Appears in 3rd International Workshop on Semantic Learning and Applications in Multimedia, Anchorage, 2008.

A Study of Query by Semantic Example

Nikhil Rasiwasia Nuno Vasconcelos
Department of Electrical and Computer Engineering
University of California, San Diego
nikux@ucsd.edu, nuno@ece.ucsd.edu

Abstract

In recent years, query-by-semantic-example (QBSE) has become a popular approach to do content based image retrieval [20, 23, 18]. QBSE extends the well established query-by-example retrieval paradigm to the semantic domain. While various authors have pointed out the benefits of QBSE, there are still various open questions with respect to this paradigm. These include a lack of precise understanding of how the overall performance depends on various different parameters of the system. In this work, we present a systematic experimental study of the QBSE framework. This can be broadly divided into three categories. First, we examine the space of low-level visual features for its effects on the retrieval performance. Second, we study the space of learned semantic concepts, herein denoted as the "semantic space", and show that not all semantic concepts are equally informative for retrieval. Finally, we present a study of the intrinsic structure of the semantic space, by analyzing the contextual relationships between semantic concepts and show that this intrinsic structure is crucial for the performance improvements.

1. Introduction

Content based image retrieval has been an active subject of research over the last decades [5], when three different retrieval paradigms have gained popularity. In the early years, the predominant paradigm was query-by-visual-example (QBVE) [11, 25, 21, 22]. Under QBVE, each image is decomposed into a number of *low-level visual features* (e.g. color, texture or shape histograms) and retrieval is based on an example (query) image. One significant limitation of this paradigm is that the similarity of low-level image descriptors does not always correlate with human judgments of similarity. This motivated the introduction of query-by-keyword paradigm [1, 6, 2, 3]. Under this paradigm, users specify their queries through a natural language description of the desired concepts. Such a paradigm requires the images to be annotated with semantic keywords. Since manual image annotation is a labor intensive process, research was focused on *semantic labeling systems* [1, 6, 2, 3]. The advantages of query-by-keyword lies in its ability to perform retrieval at a higher level of query abstraction. However, it is limited by the size of the vocabulary of concepts which the retrieval system is trained to recognize.

Realizing that the shortcomings and advantages of QVBE and query-by-keyword are in many respects complementary, several authors have proposed their combination which is rapidly gaining popularity [26, 27, 20, 24, 23, 18]. This combination extends the query-by-example paradigm to the semantic domain, and can be formulated as a two stage process. In the first stage, as is common in query-by-keyword, images are fed to a semantic labeling system which detects pre-defined semantic concepts. An image is then represented as a vector of posterior concept probabilities. These probabilities can be interpreted as *high-level semantic features*, rendered by projection of the image onto the abstract space of semantic concepts supported by the labeling system. This space is commonly referred to as the "semantic space" [24, 23] or the "model space" [26, 16]. The second stage performs all classification decisions on this higher-level semantic space, using the query-by-example principle: the concept probability vector of the query image is used to find the database images with concept distributions closest to that of the query. Using the terminology of [24], we refer to this framework as "query-by-semantic-example" (QBSE) in the remainder of this work.

While various authors have pointed out the benefits of QBSE, there are still various open questions with respect to this paradigm. These include a lack of precise understanding of how the overall performance depends on the accuracy of each of the stages, and how the performance improvements are related to the structure of the intermediate semantic space. In this work, we present the results of a systematic experimental study of the performance of a QBSE system, which addresses these questions. The experiments undertaken can be broadly divided into three categories: studies of how 1) the low-level visual space, and 2) the high-level semantic space affect the overall retrieval performance, and 3) a study of the intrinsic structure of the semantic space. To analyze the impact of the low-level visual space, we have built semantic spaces from various combinations of standard representations for color and texture. With regards to color, we consider a number of colorspaces, viz. “YBR” (luminance, normalized blue, normalized red), perceptually uniform “LAB”, “HSV” (hue, saturation, luminance) and “Y” (luminance only). In what concerns texture, we apply a standard feature transformation (in this paper we use the discrete cosine transform, although similar results were obtained with wavelets) and vary the number of dimensions in a coarse-to-fine manner. By varying the dimensionality (adding more or less high-frequencies) it is possible to vary the accuracy of the low-level visual representation, and examine its impact on the overall retrieval accuracy.

To analyze the impact of the high-level semantic space, we then vary the dimensions of the latter, by gradually eliminating non-informative semantic features. We show that the overall retrieval performance is directly proportional to the number of informative dimensions of the semantic space. Finally, we characterize the intrinsic structure of this space, by analyzing contextual relationships between concepts. We also show that these relationships play a crucial role in the retrieval operation. This is further substantiated by building a semantic space devoid of any (meaningful) structure, which is shown to obliterate the benefits (in retrieval accuracy) of QBSE over QBVE.

The paper is organized as follows. Section 2 discusses the related work on semantic spaces and QBSE. In Section 3, we review implementations of QBVE [28] and query-by-keyword [3], based on the minimum probability of error (MPE) formulation of image retrieval [28]. This MPE formulation has also been successfully applied to QBSE [23], which we review in Section 4. An extensive experimental study of the performance of QBSE is presented in Section 5. Finally, we present conclusions, and some ideas for future research in Section 6.

