Chapter 3

Image Retrieval: Query By Semantic Example

3.1 Introduction

Content-based image retrieval, the problem of searching for digital images in large image repositories according to their content, has been the subject of significant research in the recent past [133, 101, 107, 160]. Two main retrieval paradigms have evolved over the years: one based on visual queries, here referred to as *query-by-visual-example* (QBVE), and the other based on text, here denoted as *semantic retrieval* (SR). Early retrieval architectures were almost exclusively based on QBVE [61, 134, 90, 101, 107]. Under this paradigm, each image is decomposed into a number of low-level visual features (e.g. a color histogram) and image retrieval is formulated as the search for the best database match to the feature vector extracted from a query image. It was, however, quickly realized that strict visual similarity is, in most cases, weakly correlated with the measures of similarity adopted by humans for image comparison.

This motivated the more ambitious goal of designing retrieval systems with support for semantic queries [109]. The basic idea is to annotate images with semantic keywords, enabling users to specify their queries through a natural language description of the visual concepts of interest. Because manual image labeling is a labor intensive process, SR research turned to the problem of the automatic extraction of semantic descriptors from images, so as to build models of visual appearance of the semantic concepts of interest. This is usually done by the application of machine learning algorithms. Early efforts targeted the extraction of specific semantics [142, 152, 53, 45] under the framework of binary classification. More recently there has been an effort to solve the problem in greater generality, through the design of techniques capable of learning relatively large semantic vocabularies from informally annotated training image collections. This can be done with resort to both unsupervised [5, 35, 12, 41, 72] and weakly supervised learning [70, 22].

In spite of these advances, the fundamental question of whether there is an intrinsic value to building models at a semantic level, remains poorly understood. On one hand, SR has the advantage of evaluating image similarity at a higher level of abstraction, and therefore better generalization[1] than what is possible with QBVE. On the other hand, the performance of SR systems tends to degrade for semantic classes that they were not trained to recognize. Since it is still difficult to learn appearance models for massive concept vocabularies, this could compromise the generalization gains due to abstraction. This problem is seldom considered in the literature, where most evaluations are performed with query concepts that are known to the retrieval system [5, 12, 35, 41, 72, 22].

In fact, it is not even straightforward to compare the two retrieval paradigms, because they assume different levels of query specification. While a semantic query is usually precise (e.g. 'the White House') a visual example (a picture of the 'White House') will depict various concepts that are irrelevant to the query (e.g. the street that surrounds the building, cars, people, etc.). It is, therefore, possible that better SR results could be due to a better interface (natural language) rather than an intrinsic advantage of representing images semantically. This may be of little importance when the goal is to build the next generation of (more accurate) retrieval systems. However, given the complexity of the problem, it is unlikely that significant further advances can be achieved without some understanding of the intrinsic value of semantic representations. If, for example, abstraction is indeed valuable, further research on appearance models that account for image taxonomies could lead to exponential gains in retrieval accuracy. Else, if the advantages are simply a reflection of more precise queries, such research is likely to be ineffective.

In this chapter, we introduce a novel image retrieval framework based on semantic image representation, which extends the query-by-example paradigm to the semantic domain. This consists of defining a semantic feature space, where each image is represented by the vector of posterior concept probabilities assigned to it by a semantic labeling system, and performing query-by-example in this space. We refer to the combination of the two paradigms as query-by-semantic-example (QBSE), and present an extensive comparison of its performance with that of QBVE. It is shown that QBSE has significantly better performance for both

[1] Here, and throughout this work, we refer to the definition of 'generalization' common in machine learning and content-based retrieval: the ability of the retrieval system to achieve low error rates outside of the set of images on which it was trained.

concepts known and unknown to the retrieval system, i.e., it can generalize beyond the vocabulary used for training. It is also shown that, since both QBSE and QBVE share a common framework i.e. that of minimum probability of error retrieval [156], the performance gain of QBSE over QBVE is intrinsic to the semantic nature of image representation.

The chapter is organized as follows. Section 3.2 briefly reviews previous retrieval work related to QBSE. Section 3.3 discusses the limitations of the QBVE and SR paradigms, motivating the adoption of QBSE. Section 3.4 proposes an implementation of QBSE, compatible with the MPE formulation. It is then argued, in Section 3.5, that the generalization ability of QBSE can significantly benefit from the combination of multiple queries, and various strategies are proposed to accomplish this goal. A thorough experimental evaluation of the performance of QBSE is presented in Section 3.6, where the intrinsic gains of semantic image representations (over strict visual matching) are quantified.

3.2 Related Work

Although the task of building semantic image representations for image retrieval, has been on recent interest in the community, few proposals have so far been presented on how to best exploit the semantic space for the design of retrieval systems. A somewhat popular technique to construct content-based semantic spaces, is to resort to active learning based on user's relevance feedback [161, 87, 56]. The idea is to pool the images relevant to a query, after several rounds of relevance feedback, to build a model for the semantic concept of interest. Assuming that 1) these images do belong to a common semantic class, and 2) the results of various relevance feedback sessions can be aggregated, this is a feasible way to incrementally build a semantic space. An example is given in [75], where the authors propose a retrieval system based on image embeddings. Using relevance feedback, the system gradually clusters images and learns a non-linear embedding which maps these clusters into a hidden space of semantic attributes. Cox *et al.* [26] also focus on the task of learning a predictive model for user selections, by learning a mapping between 1) the image selection patterns made by users instructed to consider visual similarity and 2) those of users instructed to consider semantic similarity.

These works have focused more on the issue of learning the semantic space than that of its application to retrieval. In fact, it is not always clear how the learned semantic information could be combined with the visual search at the core of the retrieval operation. Furthermore, the use of relevance feedback to train a semantic retrieval system has various limitations. First, it can be quite time consuming, since a sizable number of examples is usually required to learn each semantic model. Second, the assumption that all queries performed in a relevance feedback session are relative to the same semantic concept is usually not realistic, even when users are instructed to do so. For example, a user searching for pictures of 'cafes in Paris' is likely to oscillate between searching for pictures of 'cafes' and pictures of 'Paris'.

The closest works in the literature, to the QBSE paradigm adopted here, are those of [137, 135, 86], where retrieval is carried out based on computing L^2 similarity between "model-vectors", a representation similar to that of semantic image representation. While laying the foundations for QBSE, [137, 135] did not investigate any of the fundamental questions that we now consider. First, because there was no attempt to perform retrieval on databases not used for training, it did not address the problem of generalization to concepts unknown to the retrieval system. As we will see, this is one of the fundamental reasons to adopt QBSE instead of the standard SR query paradigm. Second, although showing that QBSE outperformed a QBVE system, this work did not rely on the same image representation for the two query paradigms. While QBVE was based on either color or edge histogram matching, QBSE relied on a feature space composed of a multitude of visual features, including color and edge histograms, wavelet-based texture features, color correlograms and measures of texture co-occurrence. Because the representations are different, it is impossible to conclude that the improved performance of the QBSE system derives from an *intrinsic* advantage of semantic representations. In what follows, we preempt this caveat by adopting the same image representation and retrieval framework for the design of all systems.

3.3 Query by Semantic Example

Both the QBVE and SR implementations of MPE retrieval have been extensively evaluated in [156] and [21, 22]. Although these evaluations have shown that the two implementations are among the best known techniques for visual and semantic retrieval, the comparison of the two retrieval paradigms is difficult. We next discuss this issue in greater detail, and motivate the adoption of an alternative retrieval paradigm, QBSE, that combines the best properties of the two approaches.

3.3.1 Query by Visual Example vs Semantic Retrieval

Both QBVE and SR have advantages and limitations. Because concepts are learned from collections of images, SR can *generalize* significantly better than QBVE. For example, by using a large training set of images labeled with the concept 'sky', containing both images of sky at daytime (when the sky is mostly blue) and sunsets (when the sky is mostly orange), a SR system can learn that 'sky' is sometimes blue and others orange. This is a simple consequence of the fact that a large set of 'sky' images populate, with high probability, the blue and orange regions of the feature space. It is, however, not easy to accomplish with QBVE, which only has access to two images (the query and that in the database) and can only perform direct matching of visual features. We refer to this type of abstraction, as *generalization inside the semantic space*, i.e., inside the space of concepts that the system has been trained to recognize.

