Image Retrieval using Query by Contextual Example

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

ABSTRACT

Current image retrieval techniques have difficulties to retrieve images which exhibit distinct visual patterns but belong to the class of the query image. Previous attempts to improve *generalization* have shown that the introduction of semantic representations can mitigate this problem. We extend the existing query-by-semantic-example (QBSE) retrieval paradigm by adding a second layer of semantic representation. At the first level, the representation is driven by patch-based visual features. Semantic concepts, from a predefined vocabulary, are modeled as Gaussian mixtures on a visual feature space, and images as vectors of posterior probabilities of containing each of the semantic concepts. At the second level, the representation is purely semantic. Semantic concepts are modeled as Dirichlet mixtures on the semantic feature space of QBSE, and images are again represented as vectors of posterior concept probabilities. It is shown that the proposed retrieval strategy, referred to as *query-by-contextual-example* (QBCE), is able to cope with the ambiguities of patch-based classification, exhibiting significantly better generalization than previous methods. An experimental evaluation on benchmark datasets shows that QBCE retrieval systems can substantially outperform their QBVE and QBSE counterparts, achieving high precision at most levels of recall.

Categories and Subject Descriptors: I.5.4 [Pattern Recognition Applications]: Computer Vision

General Terms: Algorithms

Keywords: Dirichlet Distribution, Image Retrieval, Semantic Spaces, Query-by-example.

1. INTRODUCTION

Content-based image retrieval is a problem that involves many fundamental questions in computer vision, such as how to represent images, or to evaluate distances among them, when there is little control over 1) the image acquisition process, or 2) the classes of scenes to which the images report. It can be seen as an extension of image classification, that places greater emphasis on *generalization*: in addition to identifying the class of an image, the retrieval system should individually rank all the images from that class as more relevant to the query than images from other classes. Image retrieval has been an active subject of research over the last decades [3] when various retrieval strategies have been proposed.

Among these, query-by-example has been the most extensively used. Each image is decomposed into a number of *features vectors*, and retrieval is based on an example (query) image. In the early years, query-by-example systems invariably represented images in terms of *low-level visual features* (e.g. color, texture or shape histograms) [5, 16, 11, 12]. We refer to this retrieval strategy as *query-by-visual-example* (QBVE). Extensive evaluation of QBVE systems revealed that it lacks the generalization ability required in the retrieval setting. While most QBVE systems can easily find *iconic matches*, e.g. two images of the same scene, taken on the same day, from similar camera angles, they very rarely bridge the gap between two images from the same class which exhibit distinct visual patterns.

To address this limitation, various authors have proposed an alternative query-by-example strategy which extends QBVE to the semantic domain [17, 18, 10, 14, 13, 8]. This strategy, commonly referred to as *query-by-semantic-example* (QBSE), formulates image retrieval as a two stage process. The first stage consists of a *semantic labeling system*. A vocabulary of semantic concepts is defined and each semantic concept modeled as a probability distribution on the space of low-level visual features. Images are then fed to this semantic labeling system and represented as vectors of posterior probabilities of containing each of the semantic concepts. This is illustrated in Figure 1 (bottom right), which depicts a posterior probability vector for the image on the left and a vocabulary of 15 concepts. Posterior concept probabilities can be interpreted as *high-level semantic features*, rendered by the projection of images onto the abstract space of semantic concepts. The second stage performs all retrieval decisions on this semantic feature space, using the query-by-example principle.

Image representation by the probability vector of Figure 1 is an improvement over the representation by low-level visual features. It summarizes the information of interest for image classification: each probability can be seen as the best shot of the retrieval system at classifying the image into each of the classes. Rather than greedily attempting to classify the image, this representation captures the fact that, based on the underlying visual representation, it is only possible to rule out some class assignments, not to eliminate all ambiguities of image classification. These ambiguities are exactly what is left in the vector of posterior concept probabilities associated with the image. By exploiting the statistical structure of these ambiguities, a QBSE system is able to perform inferences at an higher level of abstraction, and significantly outperform QBVE systems. This has been confirmed by various recent studies, which

MIR'08, October 30–31, 2008, Vancouver, British Columbia, Canada.

have shown that QBSE systems can generalize much better than their QBVE counterparts [13].

Nevertheless, QBSE systems still suffer from some of the problems of QBVE, namely precision-recall (PR) curves that decay faster than acceptable, and less than ideal generalization (low precision at high levels of recall). To address this problem, we propose one further extension of QBSE along the dimension of abstraction. We argue that there exist two types of ambiguity in patch-based image classification, which are due to two types of patch co-occurrences: the co-occurrence of similar patches in images from semantically unrelated visual classes (a generalization of the notion of *polysemy* in text retrieval), and the co-occurrence of similar patches in images from semantically related visual classes (known as *synonymy* in text retrieval, and *context* in the vision literature). We then argue that the former are *accidental* and complicate the classification problem, while the latter are *stable* and facilitate it. This suggests that improved performance should be possible by extending the QBSE model by one further layer of semantic modeling.

In particular, we propose a 2-level semantic hierarchy. The first level consists of a semantic space identical to that of QBSE. The second level then models the distribution of each class in this space, through generative modeling of the semantic probability vectors derived from the images in the class. This produces a new, and more abstract, representation for the semantic concepts known to the retrieval system. We argue that this representation emphasizes stable patch co-occurrences, due to contextual relationships, and inhibits accidental ones. As in QBSE, images are represented by their posterior probabilities under these second-tier semantic concept distributions, i.e. they are projected into the second-tier semantic space. Retrieval then proceeds by example, in this space. We refer to this process as *query-by-contextual-example* (QBCE).