2. Related Work

The idea of representing documents on semantic spaces is commonly used in information retrieval [8]. In image retrieval, earliest efforts on building semantic spaces were based on semantic information extracted from metadata [12]. Later on, semantic spaces were also constructed with resort to active learning, based on user relevance feedback [9, 17]. However, it is not always clear how the learned semantic information could be combined with the visual search at the core of the retrieval system.

A solution to this was pioneered by Smith et al. [26] by extending query-by-example to the semantic domain. This was done by learning a semantic space, by learning a separate statistical model for each concept, and performing query-by-example in the space of resulting semantic concepts. They later extended QBSE to perform retrieval on video databases in [27, 20]. A semantic space representation of images was also used by Lu et al. in [16] to perform automatic image annotation, rather than image retrieval. A QBSE system based on the semantic labeling algorithm of [3] was presented in [24]. The authors highlight the superiority of QBSE over QBVE on benchmark datasets. In [23], the authors showed that this superiority also holds outside the space of learned semantic concepts, using multiple image queries. Another approach to QBSE, using the semantic labeling system of [19], is presented in [18].

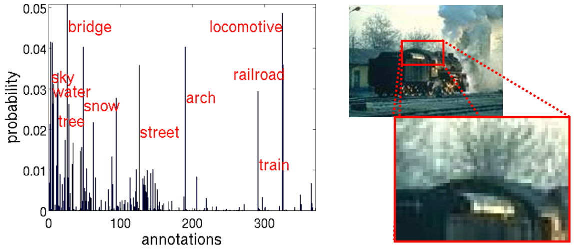


Figure 1. An image and its associated semantic representation. Note that, while most of the concepts of largest probability are present in the image, significant probability is also assigned to “bridge” and “arch”. This is due to the presence of a geometric structure similar to that of “bridge” and “arch”, shown on the image close-up.

Although laying the foundations for QBSE, these previous works lack a systematic study of the QBSE paradigm. In this work, using the QBSE implementation of [23], we address this problem by studying some of the parameters that affect the performance of a QBSE system. In particular, we examine the dependence of the retrieval performance on both the low-level visual space and the high-level semantic space. We also characterize the intrinsic structure of the semantic space, by analyzing the contextual relationships between the semantic concepts. We use the implementation of [23], because it allows the control of various parameters of the system, for example, the dimensions of the two spaces, in a systematic and fine-grained manner.

3. Minimum probability of error retrieval

The retrieval architecture adopted for the implementation of all retrieval strategies discussed in this work is that of minimum probability of error (MPE) retrieval [28]. We adopt this architecture as it has been shown to perform well in all retrieval contexts discussed herein: QBVE [28], query-by-keyword [3] and QBSE [23]. Moreover, it is also conducive to the examination of various relevant parameters of a QBSE system. We start by briefly reviewing this architecture.

3.1. Visual-level retrieval system

Under the MPE framework, images are characterized as observations from a random variable $\mathbf{X}$, defined on some visual feature space $\mathcal{X}$. The starting point for an image retrieval system is an image database $\mathcal{D} = \{\mathcal{I}_1, \ldots, \mathcal{I}_D\}$. In the absence of any labels, each image is considered an observation from a different class. The class is determined by a random variable Y defined on $\{1, \ldots, D\}$. Given a query image $\mathcal{I}_q$, the MPE decision rule for retrieval is to assign it to the class of largest posterior probability, i.e.

$$y^* = \arg\max_y P_{Y|\mathbf{X}}(y|\mathcal{I}_q). \tag{1}$$

At the visual level, each image is represented as a set of n *feature vectors* $\mathcal{I} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}, \mathbf{x}_i \in \mathcal{X}$. It is assumed that the feature vectors which compose any image $\mathcal{I}$ are sampled independently.

$$P_{\mathbf{X}|Y}(\mathcal{I}|y) = \prod_j P_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}_j|y). \tag{2}$$

Although any type of visual features are acceptable, we only consider *localized features*, i.e., features of limited spatial support.

In this work, the distributions $P_{\mathbf{X}|Y}(\mathbf{x}|y)$ are modeled as Gaussian mixtures. The parameters of the distributions are learned from the training sample (the n feature vectors $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ per image) using the well known expectation-maximization (EM) algorithm.

Image retrieval is based on the mapping $g : \mathcal{X} \rightarrow \{1, \ldots, D\}$ of (1), implemented by combining (2) and Bayes rule. Although any prior class distribution $P_Y(i)$ can be supported, we assume a uniform distribution. In the remainder of this work we refer to the nearest-neighbor operation of (1), at the visual level, as *query-by-visual-example* (QBVE).

3.2. Semantic-level retrieval system

A semantic-level retrieval system augments the database $\mathcal{D}$ with a vocabulary $\mathcal{L} = \{w_1, \ldots, w_L\}$ of semantic concepts or keywords w_i, and each image $\mathcal{I}_i$ with a pre-specified caption $\mathbf{c}_i$, making $\mathcal{D} = \{(\mathcal{I}_1, \mathbf{c}_1), \ldots, (\mathcal{I}_D, \mathbf{c}_D)\}$. Note that $\mathbf{c}_i$ is a binary L-dimensional vector such that $\mathbf{c}_{i,j} = 1$ if the i^{th} image was annotated with the j^{th} keyword in $\mathcal{L}$.