While better generalization is a strong advantage for SR, there are some limitations associated with this paradigm. An obvious difficulty is that most images have multiple semantic interpretations. 3.1 presents an example, identifying various semantic concepts as sensible annotations for the image shown. Note that this list, of relatively salient concepts, is a small portion of the keywords that could be attached to the image. Other examples include colors (e.g. 'yellow' train), or objects that are not salient in an abstract sense but could become very relevant in some contexts (e.g. the 'paint' of the markings on the street, the 'letters' in

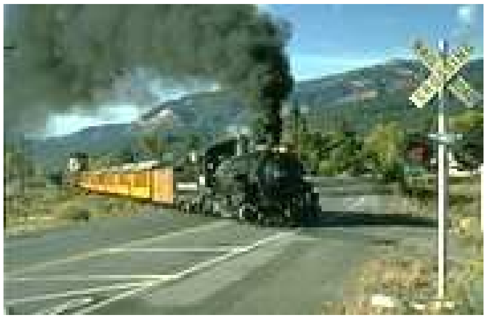


Figure 3.1: An image containing various concepts: ‘train’, ‘smoke’, ‘road’, ‘sky’, ‘railroad’, ‘sign’, ‘trees’, ‘mountain’, ‘shadows’, with variable degrees of presence.

the sign, etc.). In general, it is impossible to predict all annotations that may be relevant for a given image. This is likely to compromise the performance of a SR system. Furthermore, because queries are specified as text, a SR system is usually limited by the size of its vocabulary[2]. In summary, SR can generalize poorly *outside the semantic space*.

Since visual retrieval has no notion of semantics, it is not constrained by either vocabulary or semantic interpretations. When compared to SR, QBVE systems can generalize better outside the semantic space. In the example of 3.1, a QBVE would likely return the image shown as a match to a query depicting an industrial chimney engulfed in dark smoke (a more or less obvious query prototype for images of ‘pollution’) despite the fact that the retrieval system knows nothing about ‘smoke’, ‘pollution’, or ‘chimneys’. Obviously, there are numerous examples where QBVE correlates much worse with perceptual similarity than SR. We have already seen that when the latter is feasible, i.e. inside the semantic space, it has better generalization. Overall, it is sensible to expect that SR will perform better

[2] It is, of course, always possible to rely on text processing ideas based on thesauri and ontologies like WordNet [39] to mitigate this problem. For example, query expansion can be used to replace a query for ‘pollution’ by a query for ‘smoke’, if the latter is in the vocabulary and the former is not. While such techniques are undeniably useful for practical implementation of retrieval systems, they do not reflect an improved ability, by the retrieval system, to model the relationships between visual features and words. They are simply an attempt to fix these limitations a posteriori (i.e. at the language level) and are, therefore, beyond the scope of this work. In practice, it is not always easy to perform text-based query expansion when the vocabulary is small, as is the case for most SR systems, or when the queries report to specific instances (e.g. a person’s name).

inside the semantic space, while QBVE should fare better outside of it. In practice, however, it is not easy to compare the two retrieval paradigms. This is mostly due to the different forms of query specification. While a natural language query is usually precise (e.g. 'train' and 'smoke'), a query image like that of 3.1 always contains a number of concepts that are not necessarily relevant to the query (e.g. 'mountain', or even 'yellow' for the train color). Hence, the better performance of SR (inside the semantic space) could be simply due to higher query precision. A fair comparison would, therefore, require the optimization of the precision of visual queries (e.g. by allowing the QBVE system to rely on image regions as queries) but this is difficult to formalize.

Overall, both the engineering question of how to design better retrieval systems (with good generalization inside and outside of the semantic space) and the scientific question of whether there is a real benefit to semantic representations, are difficult to answer under the existing query paradigms. To address this problem we propose an alternative paradigm, which is denoted as *query by semantic example* (QBSE).

3.3.2 Query by Semantic Example

A QBSE system operates on a *semantic space* - the space of semantic features introduced in Chapter 2, according to a similarity mapping $f : \mathcal{S}_L \rightarrow \{1, \ldots, D\}$ such that

$$f(\boldsymbol{\pi}) = \arg\max_y s(\boldsymbol{\pi}, \boldsymbol{\pi}_y) \tag{3.1}$$

where $\mathcal{S}_L$ is the semantic space, $\boldsymbol{\pi}$ the query SMN and $\boldsymbol{\pi}_y$ the SMN that characterizes the y^{th} database image, and $s(\cdot, \cdot)$ an appropriate similarity function. As shown in 3.2 (top), the user provides a query image, for which a SMN $\boldsymbol{\pi}$ is computed, and compared to all the SMNs $\boldsymbol{\pi}_y$ previously stored for the images in the database. Note that this paradigm differs from SR, as in SR the user specifies a short natural language description which implies only a small number of concepts are assigned non-zero probability. This is illustrated in 3.2 (bottom) where queries in SR are restricted to the edges of the semantic space.

QBSE query paradigm has a number of interesting properties. As discussed

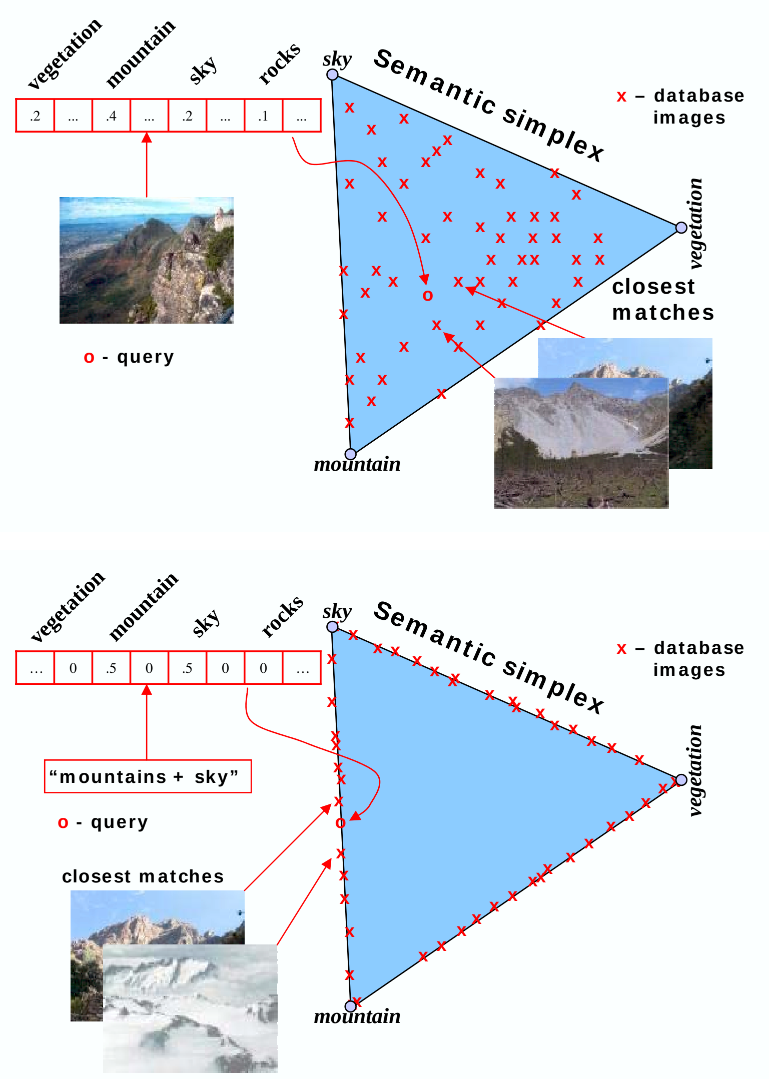


Figure 3.2: Semantic image retrieval. Top: Under QBSE the user provides a query image, probabilities are computed for all concepts, and the image represented by the concept probability distribution. Bottom: Under the traditional SR paradigm, the user specifies a short natural language description, and only a small number of concepts are assigned a non-zero posterior probability.