Overall, there are two levels of representation. At the first level, referred to as the *semantic level*, semantic concepts are modeled as distributions on the space of visual features. At the second level, the *contextual level*, they are modeled as distributions on the semantic space of posterior concept probabilities. We present an implementation of QBCE, where concepts are modeled as mixtures of Gaussian distributions on visual space, and mixtures of Dirichlet distributions on semantic space. Experimental evaluation shows that QBCE has significantly better retrieval performance than both QBVE and QBSE, and in particular generalizes much better. In fact, the generalization ability of the new representation is unlike that of any other approach that we are aware of, achieving almost flat PR curves, with high precision at high levels of recall.

The paper is organized as follows. Section 2 provides the motivation behind the proposed image and class representations. In Section 3 we review the implementations of QBVE and QBSE and describe the proposed retrieval architecture. Implementation details of a retrieval system based on this architecture are presented in Section 4. In Section 5 we present an empirical evaluation of QBVE, QBSE and QBCE on benchmark datasets. Finally, Section 6 presents some conclusions.

2. MOTIVATION

It is well known that the performance of QBVE is limited by the lack of agreement between the similarity of low-level visual features and human judgments of similarity. Substantial research has shown that this limitation is intrinsic to the QBVE strategy. For example, the high dimensionality of image space makes holistic representations infeasible, both in terms of the number of examples required to learn accurate retrieval functions, and the complexity of the surface spanned by each image class (the invariance problem).

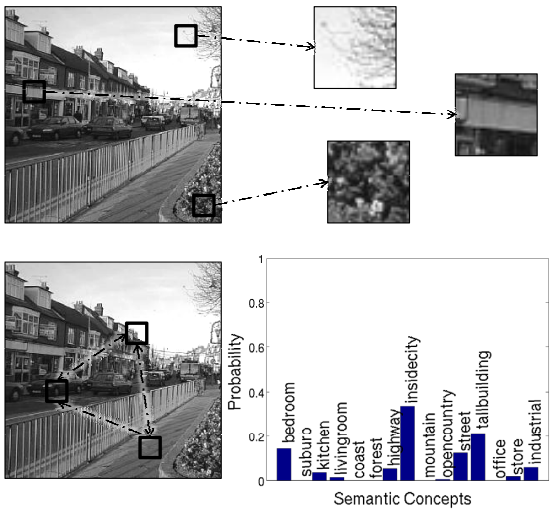

Figure 1: top row) Ambiguity co-occurrences. Image patches are frequently compatible with multiple classes. bottom left) Contextual co-occurrences. Patches of multiple other classes usually co-occur in the images of a given class. bottom right) Image representation by a vector of posterior probabilities.

This has led to the widespread use of patch-based image representations, which do not account for global image structure. These patch-based representations have two major limitations, illustrated in Figure 1. First, as shown on top, they are ambiguous: when considered in isolation, an image patch is usually compatible with many scene classes. It is unclear that even a human could confidently assign the highlighted patches to the class "Street", with which the image is labeled. Second, they lack information about the interdependence between the patches which compose the images in a class. For example, as shown on the bottom left, that images of street scenes typically contain patches of street, car wheels, and building texture. We refer to these two observations as *co-occurrences*. In the first case, a patch can co-occur with multiple classes (a property that is usually referred to as *polysemy* in the text analysis literature). In the second, patches from multiple classes typically co-occur in scenes of a given class (the equivalent to *synonymy* for text). Co-occurrences of the second type have been a subject of vigorous study, under the heading of *context*, in recent computer vision research [21], and are known to be informative clues for image classification.

The simplest possible form of context modeling is to keep track of the number of times that patches of different classes co-occurred in an image, as shown in Figure 1(bottom right). This is the rationale behind the more recent QBSE strategy. Note, however, that this probability vector, usually referred as a *semantic multinomial* (SMN), captures the two types of co-occurrences discussed above. On one hand, it reflects the ambiguity between "street scene" patches and patches of "highway", "bedroom", "kitchen" or "living room" (but not those of natural scenes, such as "mountain", "forest", "coast", or "open country", which receive close to zero probability). On the other, it reflects the likely co-occurrence, in "street scenes", of patches of "inside city", "street", "buildings", and "stores". While the various probabilities can be interpreted as *semantic features*, which account for co-occurrences due to ambiguity and context, they are not purely *contextual features*. Nevertheless, because images from the same class are expected to exhibit similar co-occurrences, it appears sensible to build retrieval systems on this semantic space.

The second stage of a QBSE system is precisely a system for query-by-example in this space. This representation is similar to those adopted by recent works in image categorization [7, 20, 15], where the use of intermediate representations for documents or images have been shown to improve performance.

In this work, we exploit the property that the two types of co-occurrences, which are left in the SMNs of QBSE, have different *stability*. While the same contextual co-occurrences are expected to be found in most images of a given class (or maybe a "mixture of contextual co-occurrences"), the same is not likely to hold for ambiguity co-occurrences. Although the "street scenes" image of Figure 1 contains some patches that could also be attributed to the "bedroom" class, it is unlikely that this will hold for most images of street scenes. By definition, ambiguity co-occurrences are *accidental*, otherwise they would reflect the presence of common objects in the two classes, and would be contextual coincidences. This leads to two observations. First, they must compromise the accuracy of the similarity matches performed by QBSE. Second, they should be possible to detect by joint inspection of all probability vectors derived from images in the same class.