The database is said to be weakly labeled if the absence of a keyword from caption $\mathbf{c}_i$ does not necessarily mean that the associated concept is not present in $\mathcal{I}_i$. For example, an image containing "sky" may not be explicitly labeled with that keyword. This is usually the case in practical scenarios, since each image is likely to be annotated with a small caption that only identifies the semantics deemed as most relevant to the labeler. We assume weak labeling throughout this work.

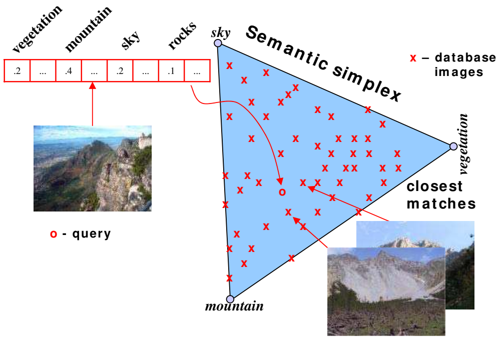


Figure 2. Under QBSE the user provides a query image, posterior probabilities (given the image) are computed for all concepts, and the image represented by the concept probability distribution.

Concepts are determined by the random variable W, which takes values in $\{1, \ldots, L\}$, so that $W = i$ if and only if $\mathbf{x}$ is a sample from the concept w_i. Each concept induces a probability density $\{P_{\mathbf{X}|W}(\mathbf{x}|i)\}_{i=1}^L$ on $\mathcal{X}$. At the *semantic level* images are assumed to be independently sampled from concept distributions,

$$P_{\mathbf{X}|W}(\mathcal{I}|w) = \prod_j P_{\mathbf{X}|\mathbf{W}}(\mathbf{x}_j|w). \tag{3}$$

For each concept w, the semantic class density $P_{\mathbf{X}|W}(\mathbf{x}|w)$ is learned from the set $\mathcal{D}_w$ of all training images labeled with the w^{th} label in $\mathcal{L}$. In the implementation of [3], this is based on a *hierarchical* procedure [29], which estimates semantic class densities directly from the image densities used for QBVE, in (2).

To support retrieval from the database using natural language queries, the unlabeled images are first annotated with the concepts of high posterior probability.

$$w^* = \arg\max_w P_{W|\mathbf{X}}(w|\mathcal{I}). \tag{4}$$

Given a query concept w_q, the optimal retrieval decision (in the MPE sense) is then to select the image for which w_q has the largest posterior annotation probability.

4. Query by Semantic Example

A QBSE retrieval system operates at the semantic level, representing images by vectors of concept counts $\mathcal{I} = (c_1, \ldots, c_L)^T$. Each feature vector extracted from an image is assumed to be sampled from the probability distribution of a semantic class (concept), and c_i is the number of feature vectors drawn from the i^{th} concept. The count vector for the y^{th} image is drawn from a multinomial variable $\mathbf{T}$ of parameters $\boldsymbol{\pi}_y = (\pi_y^1, \ldots, \pi_y^L)^T$

$$P_{\mathbf{T}|Y}(\mathcal{I}|y; \boldsymbol{\pi}_y) \quad = \quad \frac{n!}{\prod_{k=1}^L c_k!} \prod_{j=1}^L (\pi_y^j)^{c_j}, \tag{5}$$

where π_y^i is the probability that an image feature vector is drawn from the i^{th} concept. Given an image $\mathcal{I} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$, the posterior concept probabilities

$$\pi_w = P_{W|\mathbf{X}}(w|\mathcal{I}) \tag{6}$$

are maximum a posteriori estimates of the parameters $\boldsymbol{\pi}_i$, and can be computed by combining (3) and Bayes rule, assuming a uniform prior concept distribution $P_W(w)$.

The random variable $\mathbf{T}$ is the result of a feature transformation from the space of visual features $\mathcal{X}$ to the L-dimensional probability simplex $\mathcal{S}_L$. This mapping establishes a one-to-one correspondence between images and points $\boldsymbol{\pi}_y \in \mathcal{S}_L$. We refer to the probability vector $\boldsymbol{\pi}_y$ as the *semantic multinomial* (SMN) that characterizes the y^{th} image. For example, in Fig. 1 this is a 371-dimensional vector.

The QBSE system, then performs a nearest neighbor operation on the simplex $\mathcal{S}_L$, according to a similarity mapping $f : \mathcal{S}_L \rightarrow \{1, \dots, D\}$ such that

$$f(\boldsymbol{\pi}) = \arg\min_y d(\boldsymbol{\pi}, \boldsymbol{\pi}_y) \tag{7}$$

where $\boldsymbol{\pi}$ is the query SMN, $\boldsymbol{\pi}_y$ the SMN of the y^{th} database image, and $d(\cdot, \cdot)$ an appropriate dissimilarity function. In this work, the dissimilarity between two SMNs, $\boldsymbol{\pi}$ and $\boldsymbol{\pi}'$ is measured using the Kullback-Leibler divergence, i.e.

$$d(\boldsymbol{\pi}, \boldsymbol{\pi}') = \sum_{i=1}^{L} \pi_i \log \frac{\pi_i}{\pi_i'}. \tag{8}$$

This is the asymptotic limit of (1), when Y is uniformly distributed. Similarity matching in semantic space is also illustrated in Fig. 2, which depicts a query and the two *closest* database matches.

