Using Dual Cascading Learning Frameworks for Image Indexing

Joo-Hwee Lim

Institute for Infocomm Research
21 Heng Mui Keng Terrace, Singapore 119613
Email: joohwee@i2r.a-star.edu.sg

Jesse S. Jin

The University of Sydney
Sydney 2006, Australia
Email: jesse@it.usyd.edu.au

Abstract

To bridge the semantic gap in content-based image retrieval, detecting meaningful visual entities (e.g. faces, sky, foliage, buildings etc) in image content and classifying images into semantic categories based on trained pattern classifiers have become active research trends. In this paper, we present dual cascading learning frameworks that extract and combine intra-image and inter-class semantics for image indexing and retrieval.

In the supervised learning version, support vector detectors are trained on semantic support regions without image segmentation. The reconciled and aggregated detection-based indexes then serve as input for support vector learning of image classifiers to generate class-relative image indexes. During retrieval, similarities based on both indexes are combined to rank images.

In the unsupervised learning approach, image classifiers are first trained on local image blocks from a small number of labeled images. Then local semantic patterns are discovered from clustering the image blocks with high classification output. Training samples are induced from cluster memberships for support vector learning to form local semantic pattern detectors. During retrieval, similarities based on local class pattern indexes and discovered pattern indexes are combined to rank images.

Query-by-example experiments on 2400 unconstrained consumer photos with 16 semantic queries show that the combined matching aproaches are better than matching with single indexes. Both the supervised semantics design and the semantics discovery approaches also outperformed the linear fusion of color and texture features significantly in average precisions by 55% and 37% respectively.

keywords: Image Indexing, Image Retrieval, Image Classification, Pattern Discovery, Similarity Matching

1 Introduction

Started more than a decade ago, content-based image retrieval research has yet to bridge the "semantic gap between the information that one can extract from the visual data and the interpretation that the same data have for a user in a given situation" [26]. The semantic gap is due to two inherent problems. On one hand, the extraction of complete semantics from image data is extremely hard as it demands general object recognition and scene understanding. On the other hand, user interpretation of image data is inherently subjective.

In fact, broad domain such as unconstrained consumer photos pose great challenge for content-based image retrieval research. Unlike professional images or images of a narrow domain, the amount of content variations in consumer photos is usually very high due to the spontaneous and casual nature during image capturing. More often than not, the objects in the photos are ill-posed, occluded, and cluttered with poor lighting, focus, and exposure.

In particular, a challenge for computer vision is the usually very large number of object classes in polysemic images. Indeed highly accurate segmentation of objects is a major bottleneck except for selected narrow domains when few dominant objects are recorded against a clear background ([26],p.1360). The interpretation of unconstrained images is usually not unique as it may have numerous conspicuous objects, for which some of them have unknown object classes [26]. Detecting semantic objects (e.g. faces, sky, foliage, buildings etc) systematically based on trained pattern classifiers has received serious attention lately (e.g. [21, 22, 29]).

On the other hand, categorization is a powerful divide-and-conquer metaphor to organize and access images. Once the images are sorted into semantic classes, searching and browsing can be carried out in a more effective and efficient way by focusing only at relevant classes and subclasses. Moreover the classes provide context for other tasks. For instance, in medical domain, images have been grouped by pathological classes for diagnostic purpose [6] or by imaging modalities for visualization purpose [20]. For less constrained domains such as vacation photos, a progressive approach that utilized different low-level features for classifying a hierarchy of categories has been proposed [30].

In this paper, we present dual cascading learning frameworks that extract and combine intra-image and inter-class semantics for image indexing and retrieval. Fig. 1 summarizes the dual frameworks.

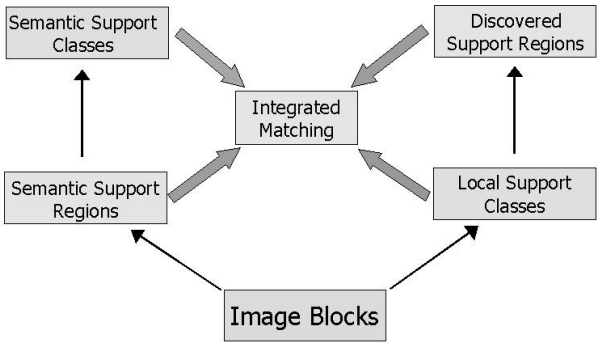

Figure 1: Dual cascading image indexing and matching frameworks

In the supervised learning version (left flow in Fig. 1), support vector detectors are trained on semantic support regions without image segmentation.