This suggests the extension of the QBSE model by one further layer of semantic modeling. By modeling the probability distribution of the SMNs derived from the images in each class, it should be possible to obtain class representations that assign substantial probability to the regions of the semantic space occupied by contextual co-occurrences, and much smaller probability to those associated with ambiguity co-occurrences. Images could then be represented by their posterior probabilities under these higher level semantic models. Such a representation would emphasize contextual co-occurrences, while suppressing the accidental coincidences due to ambiguity, enabling more reliable similarity judgments. We refer to the posterior probabilities at this higher level of semantic representation as *contextual features*, the probability vector associated with each image as a *contextual multinomial* distribution, and the space of such vectors as the *contextual space*.

The hierarchical representation is similar to recent developments in the text and object recognition literatures, namely the use on non-parametric Bayesian models with multiple hierarchical levels [1]. We emphasize two main differences. First, our goal is not to learn taxonomies or other types of hierarchical class representations. The goal of the hierarchy now proposed is simply to deal with the ambiguity of the patch-based representation, and enable more reliable contextual inferences. In fact, the semantic classes are the same at the two levels of the hierarchy. Second, rather than relying on unsupervised learning, which attempts to simultaneously determine the vocabulary and the high level features, the intermediate vocabulary of the proposed system is usually pre-specified. This enables the use of weakly supervised learning methods, for which the learning task is easier, and which can be trained from vast amounts of data.

3. REPRESENTATION HIERARCHY

The retrieval formalism adopted at all levels of the retrieval hierarchy now proposed is that of minimum probability of error (MPE) retrieval [19]. We start by briefly reviewing the application of this formalism to QBVE and QBSE, and then propose a hierarchical extension compatible with the MPE framework.

3.1 Visual representation

At the lowest level, images are characterized as observations from a random variable $\mathbf{X}$, defined on some feature space $\mathcal{X}$, of visual measurements. For examples $\mathcal{X}$ could be the space of discrete cosine transform (DCT), or SIFT descriptors. Each image is represented as a set of n *feature vectors* $\mathcal{I} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}, \mathbf{x}_i \in \mathcal{X}$ assumed to be sampled independently. The starting point for retrieval is an image dataset $\mathcal{D} = \{\mathcal{I}_1, \ldots, \mathcal{I}_{|D|}\}$, where each image is associated with a class $y_i, i \in \{1, \ldots, |D|\}$, determined by a random variable Y defined on $\{1, \ldots, K\}$. Each class induces a probability density on $\mathcal{X}$ and, under the assumption of independent sampling,

$$P_{\mathbf{X}|Y}(\mathcal{I}|y) = \prod_j P_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}_j|y). \quad (1)$$

In this work, the class distributions $P_{\mathbf{X}|Y}(\mathbf{x}|y)$ are modeled as mixtures of Gaussian distributions [19].

Under QBVE, each image is considered a class in itself, i.e. Y is defined on $\{1, \ldots, |D|\}$, and y_i indicates the index of the image. 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.

$$y^* = \arg\max_y P_{Y|\mathbf{X}}(y|\mathcal{I}_q). \quad (2)$$

Thus, retrieval is based on the mapping $g : \mathcal{X} \rightarrow \{1, \ldots, |D|\}$ of (2), implemented by combining (1) and Bayes rule. Although any prior class distribution $P_Y(i)$ can be supported, we assume a uniform distribution.

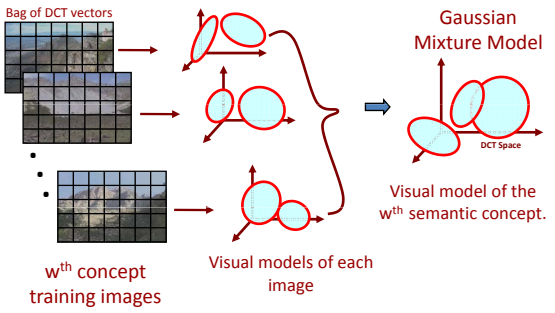

Figure 2: Learning a semantic concept density on visual space, $\mathcal{X}$, from the set $\mathcal{D}_w$ of all training images annotated with the w^{th} concept in $\mathcal{L}$.

3.2 Semantic representation

At the next level, images are represented on a *semantic space*. To enable such a representation, dataset $\mathcal{D}$ is augmented with a vocabulary $\mathcal{L} = \{w_1, \ldots, w_L\}$ of semantic concepts w_i, and each image $\mathcal{I}_i$ with a pre-specified caption $\mathbf{c}_i$. Here $\mathbf{c}_i$ is a binary L-dimensional vector such that $\mathbf{c}_{i,j} = 1$ if the i^{th} image was annotated with the j^{th} concept in $\mathcal{L}$. Concepts are drawn from a random variable W, which takes values in $\{w_1, \ldots, w_L\}$. Each concept induces a probability density on $\mathcal{X}$, from which feature vectors are drawn, and

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

For each concept w, the concept 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 $\mathcal{L}$, as shown in Figure 2. In this work, $P_{\mathbf{X}|W}(\mathbf{x}|w)$ are also modeled as mixtures of Gaussian Distributions, and directly estimated from the image densities used for QBVE in (1), using the method of [2].