The mapping of visual features to the L-dimensional probability simplex $\mathcal{S}_L$ can be seen as an abstract projection of the image onto a *semantic space* where each concept probability $\pi_w, w = 1, \dots, L$ can be thought of as a *semantic feature*, as illustrated by Fig. 2. Features (semantic concepts) that are not the part of semantic vocabulary define directions that are orthogonal to this semantic space. Their projection onto the learned semantic simplex enables QBSE to generalize beyond the known semantic concepts, and hence achieves better performance even *outside the semantic space*. This is exemplified by Fig. 8 where images of ‘construction’ (a concept absent from the semantic vocabulary) are successfully retrieved from the database. In this case, the projection of ‘construction’ images on the learned semantic simplex assigns higher probabilities to (known) concepts such as ‘people’, ‘buildings’, ‘streets’, ‘tables’ etc. Since these are an effective alternative characterization for the ‘construction’ concept, the retrieval operation succeeds.

Table 1. Retrieval and Query Database

Database	*Corel50*	*Corel15*	*Flickr18*
Semantic Space	Inside	Outside	Outside
Source	Corel CDs	Corel CDs	`flickr`
# Retrieval Images	4500	1200	1440
# Query Images	500	300	360
# Classes	50	15	18

5. Experimental evaluation

In this section, we report on the experimental study of the QBSE system. First, we examine the dependence of retrieval performance on both the low-level visual and the high-level semantic spaces. This is done by considering two cases: 1) where the query and database images contain semantic concepts known to the semantic labeling system, and 2) where this is not true. We refer to the former as *retrieval inside the semantic space* and to the latter as *retrieval outside the semantic space*. Next, we also present a study of the structure of the semantic space, showing that it captures contextual relationships between semantic concepts. This intrinsic structure is also shown to be essential for the success of the overall retrieval operation. In all cases, performance is measured with precision-recall (PR) curves and mean average precision (MAP) [7].

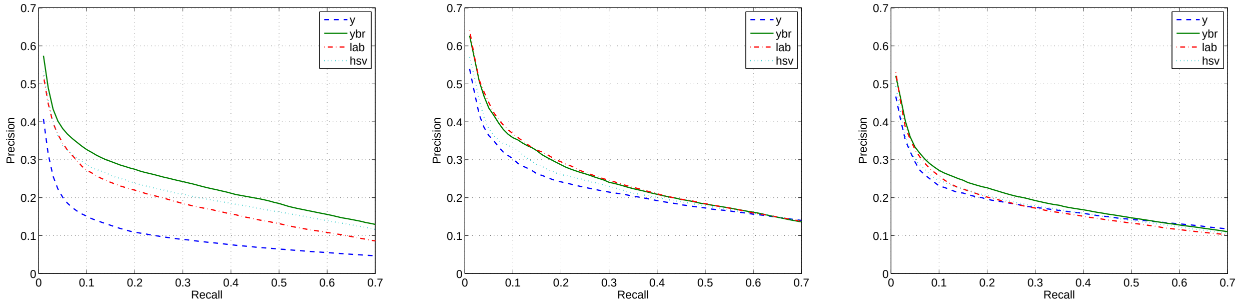
5.1. Databases

The study of a QBSE system requires three databases: a *training database*, used by the semantic labeling system to learn concept probabilities, a *retrieval database*, from which images are to be retrieved, and a database of *query images*, which plays the role of test set. All experiments are conducted on datasets used in [23]. Table 1 summarizes the composition of the databases used. The retrieval database of *Corel50* is used as the *training database* to learn the semantic space.

Note that the use of multiple-image queries has been shown to outperform single-image queries in [23]. In this work, we restrict our attention to single-image queries, as the aim is not so much to maximize performance, but to obtain a deeper understanding of the QBSE system.

5.2. Low-level visual space

In all experiments, images are normalized to a maximum of 180 pixels on the longest side, keeping the aspect ratio constant. To represent images at the low-level, they are converted to various colorspaces, including various 3-channel colorspaces (“YBR”, “HSV”, and “Lab”) and one single-channel colorspace (“Y”, luminance only). Image observations are derived from 8×8 patches obtained with a sliding window, moved in a raster-scan fashion. A feature transformation is applied by computing the 8×8 discrete cosine transform (DCT) coefficients per patch and color chan-



(a) Corel50 (b) Corel15 (c) Flickr18

Figure 3. PR curves achieved with different color spaces on the three retrieval databases. (a) Inside the semantic space (*Corel50*). (b,c) Outside the semantic space (*Corel15*, *Flickr18*).

nel. These DCT coefficients are then ordered by decreasing variance, producing a 64 dimensional feature vector. For 3-channel colorspaces, features from different channels are interleaved, e.g., the “YBR” channels are interleaved according to a “YBRYBR...” pattern. The parameters of the semantic class mixture hierarchies are learned in a subspace of these DCT coefficients. We evaluate subspaces of various dimensionalities, ranging from 3 to 64 dimensions per channel. Typically, low-dimensional subspaces capture low-frequency information, producing a coarse image representation. As the dimensionality increases, so does the accuracy of the low-level visual representation. Overall, this choice of features enables a number of possibilities for color and texture representation: from perceptual to non-perceptual color spaces, to texture only, in each case controlling the amount of texture representation by varying the subspace dimensionality.

