic space for video search," ACM Multimedia, Augsburg, Germany, 2007.