Images are then represented as vectors of concept counts, $\mathcal{I} = (c_1, \ldots, c_L)^T$. Each feature vector $\mathbf{x_i}$, extracted from an image is assumed to be sampled from the probability distribution of a semantic concept, and c_i is the number of feature vectors drawn from

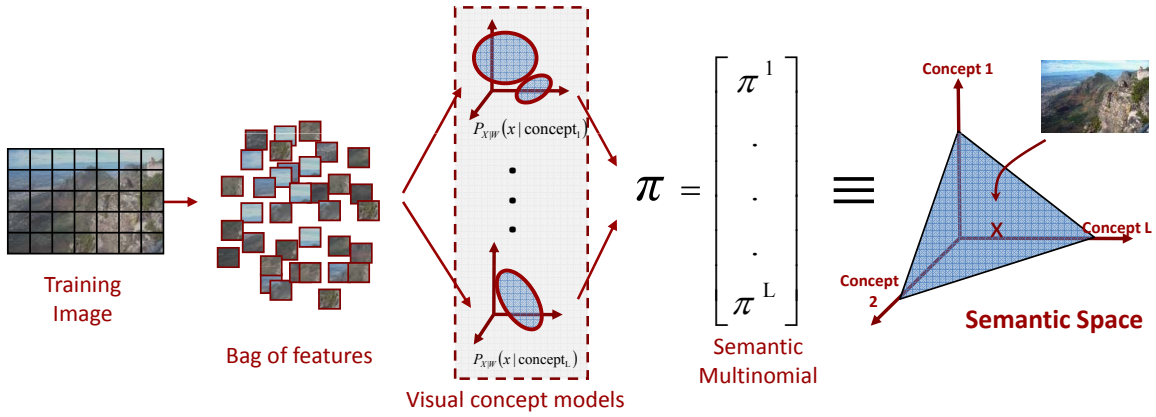

Figure 3: Image representation at the semantic level.

the i^{th} concept. The count vector for y^{th} image is drawn from a multinomial variable $\mathbf{T}_1$ of parameters $\boldsymbol{\pi}^y = (\pi_1^y, \ldots, \pi_L^y)^T$

$$P_{\mathbf{T_1}|Y}(\mathcal{I}|y;\boldsymbol{\pi}^y) = \frac{n!}{\prod_{k=1}^{L} c_k!} \prod_{j=1}^{L} (\pi_j^y)^{c_j}, \quad (4)$$

where π_i^y is the probability that an image feature vector is drawn from the i^{th} concept. Given an image $\mathcal{I} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, the posterior concept probabilities $\pi_w = P_{W|\mathbf{X}}(w|\mathcal{I})$ are maximum a posteriori estimates of the parameters π_w^y, and can be computed by combining (3) and Bayes rule, assuming a uniform prior concept distribution $P_W(w)$.

The vector $\boldsymbol{\pi}^y$ is referred to as a *s*emantic multinomial (SMN), and summarizes both the ambiguity and contextual co-occurrences of image y. The SMN vector lies on a probability simplex $\mathcal{S}_{L1}$, referred to as the *semantic space*. This representation establishes a one-to-one correspondence between images and the points $\boldsymbol{\pi}^y$ in $\mathcal{S}_{L1}$, as shown in Figure 3. A QBSE retrieval system performs a nearest neighbor operation on the simplex $\mathcal{S}_{L1}$, according to a similarity mapping $f_1 : \mathcal{S}_{L1} \rightarrow \{1, \ldots, |D|\}$ such that

$$f_1(\boldsymbol{\pi}) = \arg\min_y d(\boldsymbol{\pi}, \boldsymbol{\pi}^y) \quad (5)$$

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

$$d(\boldsymbol{\pi}, \boldsymbol{\pi}') = \sum_{i=1}^{L} \pi_i \log \frac{\pi_i}{\pi_i'}. \quad (6)$$

3.3 Contextual Representation

At the highest level, images are represented on a more stable *contextual space*, learned from the annotated dataset previously used to learn the semantic space $\mathcal{S}_{L1}$. In particular, the contextual space is built upon $\mathcal{S}_{L1}$, exploiting the fact that each concept w in $\mathcal{L}$ induces a probability distribution on this space, as illustrated in Figure 4. Since $\mathcal{S}_{L1}$ is itself a probability simplex, a suitable model is the Dirichlet distribution and, at contextual level, each concept is represented as a mixture of Dirichlet distributions

$$P_{\mathbf{\Pi}|W}(\boldsymbol{\pi}|w;\Omega^w) = \sum_k \beta_k^w \mathcal{D}ir(\boldsymbol{\pi};\boldsymbol{\alpha}_k^w). \quad (7)$$

This produces a vector of parameters $\Omega^w = \{\beta_k^w, \boldsymbol{\alpha}_k^w\}$ per concept w, where β_k is a probability mass function such that $\sum_k \beta_k^w = 1$

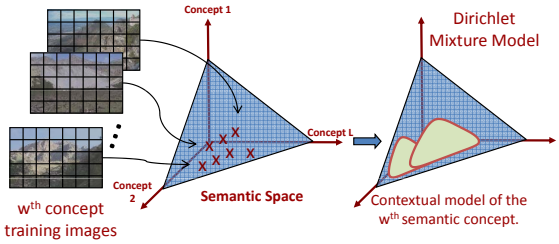

Figure 4: Learning a semantic concept density on semantic space, $\mathcal{S}_{L1}$, from the set $\mathcal{D}_w$ of all training images annotated with the w^{th} concept in $\mathcal{L}$.