5.2.1 Colorspace

Retrieval experiments were conducted with four different colorspaces, viz. “YBR”, “LAB”, “HSV”, “Y”. Fig. 3 presents the PR curves obtained on different databases. Inside the semantic space (Fig. 3(a)), the performance of 3-channel colorspaces supersedes that of luminance only colorspace significantly. This indicates, that the color correlations are a significant source of information for this database. Among the different 3-channel colorspaces, “YBR” performs better than the perceptually uniform “LAB” and the cylindrical co-ordinate based “HSV” spaces. The MAP scores for the three colorspaces are 0.197, 0.152 and 0.174 respectively, the chance performance stands at 0.0200.

Outside the semantic space, the experiments reveal a different behavior (Fig. 3(b)(c)). Interestingly, the performance of the “Y” colorspace is only marginally lower than those of the 3-channel colorspaces. That is, using just the “texture” (“Y” colorspace) information, the retrieval system performs as well as when color is also available (“texture+color” representation with any of the “YBR”, “LAB”, and “HSV” colorspaces). This suggests that only the learned “texture” correlations are informative for generalization with respect to previously unseen concepts. If true, this assertion would imply that color features capture information that, while characteristic of the images in database, is not characteristic of the underlying concepts. This could be due to the existence of certain global regularities within each class in the database (e.g. most images taken at certain times of the day or year) that create commonalities of color distribution which, although artificial, are not easily detected by visual inspection. It is an interesting assertion, given the long history of research in color-based image retrieval. While further experiments will be required to reach definitive conclusions, these results have lead us to adopt the “YBR” colorspace in the remaining experiments. Fig. 4 shows a query and the corresponding retrieved images for the ’YBR’ and ’Y’ colorspace.

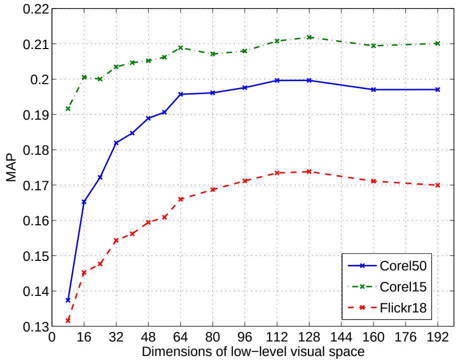


Figure 5. MAP scores of QBSE for different dimensions of the low-level visual space across all the databases.

5.2.2 Dimensionality of visual space

Since the visual representation is based on a subspace of the 192-dimensional space of DCT coefficients, it is possible to control the accuracy of visual representation by simply varying the dimension of this subspace. As the number of

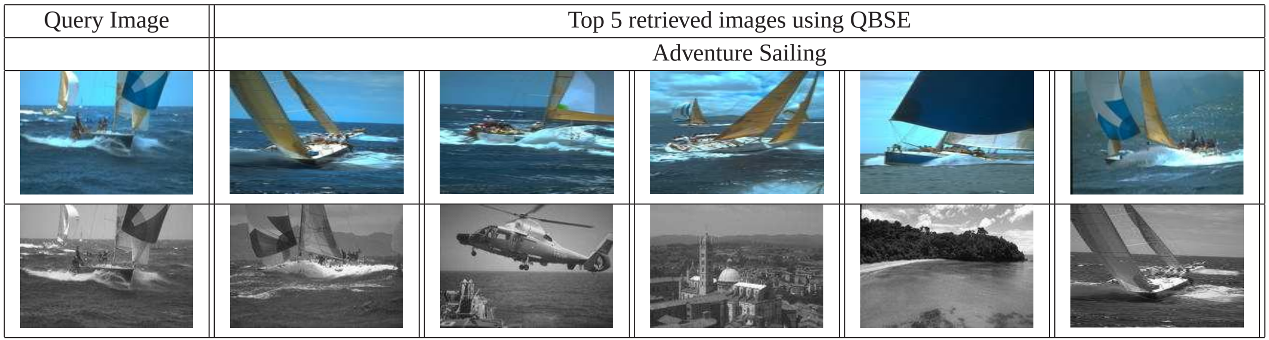


Figure 4. An example of a query and corresponding retrieved images from *Corel15* dataset. The first and the second row shows results using "YBR" and "Y" colorspace respectively. This figure is best viewed in color.

visual features decreases, the performance of QBSE, tends to degrade. Fig. 5 shows the MAP score as the subspace dimension varies from 8 to 192 for the interleaved "YBR" colorspace. The performance across the three databases are qualitatively similar, it increases rapidly from 8 to 64 dimensions and then remains fairly stable over a large range of dimensions. This suggests that 1) accuracy of low-level visual space is an important parameter for retrieval, and 2) the system is robust to the noise, introduced by the high frequency components of the DCT features. We use the first 64 dimensions of the interleaved "YBR" colorspace for rest of the experiments.