in Chapter 2, the semantic space $\mathcal{S}_L$ is defined by the concepts in the vocabulary known to the system. The semantic features, or concepts, outside the vocabulary simply define directions orthogonal to the learned semantic space. This implies that, by projecting these dimensions onto the simplex, the QBSE system can generalize beyond the known semantic concepts. In the example of 3.1, the mapping of the image onto the semantic simplex assigns high probability to (known) concepts such as ‘train’, ‘smoke’, ‘railroad’ etc. This makes the image a good match for other images containing large amounts of ‘smoke’, such as those depicting industrial chimneys or ‘pollution’ in general. The system can therefore establish a link between the image of 3.1 and ‘pollution’, despite the fact that it has no *explicit* knowledge of the ‘pollution’ concept[3]. Second, when compared to QBVE, QBSE complements all the advantages of query by example with the advantages of a semantic representation. Moreover, since in both cases queries are specified by the same examples, any differences in their performance can be directly attributed to the semantic vs. visual nature of the associated image representations[4]. This enables the objective comparison of QBVE and QBSE.

3.4 The Proposed Query by Semantic Example System

QBSE is a generic retrieval paradigm and, as such, can be implemented in many different ways. Any implementation must specify a method to estimate the SMN that describes each image, and a similarity function between SMNs. In the implementation presented herein, the SMN vectors $\boldsymbol{\pi}_i$ are learned with a semantic labeling system described in 2.2, which implements the mapping $\mathbf{\Pi}$, by computing an estimate of posterior concept probabilities given the observed feature vectors

$$\pi_w = P_{W|\mathbf{X}}(w|\mathcal{I}). \tag{3.2}$$

In the rest of this section, we describe the various similarity functions.

[3] Note that this is different from text-based query expansion, where the link between ‘smoke’ and ‘pollution’ must be *explicitly* defined. In QBSE, the relationship is instead inferred automatically, from the fact that both concepts have commonalities of visual appearance.

[4] This assumes, of course, that a common framework, such as MPE, is used to implement both the QBSE and QBVE systems.

3.4.1 Similarity Function

There are many known methods to measure the distance between two probability distributions, all of which can be used to measure the similarity of two SMNs. Furthermore, because the latter can also be interpreted as normalized vectors of counts, this set can be augmented with all measures of similarity between histograms. We have compared various similarity functions for the purpose of QBSE.

Kullback-Leibler (KL) Divergence

The KL divergence between two distributions $\boldsymbol{\pi}$ and $\boldsymbol{\pi}'$ is

$$s_{KL}(\boldsymbol{\pi}, \boldsymbol{\pi}') = KL(\boldsymbol{\pi}||\boldsymbol{\pi}') = \sum_{i=1}^{L} \pi_i \log \frac{\pi_i}{\pi'_i}. \tag{3.3}$$

It is non-negative, and equal to zero when $\boldsymbol{\pi} = \boldsymbol{\pi}'$. For retrieval, it also has an intuitive interpretation as the asymptotic limit of (2.1) when Y is uniformly distributed [158]. However, it is not symmetric, i.e. $KL(\boldsymbol{\pi}||\boldsymbol{\pi}') \neq KL(\boldsymbol{\pi}'||\boldsymbol{\pi})$. A symmetric version can be defined as

$$\begin{aligned} s_{symmKL}(\boldsymbol{\pi}, \boldsymbol{\pi}') &= KL(\boldsymbol{\pi}||\boldsymbol{\pi}') + KL(\boldsymbol{\pi}'||\boldsymbol{\pi}) && (3.4) \\ &= \sum_{i=1}^{L} \pi_i \log \frac{\pi_i}{\pi'_i} + \sum_{i=1}^{L} \pi'_i \log \frac{\pi'_i}{\pi_i}. && (3.5) \end{aligned}$$

Jensen-Shannon Divergence

The Jensen-Shannon divergence (JS) is a measure of whether two samples, as defined by their empirical distributions, are drawn from the same source distribution [25]. It is defined as

$$s_{JS}(\boldsymbol{\pi}, \boldsymbol{\pi}') = KL(\boldsymbol{\pi}||\hat{\boldsymbol{\pi}}) + KL(\boldsymbol{\pi}'||\hat{\boldsymbol{\pi}}) \tag{3.6}$$

where $\hat{\boldsymbol{\pi}} = \frac{1}{2}\boldsymbol{\pi} + \frac{1}{2}\boldsymbol{\pi}'$. This divergence can be interpreted as the average distance (in the KL sense) between each distribution and the average of all distributions.

Correlation

The correlation between two SMNs is defined as

$$s_{CO}(\boldsymbol{\pi}, \boldsymbol{\pi}') = \boldsymbol{\pi}^T \boldsymbol{\pi}' = \sum_{i}^{L} \pi_i \times \pi'_i. \tag{3.7}$$

Unlike the KL or JS divergence, which attain their minimum value (zero) for equal distributions, correlation is maximum in this case. The maximum value is, however, a function of the distributions under consideration. This limitation can be avoided by the adoption of the *normalized correlation*,

$$s_{NC}(\boldsymbol{\pi}, \boldsymbol{\pi}') = \frac{\boldsymbol{\pi}^T \boldsymbol{\pi}'}{||\boldsymbol{\pi}|| ||\boldsymbol{\pi}'||} = \frac{\sum_i^L \pi_i \times \pi'_i}{\sqrt{\sum \pi_j^2}\sqrt{\sum \pi_j'^2}}. \tag{3.8}$$

Other Similarity Measures

A popular set of image similarity metrics is that of L^p distances

$$s_{L^p}(\boldsymbol{\pi}, \boldsymbol{\pi}') = \left(\sum_{i=1}^{L} |\pi_i - \pi'_i|^p \right)^{\frac{1}{p}}. \tag{3.9}$$

These distances are particularly common in color-based retrieval, where they are used as metrics of similarity between color histograms. Another popular metric is the histogram intersection (HI) [141],

$$s_{HI}(\boldsymbol{\pi}, \boldsymbol{\pi}') = \sum_{i=1}^{L} min(\pi_i, \pi'_i), \tag{3.10}$$

the maximization of which is equivalent minimizing the L^1 norm.

3.5 Multiple Image Queries

A QBSE system can theoretically benefit from the specification of queries through multiple examples. We next give some reasons for this and discuss various alternatives for query combination.

3.5.1 The Benefits of Query Fusion

Semantic image labeling is, almost by definition, a noisy endeavor. This is a consequence of the fact that various interpretations are usually possible for a given arrangement of image intensities. An example is given in 1.1 where we show an image and the associated SMN. While most of the probability mass is assigned to concepts that are present in the image ('railroad', 'locomotive', 'train', 'street', or 'sky'), two of the concepts of largest probability do not seem related to it: 'bridge', and 'arch'. Close inspection of the image (see close-up presented in the figure), provides an explanation for these labels: when analyzed locally, the locomotive's roof actually resembles the arch of a bridge. This visual feature seems to be highly discriminant, since when used as a query in a QBVE system, most of the top matches are images with arch-like structures, not trains (see 3.6). While these types of errors are difficult to avoid, they are *accidental*. In particular, the arch-like structure of 1.1 is the result of viewing a particular type of train, at a particular viewing angle, and a particular distance. It is unlikely that similar structures will emerge consistently over a set of train images. There are obviously other sources of error, such as classification mistakes for which it is not possible to encounter a plausible explanation. But these are usually even less consistent, across a set of images, than those due to accidental visual resemblances. A pressing question is then whether it is possible to exploit the lack of consistency of these errors to obtain a better characterization of the query image set?