and $\mathcal{D}ir(\boldsymbol{\pi};\boldsymbol{\alpha})$ a Dirichlet density of parameter $\boldsymbol{\alpha} = \{\alpha_1, \ldots, \alpha_L\}$,

$$\mathcal{D}ir(\boldsymbol{\pi};\boldsymbol{\alpha}) = \frac{\Gamma(\sum_{i=1}^{L} \alpha_i)}{\prod_{i=1}^{L} \Gamma(\alpha_i)} \prod_{i=1}^{L} (\pi_i)^{\alpha_i - 1} \quad (8)$$

where $\Gamma(.)$ represents the Gamma function. The parameters Ω^w for the w^{th} concept are learned from the SMNs $\boldsymbol{\pi}^y$ of all images in $\mathcal{D}_w$, i.e. the images annotated with the w^{th} concept.

Similar to what happens at semantic level, an image is represented as a vector of concept counts $\mathcal{I} = (c_1, \ldots, c_L)^T$. However, these are now assumed to be sampled from a mixture of Dirichlet distributions on $\mathcal{S}_{L1}$, in contrast to the mixture of Gaussians on $\mathcal{X}$ used to obtain the semantic level representation. The count vector for the y^{th} image is now assumed to be drawn from a multinomial variable $\mathbf{T}_2$ of parameters $\boldsymbol{\theta}^y = (\theta_1^y, \ldots, \theta_L^y)^T$

$$P_{\mathbf{T}_2|Y}(\mathcal{I}|y;\boldsymbol{\theta}^y) = \frac{n!}{\prod_{k=1}^{L} c_k!} \prod_{j=1}^{L} (\theta_j^y)^{c_j}, \quad (9)$$

where θ_i^y is the probability that an image feature vector is drawn from the i^{th} concept. Given the image SMNs $\mathcal{I} \equiv \boldsymbol{\pi}^y = \{\pi_1^y, \ldots, \pi_L^y\}$, the posterior concept probabilities at contextual level, $\theta_w = P_{W|\Pi}(w|\boldsymbol{\pi}^y)$ are maximum a posteriori estimates of the parameters $\boldsymbol{\theta}^y$, and can be computed by combining (7) and Bayes rule, assuming a uniform prior concept distribution $P_W(w)$.

The vectors $\boldsymbol{\theta}^y$ are referred to as *contextual multinomials* (CMN) and lie on a new probability simplex $\mathcal{S}_{L2}$, here referred to as the *contextual space*. In this way, the contextual representation establishes a one-to-one correspondence between images and the points $\boldsymbol{\theta}^y$ in $\mathcal{S}_{L2}$, as shown in Figure 5. The retrieval operation is similar

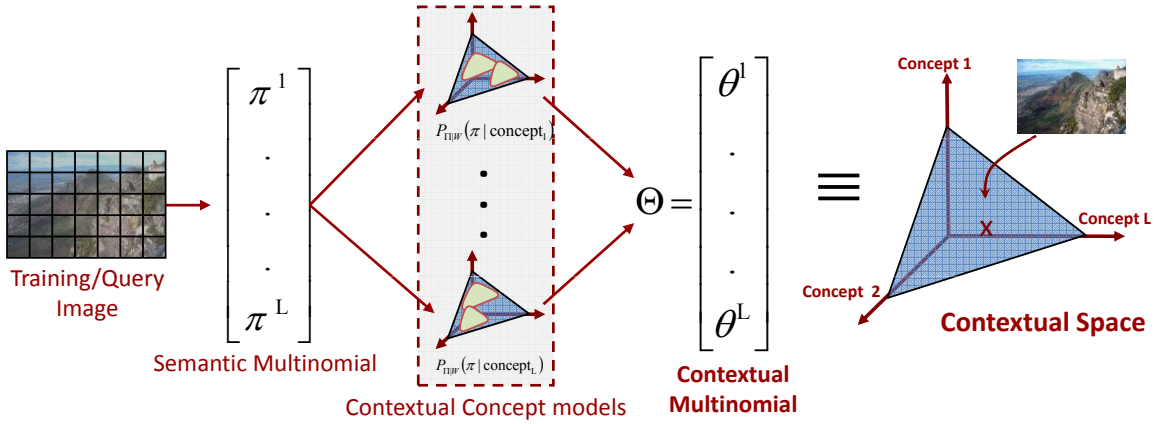

Figure 5: Image representation at the contextual level.

to that of QBSE, i.e. based on a similarity mapping $f_2 : \mathcal{S}_{L2} \rightarrow \{1, \ldots, |D|\}$ such that

$$f_2(\boldsymbol{\theta}) = \arg\min_y d(\boldsymbol{\theta}, \boldsymbol{\theta}^y) \tag{10}$$

We refer to nearest-neighbor operation of (10) as *query-by-contextual-example* (QBCE).

4. IMPLEMENTATION DETAILS

In this section we report on the implementation of a QBCE system. For brevity, we limit the discussion to the implementation details of contextual level. The visual and semantic levels are those proposed in [19] and [13], where they were shown to achieve better performance than a number of other state of the art image retrieval systems.