5.3. High-level Semantic space

We next study the dependence of QBSE performance on the number of informative semantic dimensions. Assuming that the dimensions of the learned semantic space are not equally useful for retrieval, it should be possible to achieve improved performance with feature selection. It should, nevertheless, be noted that standard feature extraction techniques, such as principal component analysis or latent semantic indexing, do not preserve the semantic meaning of the dimensions of the space. To avoid this problem, we investigated the benefits of feature selection by simply 1) ordering the semantic features by decreasing variance of their posterior probabilities (over the retrieval database) and 2) selecting the top k, for values of k ranging from 4 to 371.

Fig 6 shows the MAP score obtained on the three databases, as a function of k. In each figure, the right vertical-axis shows the percent of the variance (over the retrieval database) explained by the top k features, as a function of k. It can be observed that retrieval performance improves proportionally to the increase in the number of informative semantic dimensions. This is explained by the fact that more features enable a greater diversity of contextual correlations between concepts, and the similarity judgments are more robust. However, there is a saturation effect, i.e. not all 371 semantic concepts are equally informative. In particular, the MAP score saturated faster on *Flickr18*, *Corel15* than on *Corel50*, as almost all the variability is explained by less than 100 concepts for the former while more than 200 are needed for the latter. However, unlike most learning problems, the inclusion of uninformative features does not seem to degrade retrieval performance. We have, therefore, used all semantic features in the remaining experiments.

Table 2. Semantic feature pairs with highest mutual information.

Feature Pair	MI	Feature Pair	MI
'polar-bear'	0.1949	'sun-sunset'	0.1684
'beach-sand'	0.1579	'stone-ruins'	0.1566
'plane-jet'	0.1297	'leaf-flowers'	0.1075
'sun-sea'	0.0976	'light-restaurant'	0.0881
'sky-tree'	0.0852	'restaurant-tables'	0.0832
'sunset-sea'	0.0734	'statue-pillar'	0.0700
'sky-beach'	0.0690	'petals-leaf'	0.0687
'sky-mountain'	0.0583	'tree-mountain'	0.0568

5.4. Structure of the semantic space

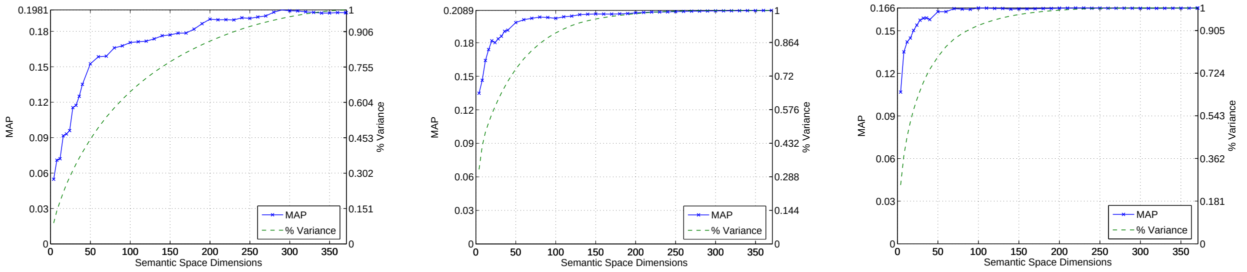
In this section we demonstrate that 1) the labeling process does seem to produce a space with semantic structure, and 2) this semantic structure is a necessary condition for the success of QBSE.

5.4.1 Relationship between semantic features

To unveil some of the structure of the semantic space, we analyzed the relationship between pairs of semantic features, by measuring their mutual information (MI) [4]

$$I(w_1; w_2) = \sum_{w_2 \in \mathcal{L}} \sum_{w_1 \in \mathcal{L}} p(w_1, w_2) \log \frac{p(w_1, w_2)}{p(w_1)\, p(w_2)}, \quad (9)$$

where $p(w_i)$ is estimated from the posterior probability of the semantic feature w_i in a given set of SMNs. Since MI is a measure of the statistical dependence between variables, it should be strong for pairs of concepts that are either synonyms or frequently appear together in natural imagery. Table 2 presents the most dependent concept pairs for the SMNs in the retrieval dataset of *Corel15*. Note that, even



(a) Corel50 (b) Corel15 (c) Flickr18

Figure 6. MAP scores for all the databases, as it varies with the dimensions of the semantic space. (a) Inside the semantic space (*Corel50*). (b,c) Outside the semantic space (*Corel15*, *Flickr18*). Also shown are the % variance of the semantic dimensions, as it varies across the respective retrieval database (on the right Y-axis).