The reconciled and aggregated detection-based indexes then serve as input for support vector learning of image classifiers to generate class-relative image indexes. During retrieval, similarities based on both indexes are combined to rank images.

Town and Sinclair [29] described a semantic labeling approach to image retrieval. An image is segmented into non-overlapping regions and each subject to classification into 11 visual categories suited to outdoor scenes by neural networks. Similarity between a query and an image is computed as either the sum over all grids of the Euclidean distance between classification vectors, or their cosine of correlation. The evaluation was carried out on over 1000 Corel Photo Library images and about 500 home photos, with better results obtained for the Corel images. In another effort to detect 34 visual concepts in videos [22], support vector machines are trained on segmented regions of key frames using color and texture features.

A key innovation in our method is that no region segmentation is needed. Instead, visual concepts are learned and detected during image indexing from tessellated image blocks, as inspired by multi-scale view-based object recognition framework [23]. Moreover, the local detection decisions are reconciled across multiple resolutions and aggregated over spatial areas as semantic indexes.

While Naphade [21] proposed a probabilistic framework that enhances the detection of probabilistic multimedia objects called multijects in videos by modeling their inter-relationship in an explicit network form (multinet), our approach is simpler as both the local semantics and their implicit co-occurrence context are trained separately, hence simplifying the learning problem. In addition, segmented objects and sites (e.g. outdoor scene) are treated as equal entities as multijects [21]. In our case, segmentation-free block regions and image classes are represented at different levels of semantics as content and context respectively.

However, a major drawback of the supervised learning approach is the human effort required to provide labeled training samples, especially at the image region level. Lately there are two promising trends that attempt to achieve semantic indexing of images with minimal or no effort of manual annotation (i.e. semi-supervised or unsupervised learning).

In the field of computer vision, researchers have developed object recognition systems from unlabeled and unsegmented images [8, 25, 32]. In the context of relevance feedback, unlabeled images have also been used to bootstrap the learning from very limited labeled examples (e.g. [31, 33]). For the purpose of image retrieval, unsupervised models based on "generic" texture-like descriptors without explicit object semantics can also be learned from images without manual extraction of objects or features [24]. As a representative of the state-of-the-art, sophiscated generative and probabilistic model has been proposed to represent, learn, and detect object parts, locations, scales, and appearances from fairly cluttered scenes with promising results [8].

Motivated from a machine translation perspective, object recognition is posed as a lexicon learning problem to translate image regions to corresponding words [7]. More generally, the joint distribution of meaningful text descriptions and entire or local image contents are learned from images or categories of images labeled with a few words [1, 3, 10, 11]. The lexicon learning metaphor offers a new way of looking at object recognition [7] and a powerful means to annotate entire images with concepts evoked by what is visible in the image and specific words (e.g. fitness, holiday, Paris etc [11]). While the results for the annotation problem on entire images look promising [11], the correspondence problem of associating words with segmented image regions remains very challenging [3] as segmentation, feature selection, and shape representation are critical and non-trivial choices [2].

We address the issue of minimal supervision differently. We do not assume availability of text descriptions for image or image classes as in [3, 11]. Neither do we know the object classes to be recognized as in [8]. We wish to discover and associate local unsegmented regions with semantics and generate their samples to construct models for content-based image retrieval, all with minimal manual intervention. This is realized as a novel three-stage hybrid framework that interleave supervised and unsupervised learnings.

In the unsupervised learning approach (right flow in Fig. 1), image classifiers are first trained on local image blocks from a small number of labeled images. Then local semantic patterns are discovered from clustering the image blocks with high classification output. Training samples are induced from cluster memberships for support vector learning to form local semantic pattern detectors. During retrieval, similarities based on local class pattern indexes and discovered pattern indexes are combined to rank images.

Query-by-example experiments on 2400 unconstrained consumer photos with 16 semantic queries show that the combined matching aproaches are better than matching with single indexes. Both the supervised semantics design and the semantics discovery approaches also outperformed the linear fusion of color and texture features significantly in average precisions by 55% and 37% respectively.

The rest of the paper is presented as follows. The next two sections are devoted to the description of the semantics design (i.e. supervised) and semantics discovery (i.e. unsupervised) approaches respectively. Then the mechanism to integrate intra-image and inter-class semantics in similarity matching is presented followed by experimental results.

2 Semantics Design Approach

From a semantics design perspective, we propose a cascading framework to combine intra-image and inter-class semantics that are learned and extracted from images with two layers of binary pattern classifiers (Fig. 2).