4.1 Learning the contextual space

The parameters of all Dirichlet mixture models are estimated by maximum likelihood, via the generalized expectation-maximization (GEM) algorithm. GEM is an extension of the well known EM algorithm, applicable when the M-step of the latter is intractable. It consists of two steps. The E-Step is identical to that of EM, computing the expected values of the component probability mass β_k. The generalized M-step estimates the parameters $\boldsymbol{\alpha}_k$. Rather than solving for the parameters of maximum likelihood, it simply produces an estimate of higher likelihood than that available in the previous iteration. This is known to suffice for convergence of the overall EM procedure. We resort to the Newton-Raphson algorithm to obtain these improved parameter estimates, as suggested in [9] (for single component Dirichlet distributions). Figure 6(left) shows a 3-component Dirichlet mixture model learned for the semantic concept "street". This model is estimated from 100 images (shown as data points on the figure), on a three concept semantic space. Notice that the Dirichlet model captures the contextual co-occurrences of the concepts "street" and "store" as the distribution assigns high probabilities to both of them. Also shown in the figure are the three components of the Dirichlet mixture.

5. EXPERIMENTAL EVALUATION

An empirical evaluation of the proposed QBCE retrieval paradigm was performed by comparing its performance with those of QBVE and QBSE, for two publicly available datasets.

5.1 Datasets

Image retrieval results are presented on two public datasets - 1) Natural15: 15-Natural scene classes [6] and 2) Corel15: 15-Corel stock photo CD's, used in [13] for QBSE. The Natural15 dataset contains 200-400 images per class with an average image size of 300×250px. This dataset has also been widely used in the scene classification literature. The 15-Corel dataset has 100 high resolution images per class, which we resize to an average of 180×120 pixels. In all cases, performance is measured with precision-recall (PR) curves and mean average precision (MAP) [4]. Given a query and the top n database matches, recall is the percentage of all relevant images contained in the retrieved set, and precision the percentage of the n which are relevant (where relevant means belonging to the class of the query). The MAP is then defined as the average precision, over all queries, at the ranks where recall changes (i.e., where relevant items occur).

Learning the semantic and contextual spaces requires a vocabulary of concepts and an annotated dataset. In general, concepts are different from image classes. For example, the image from the "Street" class of Figure 7 (top left) contains themes such as "Road", "Sky", "Buildings", and "Cars". However, in the absence of "concept" annotations in the training dataset, the image class (e.g. "Street") can serve as a proxy for the concept vocabulary. In this case, each image is only explicitly annotated with one "concept", even though it may depict multiple. We refer to this limited type of image labeling as *casual annotation*. This is the annotation mode for all results reported in this work.

5.2 Experimental Protocol

At the visual level, images are represented as bags of 8×8 vectors of DCT coefficients sampled on a uniform grid. Corel15 consists of color images which are converted from RGB to YCrCb colorspace. Natural15 consist of grayscale images. Semantic concept (class)[1] densities are learned on a 64 dimensional subspace of DCT coefficients, and each class modeled as a mixture of 128 Gaussian components. A varying number of Dirichlet components (1 to 20) is used to model the classes at contextual level. 100 (80) randomly selected images per class are used to learn the class mixture models for Natural15 (Corel15). Note that this set also serves as the dataset from which the images are retrieved. 50 randomly selected (20 remaining), but previously unused images per class serve as the

[1]Since we use a *casually* annotated dataset, semantic concepts are substituted with image classes.

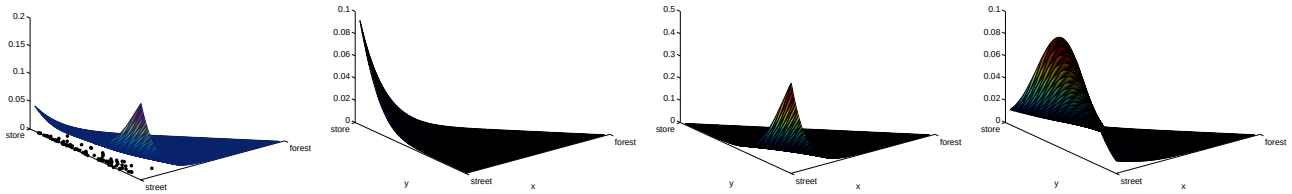

Figure 6: left) 3-component Dirichlet mixture model learned for the semantic concept "street". Also shown are the semantic multinomials associated with each image from this concept. remainder) The three components of the Dirichlet mixture model. The component probability masses β_k are 0.437, 0.332 and 0.231, from left to right.