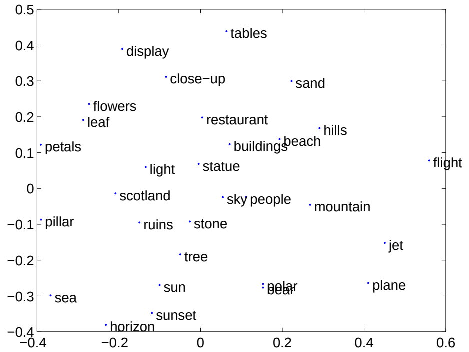
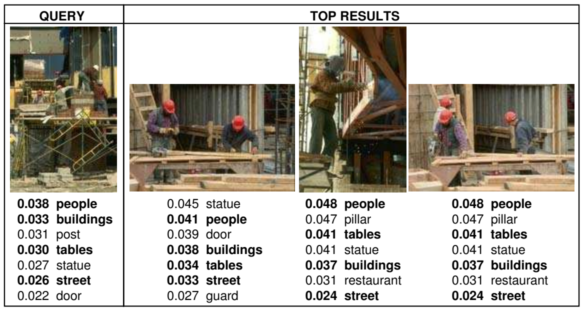


Figure 7. A visualization of the semantic correlations in *Corel15* dataset. The mutual information of the top 30 concepts (sorted according to their variance), is used to learn an embedding in a two-dimensional space, by non-metric multidimensional scaling.



QUERY	TOP RESULTS		
0.038 people	0.045 statue	**0.048 people**	**0.048 people**
0.033 buildings	**0.041 people**	0.047 pillar	0.047 pillar
0.031 post	0.039 door	**0.041 tables**	**0.041 tables**
0.030 tables	**0.038 buildings**	0.041 statue	0.041 statue
0.027 statue	**0.034 tables**	**0.037 buildings**	**0.037 buildings**
0.026 street	**0.033 street**	0.031 restaurant	0.031 restaurant
0.022 door	0.027 guard	**0.024 street**	**0.024 street**

Figure 8. Query from class 'commercial construction' and top QBSE matches. Shown below each image are the semantic features of highest posterior probability.

though none of the images in this set was used to train the semantic space, all pairs consist of words which are, indeed, semantically correlated. Fig.7 presents a visualization of the semantic correlations amongst the top 30 concepts (selected according to highest variance) in *Corel15*. To obtain this visualization, the mutual informations between concepts were used to learn a two-dimensional embedding of the semantic space, with non-metric multidimensional scaling [13]. These correlations show that the semantic space encodes contextual relationships.

To further substantiate this claim, Fig 8 shows a query image from the class 'Commercial construction' (*Corel15*). Although the 'construction' concept is absent from the semantic vocabulary, the top retrieved images are all in this class. This illustrates how the QBSE system is effectively able to rely on contextual correlations to retrieve semantically similar images. Analyzing the SMN's of the query and retrieved images, it is clear that the semantic features of largest probability (shown below each image) include various words that are contextually related to the concept of 'construction'. This shows that outside the semantic space, retrieval success is purely due to the effectiveness of such contextual relationships.

5.4.2 Meaningless semantic space

The fact that QBSE significantly outperforms QBVE both inside and outside the semantic space is strong evidence for the benefits of image retrieval on semantic spaces. To study the benefits of the contextual structure of the semantic space, QBSE was applied to a *meaningless semantic space* - a semantic space without any contextual structure. This was achieved by replicating the QBSE experiments with random image groupings. That is, instead of a semantic space composed of concepts like 'sky' (learned from images containing sky), we created a semantic space of nameless concepts learned from random collections of images. Fig. 9 compares (on *Corel50*) the PR obtained with QBSE on this "meaningless semantic space", with the previous results of QBVE and QBSE. Although, as before, the classification is performed on a *semantic space* (albeit meaningless), the absence of true semantic structure leads to very poor QBSE performance, even clearly inferior to that of QBVE. This suggests that the gains previously observed for QBSE are intrinsic to the semantic nature of the image representation, and strengthens the claim that the contextual correlations of the underlying semantic space are the reason for its advantages over QBVE.

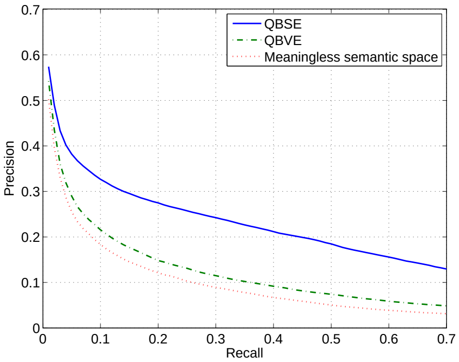


Figure 9. Comparison of precision-recall curve for the retrieval results using meaningless semantic space to that of QBSE and QBVE inside the semantic space (*Corel50*).

6. Conclusion

We have presented an extensive study of the QBSE image retrieval framework. This study supports various conclusions. First, experiments on the low-level visual space, reveal that 1) inside the semantic space colorspaces play an important role in retrieval performance, with the “YBR” color space achieving the best results, but 2) outside the semantic space there are only small differences across colorspaces. Second, experiments on the high-level semantic space, reveal that 1) semantic features are not all equally informative for retrieval, and 2) the number of informative features grows proportionally to the variance of the semantic multinomials. Third, a study of the intrinsic structure of the semantic space revealed the presence of contextual relationships between concepts, that seems to substantially improve the robustness of similarity judgments. Finally, it was shown that, in the absence of meaningful semantic structure, QBSE performs worse than QBVE.