We approach this question from a *multiple instance* learning perspective [92, 2], formulating the problem as one of learning from *bags of examples*. In QBSE, each image is modeled as a bag of feature vectors, which are drawn from the different concepts according to the probabilities $\boldsymbol{\pi}_i$. When the query consists of multiple images, or bags, the negative examples that appear across those bags are inconsistent (e.g. the feature vectors associated with the arch-like structure which is prominent in 1.1 but does not appear consistently in all train images), and tend to be spread over the feature space (because they also depict background concepts, such as roads, trees, mountains, etc., which vary from image to image). On the other hand, feature vectors corresponding to positive examples are likely to be concentrated within a small region of the space. It follows that, although the distribution of positive examples may not be dominant in any individual bag, the consistent appearance in all bags makes it dominant over the entire query ensemble. This suggests that a better estimate of the query SMN should be possible by considering a set of multiple query images.

In addition to higher accuracy, a set of multiple queries is also likely to have better generalization, since a single image does not usually exhibit all possible visual manifestations of a given semantic class. For example, images depicting 'bikes on roads' and 'cars in garage' can be combined to retrieve images from the more general class of 'vehicles'. A combination of the two query image sets enables the retrieval system to have a more complete representation of the vehicle class, by simultaneously assigning higher weights to the concepts 'bike', 'cars', 'road', and 'garage'. This enables the retrieval of images of 'bikes in garage' and 'cars on roads', matches that would not be possible if the queries were used individually.

3.5.2 Query Combination

Under MPE retrieval, query combination is relatively straightforward to implement by QBVE systems. Given two query images $\mathcal{I}_q^1 = \{\mathbf{x}_1^1, \mathbf{x}_2^1, \ldots, \mathbf{x}_n^1\}$ and $\mathcal{I}_q^2 = \{\mathbf{x}_1^2, \mathbf{x}_2^2, \ldots, \mathbf{x}_n^2\}$, the probability of the composite query $\mathcal{I}_q^C = \{\mathbf{x}_1^1, \mathbf{x}_2^1, \ldots, \mathbf{x}_n^1, \mathbf{x}_1^2, \mathbf{x}_2^2, \ldots, \mathbf{x}_n^2\}$ given class $Y = y$ is

$$\begin{aligned} P_{\mathbf{X}|Y}(\mathcal{I}_q^C|y) &= \prod_{k=1}^{n} P_{\mathbf{X}|Y}(\mathbf{x}_k^1|y) \prod_{l=1}^{n} P_{\mathbf{X}|Y}(\mathbf{x}_l^2|y) \\ &= P_{\mathbf{X}|Y}(\mathcal{I}_q^1|y) P_{\mathbf{X}|Y}(\mathcal{I}_q^2|y). \end{aligned} \tag{3.11}$$

The MPE decision of (2.1) for the composite query is obtained by combining (3.11) with (2.4) and Bayes rule.

In the context of QBSE, there are at least three possibilities for query combination. The first is equivalent to (3.11), but based on the probability of the composite query $\mathcal{I}_q^C$ given semantic class $W = w$,

$$P_{\mathbf{X}|W}(\mathcal{I}_q^C|w) = \prod_{k=1}^{n} P_{\mathbf{X}|W}(\mathbf{x}_k^1|w) \prod_{l=1}^{n} P_{\mathbf{X}|W}(\mathbf{x}_l^2|w) \tag{3.12}$$
$$= P_{\mathbf{X}|W}(\mathcal{I}_q^1|w) P_{\mathbf{X}|W}(\mathcal{I}_q^2|w),$$

which is combined with (2.9) and Bayes rule to compute the posterior concept probabilities of (3.2). We refer to (3.12) as the 'LKLD combination' strategy for query combination. It is equivalent to taking a geometric mean of the probabilities of the individual images given the class.

A second possibility is to represent the query as a mixture of SMNs. This relies on a different generative model than that of (3.12): the i^{th} query is first selected with probability λ_i and a count vector is then sampled from the associated multinomial distribution. It can be formalized as

$$P_{\mathbf{T}}(\mathcal{I}_q^C; \boldsymbol{\pi}_q) = \frac{n!}{\prod_{k=1}^{L} c_k!} \prod_{j=1}^{L} (\lambda_1 \pi_1^j + \lambda_2 \pi_2^j)^{c_j}, \tag{3.13}$$

where $P_{\mathbf{T}}(\mathcal{I}_q^C; \boldsymbol{\pi}_q)$ is the multinomial distribution for the query combination, of parameter $\boldsymbol{\pi}_q = \lambda_1 \boldsymbol{\pi}_1 + \lambda_2 \boldsymbol{\pi}_2$. $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_2$ are the parameters of the individual multinomial distribution, and $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)^T$ the vector of query selection probabilities. If $\lambda_1 = \lambda_2$, the two SMNs are simply averaged. We adopt the uniform query selection prior, and refer to this strategy as 'SMN combination'. Geometrically, it sets the combined SMN to the centroid of the simplex that has the SMNs of the query images as vertices. This ranks highest the database SMN which is closest to this centroid.

The third possibility, henceforth referred to as 'KL combination', is to execute the multiple queries separately, and combine the resulting image rankings. For example, when similarity is measured with the KL divergence, the divergence between the combined image SMN, $\boldsymbol{\pi}_q$, and database SMNs $\boldsymbol{\pi}_y$ is,

$$s_{KL}(\boldsymbol{\pi}_q, \boldsymbol{\pi}_y) = \frac{1}{2} KL(\boldsymbol{\pi}_1 || \boldsymbol{\pi}_y) + \frac{1}{2} KL(\boldsymbol{\pi}_2 || \boldsymbol{\pi}_y). \tag{3.14}$$

It is worth noting that this combination strategy is closely related to that used in QBVE. Note that the use of (3.11) is equivalent to using the arithmetic average (mean) of log-probabilities which, in turn, is identical to combining image rankings, as in (3.14). For QBVE the two combination approaches are identical.

3.6 Experimental Evaluation

In this section we report on an extensive evaluation of QBSE. We start by describing the evaluation procedure and the various databases used. This is followed by some preliminary tuning of the parameters of the QBSE system and the analysis of a number of retrieval experiments, that can be broadly divided into two classes. Both compare the performance of QBSE and QBVE, but while the first is performed inside the semantic space, the second studies retrieval performance outside of the latter.

3.6.1 Evaluation Procedure

In all cases, performance is measured with *precision* and *recall*, a classical measure of information retrieval performance [125], which is also widely used by the image retrieval community [136], and one of the metrics adopted by the TRECVID evaluation benchmark. Given a query and the top 'N' database matches, also called as *scope*, if 'r' of the retrieved objects are relevant (where relevant means belonging to the class of the query), and the total number of relevant objects in the database is 'R', then precision is defined as 'r/N', i.e. the percentage of N which are relevant and recall as 'r/R', which is the percentage of all relevant images contained in the retrieved set. Precision-recall is commonly summarized by the mean average precision (MAP)[41]. This consists of averaging the precision at the ranks where recall changes, and taking the mean over a set of queries. Because some authors [123] consider the characterization of retrieval performance by curves of *precision-scope* more expressive for image retrieval, we also present results with this measure.

3.6.2 Databases

The evaluation of a QBSE system requires three different databases. The first is a *training database*, used by the semantic labeling system to learn concept probabilities. The second is a *retrieval database* from which images are to be retrieved. The third is a database of *query images*, which do not belong to either

Table 3.1: Retrieval and Query Database

Database	*Corel371*	*Corel15*	*Flickr18*
Semantic Space	Inside	Outside	Outside
Source	Corel CDs	Corel CDs	`flickr.com`
# Retrieval Images	4500	1200	1440
# Query Images	500	300	360
# Classes	50	15	18

the training or retrieval databases. In the first set of experiments, the training and retrieval databases are identical, and the query images are inside the semantic space. This is the usual evaluation scenario for semantic image retrieval [35, 72, 41]. In the second, designed to evaluate generalization, both query and retrieval databases are outside the semantic space.