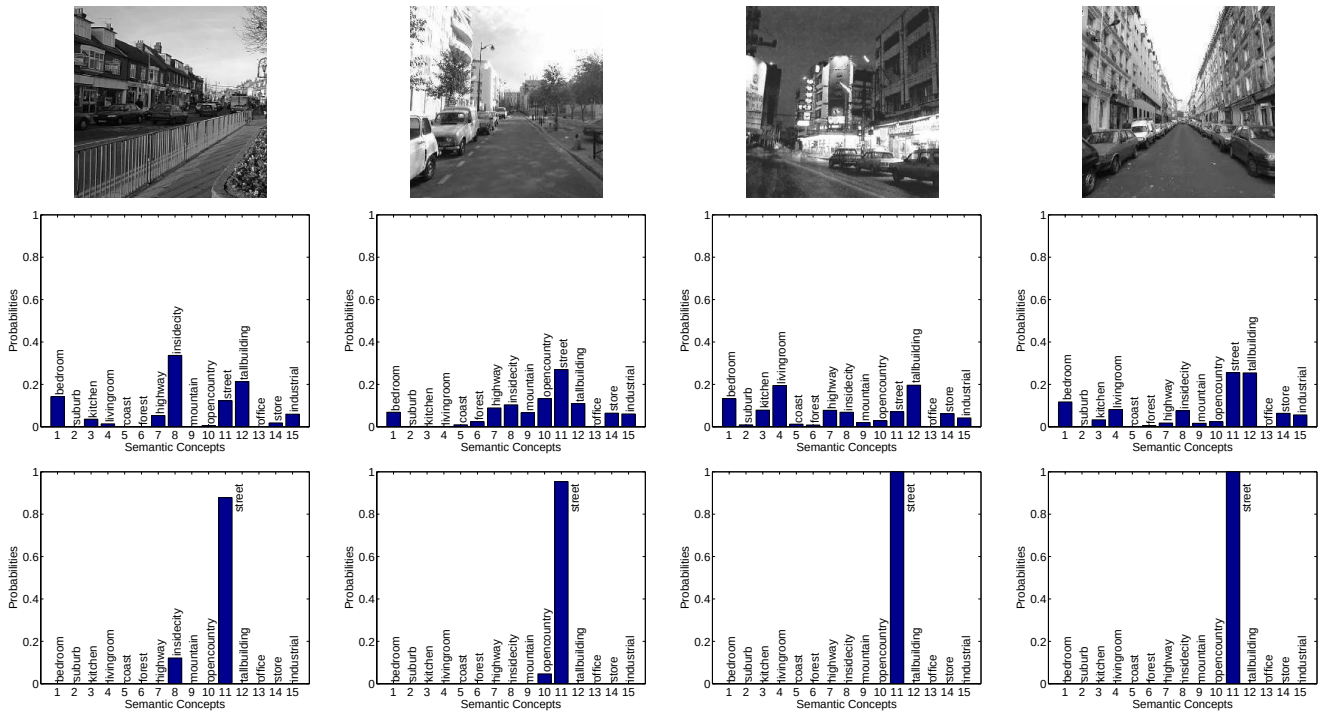

Figure 7: top row) Four images from the "Street" class of Natural15. middle row) Semantic multinomials of the images shown in the top row. bottom) Contextual multinomials of the images shown in the top row.

query set for Natural15 (Corel15). Finally we also show results obtained with SIFT descriptors on Natural-15, as these descriptors are quickly gaining popularity.

5.3 Results

In this section we report on the image retrieval results. We first provide some examples of the proposed representation, which illustrate its capacity to capture contextual co-occurrences. Next, we compare the performance of QBVE, QBSE and QBCE on the datasets discussed above.

5.3.1 Image representation

Figure 7 (top row) shows four images from the "Street" class of Natural15. The SMN and CMN vectors corresponding to each image are also shown (second and the third rows respectively). As discussed in Sec. 1, the SMN vectors capture the co-occurrence of visual patches across different concepts. It is evident that the visual patches of the given images, although belonging to the "Street" class, have high probability of occurrence under various other concepts, including "bedroom", "livingroom", "kitchen", "inside city", "tall building", etc. Some of these concepts, such as "bedroom", "livingroom", "kitchen" etc., are due to accidental co-occurrences while the others, such as "inside city", "tall building", result from contextual co-occurrences. Based on the SMN of a single image, it is impossible to distinguish these two types of co-occurrence.

The image representation is substantially more robust in contextual space, because the retrieval system learns the statistical structure of the co-occurrences associated with a given class from all SMNs in that class. For example, even though the image in column one has a low probability under the "street" class, at semantic level, the complete co-occurrence pattern (its SMN) has a high probability, under the same class, at the contextual level. Furthermore, class models at the contextual level mitigate accidental co-occurrences, while accentuating contextual co-occurrences. For example, the accidental co-occurrences of "street" with "kitchen" and "livingroom" are not consistent across all class images, while the contextual co-occurrences with "tall building" and "inside city" appear consistently throughout the class.

Figure 8: Some examples of query and retrieved images from the Corel15 dataset("Greece", "Helicopter", "Military Vehicles"). The query image is shown in the first column and the top 6 retrieved images in the other columns. Notice the visual variability of the query and retrieved images. This figure is best viewed in color.

Figure 9: Some examples of query and retrieved images from the Natural15 dataset("Street", "Open-country", "Industrial"). The query image is shown in the first column and the top 6 retrieved images in the other columns. Notice the visual variability of the query and retrieved images.

5.3.2 Retrieval Performance

The retrieval performance of the three systems - QBVE, QBSE and QBCE is summarized by Figure 10. Figure 10 (left, middle) present PR curves for the two datasets considered. It is evident that the precision of QBCE is *significantly* higher than that of QBSE and QBVE at almost all levels of recall. Only at very low levels of recall QBSE achieves performance similar to that of QBCE. This is due to the fact that, in most classes, there are images which are close visual matches to the query. Accurate retrieval of the remaining images (from the same class) requires matching at a much higher level of generalization. This is clearly exemplified by Figure 8, which shows retrieval results of QBCE on Corel15. The first column shows three queries (from three different classes - "Greece", "Helicopter", "Military Vehicles") and the remaining columns show the top six retrieved images. Notice that, in spite of high variability of visual appearance, QBCE is successful in retrieving images from the class of the query. A similar set of results, from Natural15, is shown in Figure 9. Comparing the performance of QBSE and QBVE, the former performs better than the latter. This is due to the higher level of abstraction of the representation used by QBSE [13]. Figure 10 (right) presents the MAP scores obtained on both datasets, as a function of the number of components in the Dirichlet mixture model. Performance increases initially, until about 12 components, but stabilizes as the number of components is further increased.