It should be noted that our current implementation does not incorporate spatial scene information, current evidence [14] favoring integration of weak spatial information. Furthermore, although our visual representations is based on DCT features, the current success of scale invariant features such as SIFT [15] warrants a preference for them. At the semantic level, instead of using variance based feature selection, more sophisticated feature extraction techniques which conserve the semantic meaning of the space, such as probabilistic latent semantic indexing [10], can also be used. We intend to investigate these question in future work.

References

[1] K. Barnard and D. Forsyth. Learning the semantics of words and pictures. In *ICCV*, volume 2, pages 408–415, Vancouver, 2001.

[2] D. Blei and M. Jordan. Modeling annotated data. In *Proc. ACM SIGIR*, 2003.

[3] G. Carneiro, A. Chan, P. Moreno, and N. Vasconcelos. Supervised learning of semantic classes for image annotation and retrieval. *IEEE PAMI*, 29(3):394–410, March, 2007.

[4] T. Cover and J. Thomas. *Elements of Information Theory*. John Wiley, 1991.

[5] R. Datta, D. Joshi, J. Li, and J. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys*, 39:65, 2007.

[6] P. Duygulu, K. Barnard, N. Freitas, and D. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *ECCV*, Copenhagen, Denmark, 2002.

[7] S. Feng, R. Manmatha, and V. Lavrenko. Multiple bernoulli relevance models for image and video annotation. In *IEEE CVPR*, Washington DC, 2004.

[8] S. G., W. A., and Y. C. A vector space model for automatic indexing. *Communications of the ACM*, 18:613–620, 1975.

[9] X. He, O. King, W. Ma, M. Li, and H. Zhang. Learning a semantic space from user’s relevance feedback for image retrieval. 13(1):39–48, January 2003.

[10] T. Hofmann. Probabilistic latent semantic indexing. *ACM SIGIR*, pages 50–57, 1999.

[11] A. Jain and A. Vailaya. Image retrieval using color and shape. *Pattern Recognition Journal*, 29, August 1996.

[12] Y. Kiyoki, T. Kitagawa, and T. Hayama. A metadatabase system for semantic image search by a mathematical model of meaning. *SIGMOD Rec.*, 23(4):34–41, 1994.

[13] J. Kruskal. Nonmetric multidimensional scaling: A numerical method. *Psychometrika*, 29(2):115–129, 1964.

[14] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. *Proc. IEEE CVPR*, 2005.

[15] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[16] J. Lu, S. ping Ma, and M. Zhang. Automatic image annotation based-on model space. In *IEEE NLP-KE*, pages 455– 460, 2005.

[17] Y. Lu, H. Zhang, L. Wenyin, and C. Hu. Joint semantics and feature based image retrieval using relevance feedback. *IEEE Transactions on Multimedia*, 5(3):339–347, 2003.

[18] J. Magalhães, S. Overell, and S. Rüger. A semantic vector space for query by image example. *ACM SIGIR*, 2007.

[19] J. Magalhães and S. Rüger. Information-theoretic semantic multimedia indexing. *Proceedings of the 6th ACM CIVR*, pages 619–626, 2007.

[20] A. Natsev, M. Naphade, and J. Smith. Semantic representation: search and mining of multimedia content. *Proceedings of the 2004 ACM SIGKDD*, pages 641–646, 2004.

[21] W. Niblack and et al. The qbic project: Querying images by content using color, texture, and shape. In *Storage and Retrieval for Image and Video Databases*, pages 173–181, SPIE, Feb. 1993.

[22] A. Pentland, R. Picard, and S. Sclaroff. Photobook: Content-based manipulation of image databases. *Int. Journal of Computer Vision*, Vol. 18(3):233–254, June 1996.

[23] N. Rasiwasia, P. Moreno, and N. Vasconcelos. Bridging the gap: Query by semantic example. *Multimedia, IEEE Transactions on*, 9(5):923–938, 2007.

[24] N. Rasiwasia, N. Vasconcelos, and P. J. Moreno. Query by semantic example. In *CIVR*, pages 51–60, 2006.

[25] J. Smith and S. Chang. Visualseek: a fully automated content-based image query system. In *ACM Multimedia, Boston, Massachussetts*, pages 87–98, 1996.

[26] J. Smith, M. Naphade, and A. Natsev. Multimedia semantic indexing using model vectors. *ICME*, pages 445–448, 2003.

[27] J. R. Smith, C.-Y. Lin, M. R. Naphade, A. Natsev, and B. L. Tseng. Validity-weighted model vector-based retrieval of video. In *Proceedings of the SPIE.*, pages 271–279, 2003.

[28] N. Vasconcelos. Minimum probability of error image retrieval. *IEEE Trans. on Signal Processing*, 52(8), August 2004.

[29] N. Vasconcelos. Image indexing with mixture hierarchies. In *Proc. IEEE CVPR.*, Kawai, Hawaii, 2001.