Training Database

We relied on *Corel371* dataset as the training database for all experiments. A detailed description of*Corel371* dataset is provided in Appendix A.1.3. $4,500$ training images are used to learn the semantic space. Since overall there are 371 concepts, this leads to a 371-dimensional semantic space. With respect to image representation, all images were normalized to size 181×117 or 117×181 and converted from RGB to the YBR color space. Image observations were derived from 8×8 patches obtained with a sliding window, moved in a raster-scan fashion. A feature transformation was applied to this space by computing the 8×8 discrete cosine transform (DCT) of the three color components of each patch. The parameters of the semantic class mixture hierarchies were learned in the subspace of the resulting 192-dimension feature space composed of the first 21 DCT coefficients from each channel. In all experiments, the SMN associated with each image was computed with these semantic class-conditional distributions.

Retrieval and Query Database

To evaluate retrieval performance we carried out tests on three databases *Corel371, Corel15* and *Flickr18*, the details of which are provided in Appendix A.1.3 and Appendix A.1.4.

Inside the Semantic Space Retrieval performance inside the semantic space was evaluated by using *Corel371* as both retrieval and query database. More precisely, the 4500 training images served as the *retrieval database* and the remaining 500 as the *query database.* This experiment relied on clear ground truth regarding the relevance of the retrieved images, based on the theme of the CD to which the query belonged.

Outside the Semantic Space To test performance outside the semantic space, we relied on two additional databases viz *Corel15* and *Flickr18.* For both databases, 20% of randomly selected images served as *query images* and the remaining 80% as the *retrieval database.* 3.1 summarizes the composition of the databases used.

A QBVE system only requires a query and a retrieval database. In all experiments, these were made identical to the query and retrieval databases used by the QBSE system. Since the performance of QBVE does not depend on whether queries are inside or outside the semantic space, this establishes a benchmark for evaluating the generalization of QBSE.

3.6.3 Model Tuning

All parameters of our QBVE system have been previously optimized, as reported in [156]. Here, we concentrate on the QBSE system, reporting on the impact of 1) SMN regularization, and 2) choice of similarity function on the retrieval performance. The parameters resulting from this optimization were used in all subsequent experiments.

Table 3.2: Effect of SMN regularization on the MAP score of QBSE

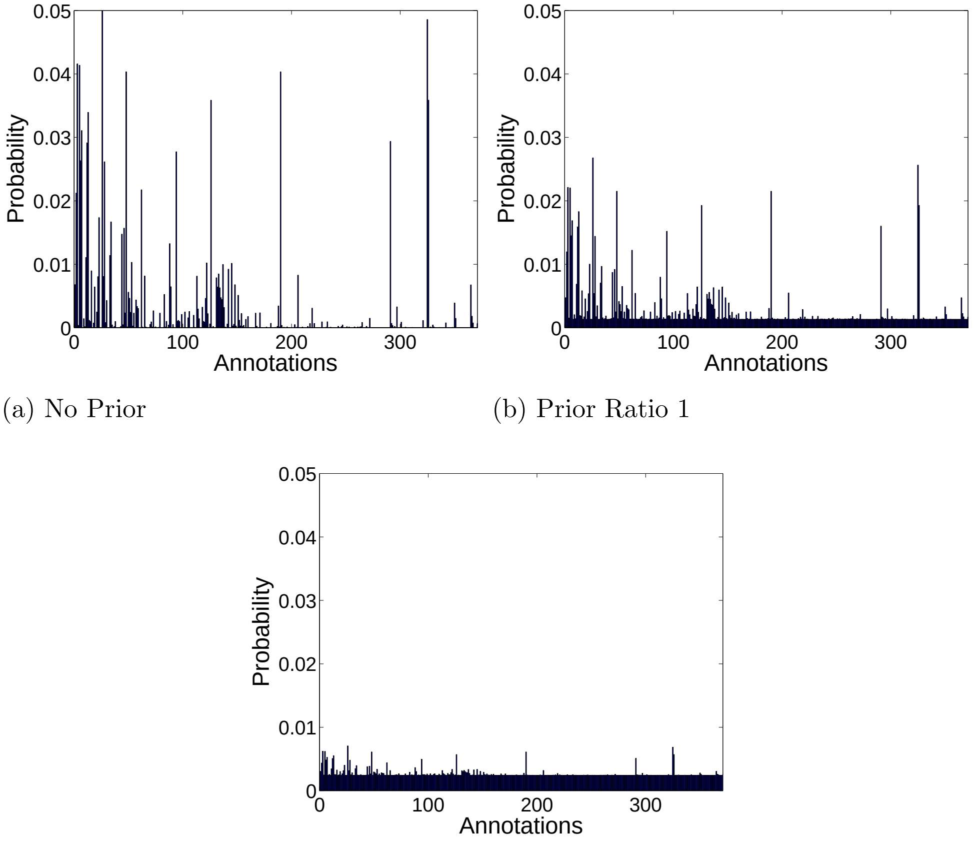
Ratio	**MAP score**		
	Corel371	*Corel15*	*Flickr18*
100	0.1544	0.1878	0.1447
10	0.1744	0.2030	0.1557
1	**0.1833**	0.2156	**0.1625**
0.1	0.1768	**0.2175**	0.1615
0.01	0.1709	0.2160	0.1594
0.001	0.1683	0.2150	0.1578
0.0001	0.1672	0.2144	0.1569
0.00001	0.1667	0.2141	0.1564

Effect of regularization on QBSE

3.2 presents the MAP obtained with values of $L\pi_0$ (2.18), ranging from 10^{-5} to 100. 3.3 presents the SMN of the *train* query of 3.6, for some of the values of $L\pi_0$. It can be seen that very large values of α force the SMN towards a uniform distribution, e.g. 3.3c, and almost all semantic information is lost. 3.3b shows the SMN regularized with the optimal value of $\pi_0 = 1/L$, where exceedingly low concept probabilities are lower-bounded by the value of 0.001. This regularization is instrumental in avoiding very noisy distance estimates during retrieval.

Effect of the Similarity Function on QBSE

3.3 presents a comparison of the seven similarity functions discussed in the text. It is clear that L^2 distance and histogram intersection do not perform well. All information theoretic measures, KL divergence, symmetric KL divergence and Jensen-Shanon divergence, have superior performance, with an average improvement of 15%. Among these the KL divergence performs the best. The closest competitors to KL divergence are the correlation and normalized correlation metrics. Although, they outperform KL divergence inside the semantic space (*Corel371*), their performance is inferior for databases outside the semantic space (*Flickr18,*



(a) No Prior

(b) Prior Ratio 1

(c) Prior Ratio 10

Figure 3.3: SMN of the *train* query of 3.6 as a function of the ratio $\frac{L(\alpha-1)}{n}$ adopted for its regularization.

Table 3.3: Effect of the similarity function on the MAP score of QBSE

Similarity Function	**MAP score**		
	Corel371	*Corel15*	*Flickr18*
KL divergence	0.1768	**0.2175**	**0.1615**
Symmetric KLD	0.1733	0.2164	0.1602
Jensen-Shanon	0.1740	0.2158	0.1611
Correlation	**0.2108**	0.1727	0.1392
Normalized Correlation	0.1938	0.2041	0.1595
L2 distance	0.1461	0.1830	0.1408
Histogram Intersection	0.1692	0.2119	0.1600

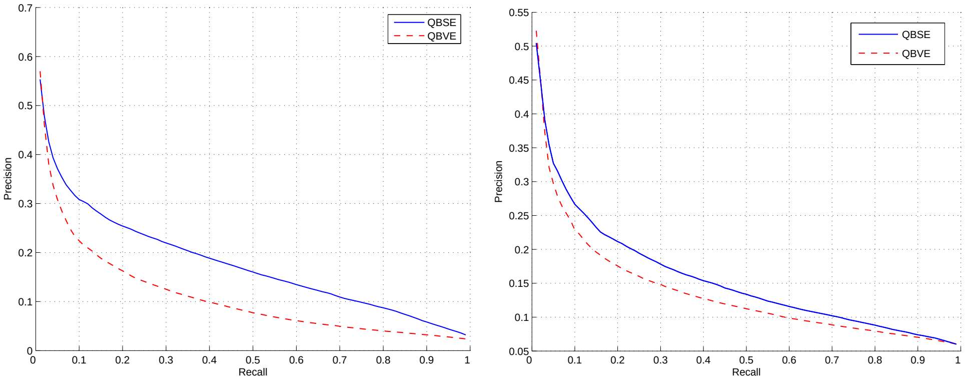


Figure 3.4: Average precision-recall of single-query QBSE and QBVE, Left: Inside the semantic space (*Corel371*), Right: Outside the semantic space (*Flickr18*).