Figure 11 shows the MAP scores for the two datasets, across different retrieval paradigms, and compares them to the chance MAP score. For Corel15, the best MAP score of 0.610 is obtained with QBCE, a gain of 27.08% over QBSE. For Natural15, QBCE yields a MAP score of 0.563, which is 46.50% higher than that of QBSE. Retrieval results obtained with SIFT descriptors on Natural15 are also shown in Figures 10 (middle) and 11. The QBCE MAP score obtained with SIFT is 0.627, an improvement of 18.25% over the corresponding QBSE score. This is an indication that the gains of QBCE are not critically dependant on the low-level visual features, although careful feature selection can improve the absolute retrieval performance at all levels of the hierarchical representation.

6. CONCLUSION

We have presented a new approach to content based image retrieval. This approach extends the existing QBSE retrieval paradigm by addition of a second layer of semantic modeling, to produce a second-tier semantic space. Semantic classes at the highest level, referred to as the *contextual level*, are represented as mixtures of Dirichlet distributions, and model the statistical structure of the contextual co-occurrences between concepts. Images are then represented as posterior probabilities under these distributions. It is argued that the proposed representation emphasizes contextual co-occurrences, while suppressing accidental coincidences due to patch classification ambiguity, and enables more reliable similarity

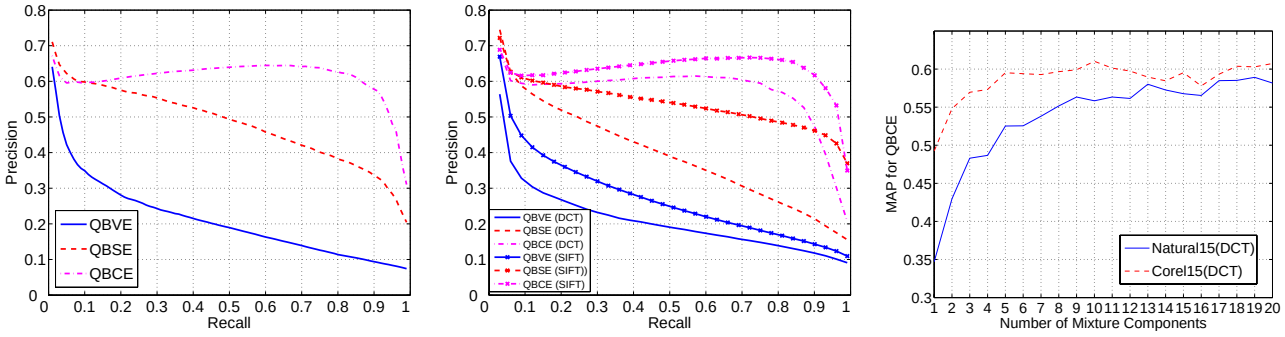

Figure 10: Comparison of precision-recall curves for QBVE, QBSE and QBCE. left) Corel15, middle) Natural15. right) The variation of the MAP scores with the number of mixture components of the Dirichlet distribution.

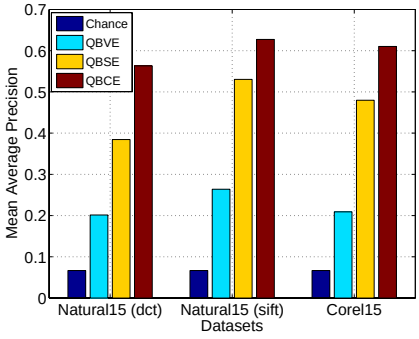

Figure 11: Comparison of the MAP scores for the various evaluations considered in the text. Also shown is the chance retrieval performance.

judgments. Experimental results on benchmark datasets show that QBCE significantly outperforms the previously available QBVE and QBSE retrieval strategies, and is able to achieve high levels of precision at almost all levels of recall, thus leading to much better generalization.

7. REFERENCES

[1] D. Blei and M. Jordan. Modeling annotated data. In *Proc. ACM SIGIR conf. on Research and development in information retrieval*, 2003.

[2] 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.

[3] 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.

[4] S. Feng, R. Manmatha, and V. Lavrenko. Multiple bernoulli relevance models for image and video annotation. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Washington DC, 2004.

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

[6] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. *Proc. IEEE Conf. Comp. Vision Patt. Recog*, 2005.

[7] F.-F. Li and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *IEEE CVPR*, pages 524–531, 2005.

[8] J. Magalhães, S. Overell, and S. Rüger. A semantic vector space for query by image example. *ACM SIGIR Special Interest Group on Information Retrieval*, 2007.

[9] T. Minka. Estimating a Dirichlet distribution. *http://research.microsoft.com/ minka/papers/dirichlet/*, 1:3, 2000.

[10] 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.

[11] 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, San Jose, California.

[12] 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.

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

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

[15] J. Sivic, B. Russell, A. Efros, A. Zisserman, and W. Freeman. Discovering object categories in image collections. *Proc. ICCV*, 1:65, 2005.

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

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

[18] 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, Volume 5307, pp. 271-279 (2003).*, pages 271–279, 2003.

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

[20] J. Vogel and B. Schiele. A semantic typicality measure for natural scene categorization. *DAGMŠ04 Annual Pattern Recognition Symposium*.

[21] L. Wolf and S. Bileschi. A critical view of context. *International Journal of Computer Vision*, 69(2):251–261, 2006.