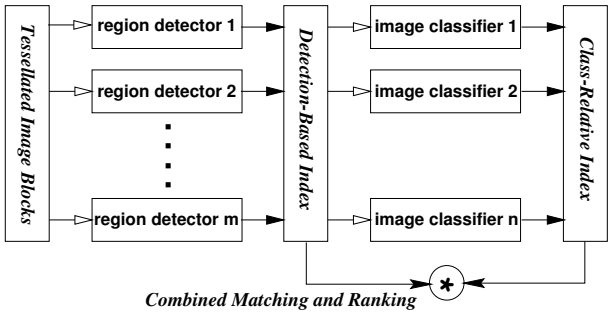

Figure 2: A cascading image indexing framework that combines intra-image and inter-class semantics

At the region level, local support vector detectors are trained on semantic support regions without image segmentation. The detection results are reconciled across multiple resolutions and aggregated spatially as image indexes. They also serve as input patterns for support vector image classifiers to learn and generate class-relative image indexes. During retrieval, similarities based on both types of indexes are

combined to rank images. Query-by-example experiments on 2400 heterogeneous consumer photos with 16 semantic queries show that the combined matching aproach is better than matching with single index. The proposed approach also achieved a very significant improvement of 55% in average precision than using combined matching of color and texture features.

2.1 Detection-Based Indexing

To realize strong semantic interpretation of image content, we propose the use of salient image regions that exhibit semantic meanings to human users as support for image indexing. They are similar to the *signs* designed for domain-specific applications ([26],p.1359) and the *visual keywords* built for explicit query specification [12, 13, 14].

Semantic support regions (SSR) are salient image patches that exhibit semantic meanings to us. A cropped face region, a typical grass patch, and a patch of swimming pool water etc can all be treated as their instances. In this paper, we train local support vector detectors on multi-scale block-based image regions, as inspired by multi-resolution view-based object recognition framework [23], hence without a region segmentation step. Given a local image patch with feature vector z, a support vector classifier $\mathcal{S}_i$ is a detector for SSR i on z. The classification vector T for region z can be computed via the softmax function [4] as

$$T_i(z) = \frac{\exp^{\mathcal{S}_i(z)}}{\sum_j \exp^{\mathcal{S}_j(z)}}. \quad (1)$$

As each support vector machine is regarded as an expert on a local semantic class, the outputs of $\mathcal{S}_i \ \forall i$ is set to 0 if there exist $\mathcal{S}_j, j \neq i$ that has a positive output.

As we are dealing with heterogeneous consumer photos, we adopt color and texture features to characterize SSR. A feature vector z has two parts, namely, a color feature vector z^c and a texture feature vector z^t. For the color feature, we compute the mean and standard deviation of each color channel (i.e. z^c has 6 dimensions). We use the YIQ color space over other color spaces as it performed better in our experiments. For the texture feature, we adopted the Gabor coefficients [19]. Similarly, the means and standard deviations of the Gabor coefficients (5 scales and 6 orientations) in an image block are computed as z^t (60 dimensions). Zero-mean normalization was applied to both the color and texture features. In this paper, we used polynomial kernels with a modified dot product similarity measure between feature vectors y and z,

$$y \cdot z = \frac{1}{2}\left(\frac{y^c \cdot z^c}{|y^c||z^c|} + \frac{y^t \cdot z^t}{|y^t||z^t|}\right) \quad (2)$$

To detect SSR with translation and scale invariance in an image to be indexed, the image is scanned with windows of different scales, following the strategy in view-based object detection [23]. In our experiments, we progressively increase the window size from 20×20 to 60×60 at a step of 10 pixels, on a 240×360 size-normalized image. That is, after this detection step, we have 5 maps of detection.

To reconcile the detection maps across different resolutions onto a common basis, we adopt the following principle: If the most confident classification of a region at resolution r is less than that of a larger region (at resolution $r+1$) that subsumes the region, then the classification output of the region should be replaced by those of the larger region at resolution $r+1$. Using this principle, we start the reconciliation from detection map based on largest scan window (60×60) to detection map based on next-to-smallest scan window (30×30). After 4 cycles of reconciliation, the detection map that is based on the smallest scan window (20×20) would have consolidated the detection decisions obtained at other resolutions.

Suppose a region Z comprises of n small equal regions with feature vectors $z_1, z_2, \cdots, z_n$. To account for the size of detected SSR in the spatial area Z, the SSR detection vectors of the reconciled detection map is aggregated as

$$T_i(Z) = \frac{1}{n}\sum_k T_i(z_k). \quad (3)$$

The segmentation-free indexing process is summarized in Fig. 3.