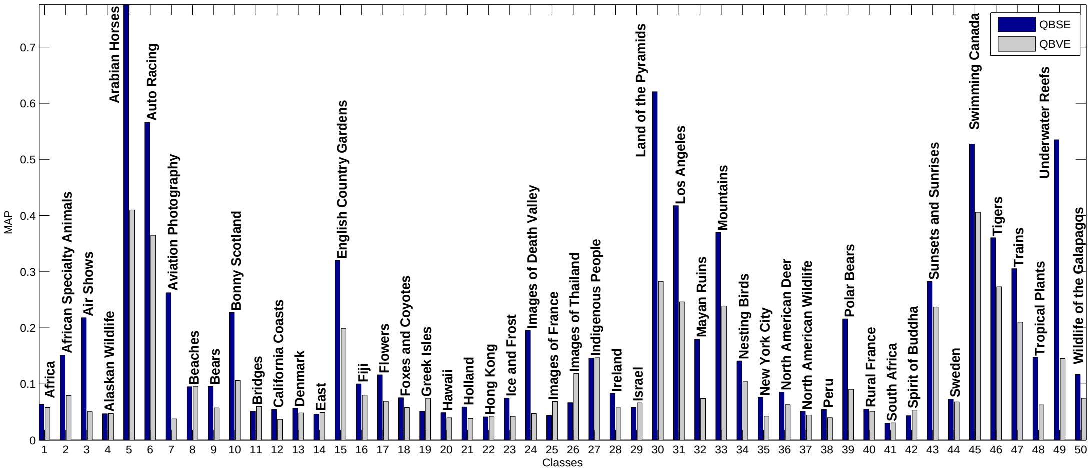


Figure 3.5: MAP scores of QBSE and QBVE across the 50 classes of *Corel371*.

Corel15).This indicates that the KL divergence is likely to have better generalization. While further experiments will be required to reach definitive conclusions, this has led us to adopt the KL divergence in the remaining experiments.

3.6.4 Performance Within the Semantic Space

3.4 (left) presents the precision-recall curves obtained on *Corel371* with QBVE and QBSE. It can be seen that the precision of QBSE is significantly higher than that of QBVE, at most levels of recall. The competitive performance of QBVE at low recall can be explained by the fact that there are always some database

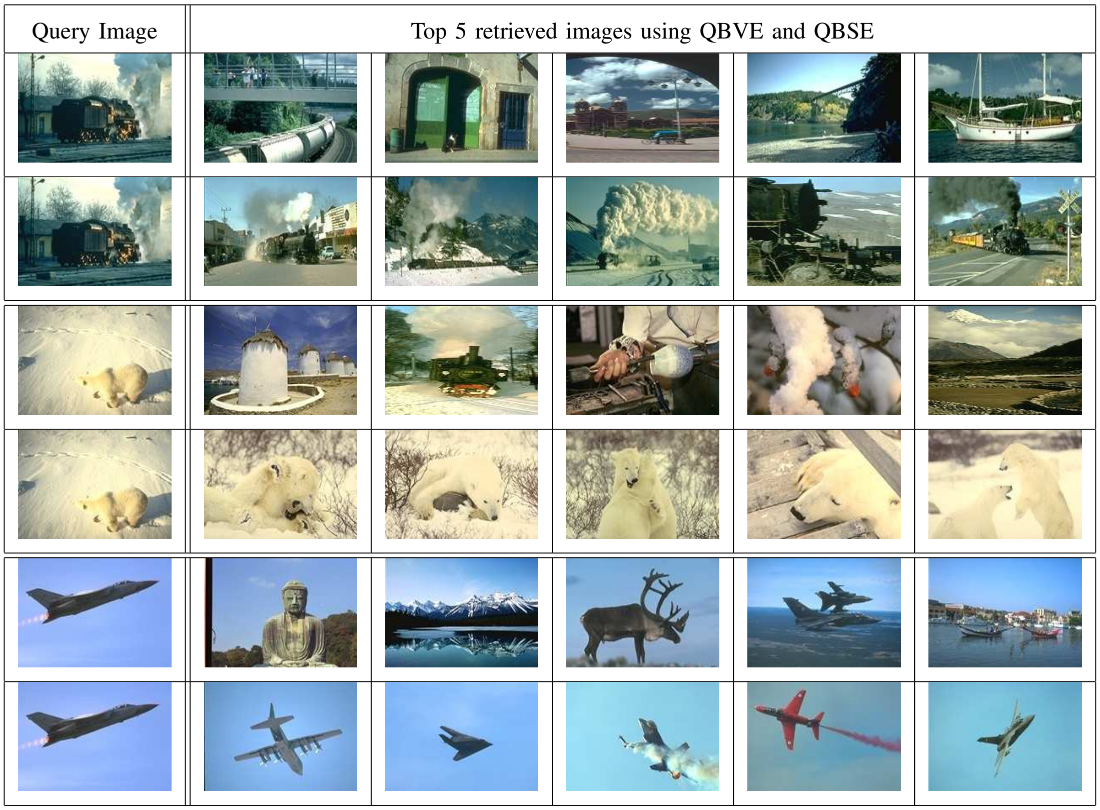


Figure 3.6: Some examples where QBSE performs better than QBVE. The second row of every query shows the images retrieved by QBSE.

images which are visually similar to the query. However, performance decreases much more dramatically than that of QBSE, as recall increases, confirming the better generalization ability of the latter. The MAP scores for QBSE and QBVE are 0.1665 and 0.1094 respectively and the chance MAP performance is 0.0200. 3.5 presents a comparison of the performance on individual classes, showing that QBSE outperforms QBVE in almost all cases.

The advantages of QBSE are also illustrated by 3.6, where we present the results of some queries, under both QBVE and QBSE. Note, for example, that for the query containing *white smoke* and a large area of *dark train*, QBVE tends to retrieve images with *whitish* components, mixed with *dark* components, that have little connection to the *train* theme. Furthermore, the arch-like structure highlighted in 1.1 seems to play a prominent role in visual similarity, since three of the five top matches contain arches. Due to its higher level of abstraction, QBSE is successfully able to generalize the main semantic concepts of *train, smoke* and *sky*, realizing that the white color is an irrelevant attribute to this query (as can be seen in the last column, where an image of *train with black smoke* is successfully retrieved).

3.6.5 Multiple Image Queries

Since the test set of *Corel371* contains 9 to 11 images from each class, it is possible to use anywhere from 1 to 9 images per query. When the number of combinations was prohibitively large (for example, there are close to $13,000$ combinations of 5 queries), we randomly sampled a suitable number of queries from the set. 3.7 (left) shows the MAP values for multiple image queries, as a function of query cardinality, under both QBVE and QBSE for *Corel371*. In the case of QBSE, we also compare the three possible query combination strategies: '*LKLD*','*SMN*', and '*KL Combination*'. It is clear that, inside the semantic space, the gains achieved with multiple QBSE queries are unparalleled on the visual domain. In [143], the authors have experimented with multiple query images on a QBVE system. They show that, using two examples, precision increases by around 15% at 10% recall (over single example queries) but no further improvements are observed for three or more images. We have found that, while the MAP of QBSE increases with the number of images, no gain is observed under QBVE. For QBSE, among the various combination methods, combining SMNs yields best results, with a gain of 29.8% over single image queries. '*LKLD*' and '*KL Combination*' exhibit a gain of 17.3% and 26.4% respectively. For QBSE, the increase of precision with query cardinality is experienced at all levels of recall.

3.8 shows the performance of 1-9 image queries for the best and the worst ten classes, sorted according to the gain in MAP score. It is interesting to note that in all of the best 10 classes, single image query performs well above chance, while the opposite holds for the worst 10. This means that moderate performance of a QBSE system can be considerably enhanced by using multiple query images, but this is not a cure for fundamental failures. Overall, the MAP score increases with the number of queries for 76% of the classes. For the classes with unsatisfactory

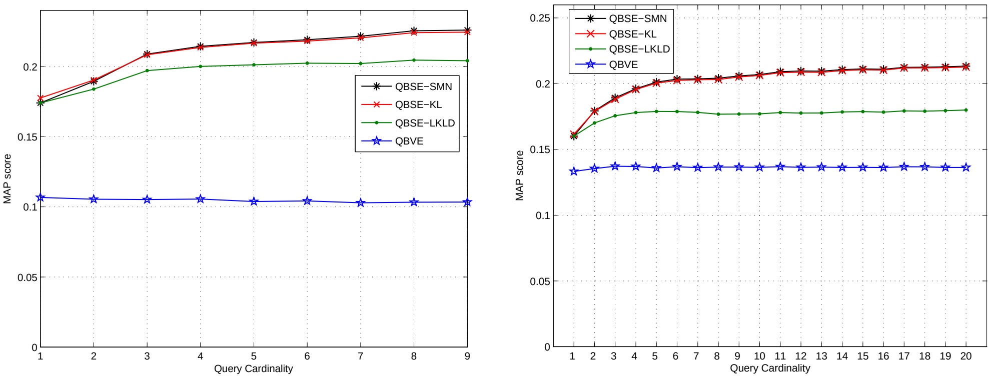


Figure 3.7: MAP as a function of query cardinality for multiple image queries. Comparison of QBSE, with various combination strategies, and QBVE. Left: Inside the semantic space (*Corel371*), Right: Outside the semantic space (*Flickr18*).

MAP score, poor performance can be explained by 1) significant inter-concept overlap (e.g., 'Air Shows' vs. 'Aviation Photography'), 2) incongruous concepts that would be difficult even for a human labeler (e.g. 'Holland' and 'Denmark'), or 3) failure to learn semantic homogeneity among the images, e.g. 'Spirit of Buddha'. Nevertheless, for 86% of the classes QBSE outperforms QBVE by an average MAP score of 0.136. On the remaining QBVE is only marginally better than QBSE, by an average MAP score of 0.016. 3.9 (Left) presents the average precision-recall curves, obtained with the number of image queries that performed best, for QBSE and QBVE on *Corel371*. It is clear that QBSE significantly outperforms QBVE at all levels of recall, the average MAP gain being of 111.73%.

3.6.6 Performance Outside the Semantic Space

3.4 (Right) presents precision-recall curves obtained on *Flickr18*[5], showing that outside the semantic space single-query QBSE is marginally better than QBVE. When combined with 3.4 (Left), it confirms that, overall, single-query QBSE has better generalization than visual similarity: it is substantially better inside the semantic space, and has slightly better performance outside of it.

[5] For brevity, we only document the results obtained with *Flickr18*, those of *Corel15* were qualitatively similar

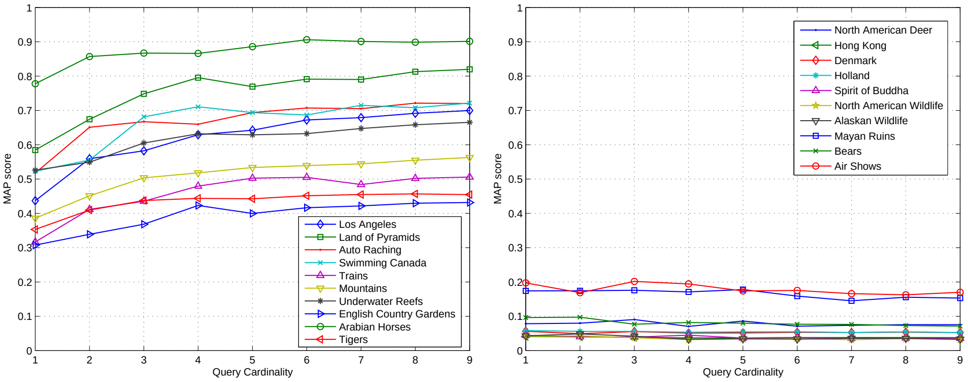


Figure 3.8: Effect of multiple image queries on the MAP score of various classes from *Corel371*. Left: Classes with highest MAP gains, Right: Classes with lowest MAP gains

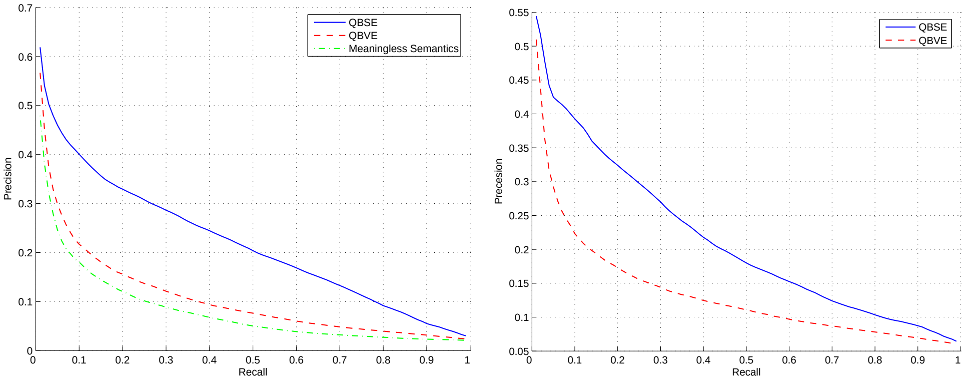


Figure 3.9: Best precision-recall curves achieved with QBSE and QBVE on *Corel371*. Left: Inside the semantic space (*Corel371*), also shown is the performance with meaningless semantic space. Right: Outside the semantic space (*Flickr18*).

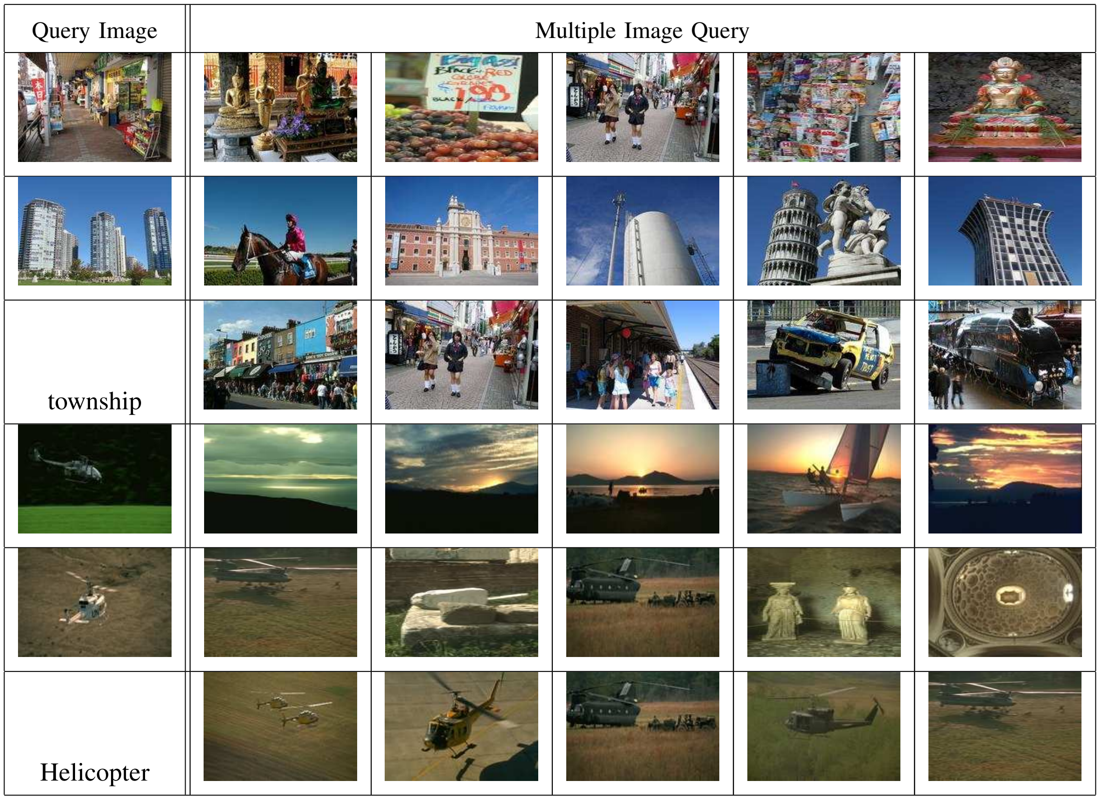


Figure 3.10: Examples of multiple-image QBSE queries. Two queries (for “Township” and “Helicopter”) are shown, each combining two examples. In each case, two top rows presents the single-image QBSE results, while the third presents the combined query.

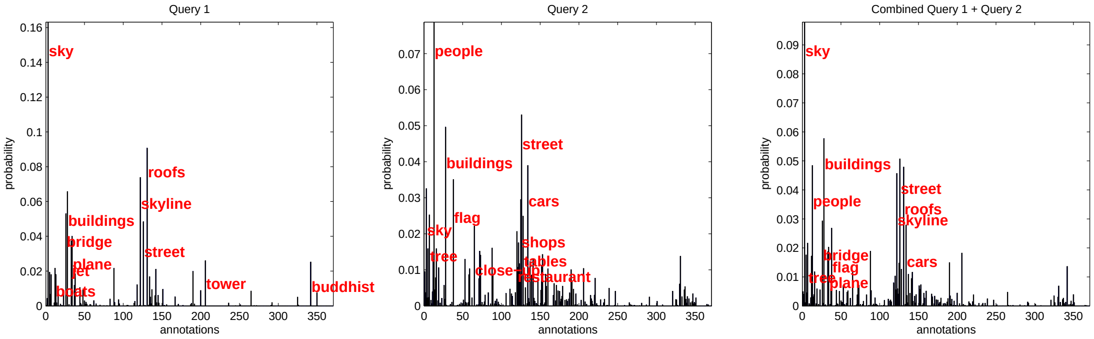


Figure 3.11: SMN of individual and combined queries from class 'Township' of 3.10. Left column shows the first query SMN, center the second and, right the combined query SMN.

For multiple image queries we performed experiments with up to 20 images per query (both databases contain 20 test images per class). As was the case for *Corel371*, multiple image queries benefit QBSE substantially but have no advantage for QBVE. This is shown in 3.7 (Right), where we present the MAP score as a function of query cardinality. With respect to the combination strategy, 'SMN' once again outperforms 'KL'(slightly) and 'LKLD Combination' (significantly).

An illustration of the benefits of multiple image queries is given in 3.10. The two top rows present query images from the class 'Township'(*Flickr18*) and single-query QBSE retrieval results. The third row presents the result of combining the two queries by 'SMN *combination*'. It illustrates the wide variability of visual appearance of the images in the 'Township' class. While single-image queries fail to express the semantic richness of the class, the combination of the two images allows the QBSE system to expand 'indoor market scene' and 'buildings in open air' to an 'open market street' or even a 'railway platform'. This is revealed, by the SMN of the combined query, presented in 3.11 (right), which is a semantically richer description of the visual concept 'Township', containing concepts (like 'sky', 'people', 'street', 'skyline') from both individual query SMNs. The remaining three rows of 3.10 present a similar result for the class 'Helicopter' (*Corel15*).

Finally, 3.9 presents the best results obtained with multiple queries under both the QBSE and QBVE paradigms. A similar comparison, using the precision-

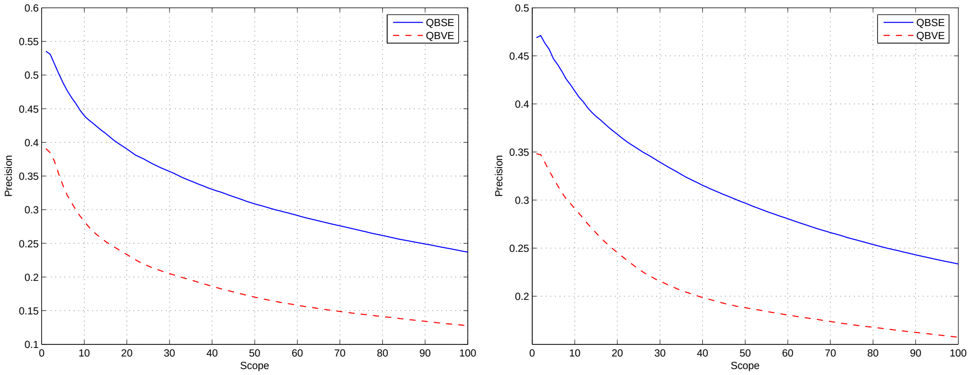


Figure 3.12: Performance of QBSE compared to QBVE, based on precision-scope curve for $N = 1$ to 100, Left: Inside the semantic space (*Corel371*), Right: Outside the semantic space (*Flickr18*).

Table 3.4: MAP of QBVE and QBSE on all datasets considered.

Database	**Chance**	**QBVE**	**QBSE**	**% increase**
Corel371	0.0200	0.1067	0.2259	111.73
Corel15	0.0667	0.2176	0.2980	36.95
Flickr18	0.0556	0.1373	0.2134	55.47

scope curve is shown in 3.12. It is clear that, when multiple image queries are adopted, QBSE significantly outperforms QBVE, even outside the semantic space. 3.4 summarizes the MAP gains of QBSE, over QBVE, for all datasets considered. In the case of *Flickr18* the gain is of 55.47%. Overall, the table emphatically points out that QBSE significantly outperforms QBVE, both inside and outside the semantic space. Since the basic visual representation (DCT features and Gaussian mixtures) is shared by the two approaches, this is strong indication that *there is a benefit* to the use of semantic representations in image retrieval. To further investigate this hypothesis we performed a final experiment, based on QBSE with a semantically meaningless space. Building on the fact that all semantic models are learned by grouping images with a common semantic concept, 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. 3.9 (left) compares (on *Corel371*) the precision-recall obtained with QBSE on this ‘meaningless semantic space’, with the previous results of QBVE and QBSE. It is clear that, in the absence of semantic structure, QBSE has *very poor* performance, and is *clearly inferior* to QBVE.

3.7 Acknowledgments

The author would like to thank Gustavo Carneiro and Pedro Moreno for helpful discussions and comments.

The text of Chapter 3, in part, is based on the material as it appears in: N. Rasiwasia, P. J. Moreno and N. Vasconcelos, *‘Bridging the Semantic Gap: Query by Semantic Example’*, IEEE Transactions on Multimedia, 9(5), 923-938, August 2007, N. Rasiwasia, P. J. Moreno and N. Vasconcelos, *‘Query by Semantic Example’*, ACM International Conference on Image and Video Retrieval, LNCS 51-60, Phoenix, 2006, and N. Rasiwasia and N. Vasconcelos, *‘A Systematic Study of the role of Context on Image Classification’*, IEEE Conference on Image Processing, 1720-1723, San Diego, Oct 2008. The dissertation author was a primary researcher and an author of the cited material.