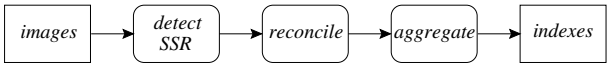

Figure 3: A schematic digram of detection-based image indexing

For the data set and experiments reported in this paper, we designed 26 classes of SSR (i.e. $S_i, i = 1, 2, \cdots, 26$ in Eq. (1)), organized into 8 superclasses as illustrated in Fig. 4. We cropped 554 image regions from 138 images and used 375 of them (from 105 images) as training data and the remaining one-third for validation. Among all the kernels evaluated, those with better generalization result on the validation set are used for the indexing and retrieval tasks. A polynomial kernel with degree 2 and constant 1 ($C = 100$ [9]) produced the best result on precision and recall. Hence it was adopted in our experiments.

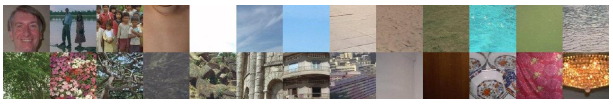

Figure 4: Examples of semantic support regions (top-down, left-to-right): people (face, figure, crowd, skin), sky (clear, cloudy, blue), ground (floor, sand, grass), water (pool, pond, river), foliage (green, floral, branch), mountain (far, rocky), building (old, city, far), interior (wall, wooden, china, fabric, light)

For query by examples, the content-based similarity λ between a query q and an image x can be computed in terms of the similarities between their corresponding local regions. For example, the similarity based on L_1 distance measure (city block distance) between query q with m local regions Y_j and image x with m local regions Z_j is defined as

$$\lambda(q, x) = 1 - \frac{1}{2m}\sum_j \sum_i |T_i(Y_j) - T_i(Z_j)| \quad (4)$$

2.2 Class-Relative Indexing

Beyond the local semantics of an image, it is also important to capture the categorical context of a query. As a means to probe the relevant class of images of a query, we can define prior semantic image categories as prototypical instances of the relevance class and use categorical memberships in a similarity measure. That is, images are compared to a query image based on their relative memberships to prototypical image classes learned a priori. This is related to but different

from class-based retrieval where images are ranked according to their class memberships to a query class [17, 16, 15].

Research on image categorization has received more attention lately [5, 18, 28, 30]. In particular, the efforts to classify photos based on contents have been devoted to: indoor versus outdoor [5, 28], natural versus man-made [5, 30], and categories of natural scenes [18, 30]. In general, the classifications were made based on low-level features such as color, edge directions etc and [30] presented the most comprehensive coverage of the problem by dealing with a hierarchy of 8 categories (plus 3 "others") though the vacation photos used in their experiments are a mixture of Corel photos, personal photos, video key frames, and photos from the web. In our case, image classification is not the end but a means to compute categorical similarity so as to provide contextual estimation of the relevance class of a query.

As our test images are consumer photos, we designed a taxonomy for consumer photos as shown in Fig. 5. This hierarchy of categories is more comprehensive than that addressed in [30]. We trained support vector classifiers on the 7 disjoint categories represented by the leaf nodes (except the `miscellaneous` category) in Fig. 5. A support vector classifier $\mathcal{C}_k, k = 1, \cdots, 7$ is trained to differentiate each category from other categories. Using the softmax function, the output of classification C_k given an image x is computed as,

$$R_k(x) = \frac{\exp^{\mathcal{C}_k(x)}}{\sum_j \exp^{\mathcal{C}_j(x)}}. \tag{5}$$

For each class, a human subject was asked to define the list of ground truth images from the 2400 collection and 20% of the lists was used for training. To ensure unbiased training samples, we generated 10 different sets of positive training samples from the ground truth list for each class based on uniform random distribution. The negative training (test) examples for a class are the union of positive training (test) examples of the other 6 classes and the `miscellaneous` class. The classifier training for each class was carried out 10 times on these different training sets and the support vector classifier of the best run was retained.

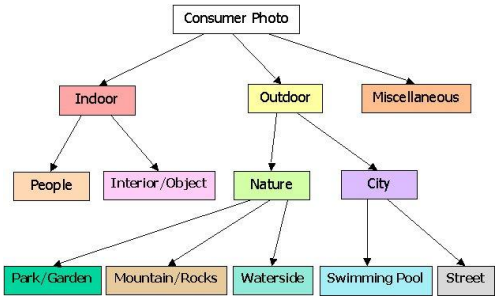

Figure 5: Taxonomy for consumer photos

The feature vector for classification is the detection-based image index as described above. To be consistent with the SSR training, we adopted the polynomial kernels with degree 2 and constant 1 ($C = 100$ [9]) and the following modified dot product similarity measure between image indexes $u = T_i(Y_j)$ and $v = T_i(Z_j)$ is computed as

$$u \cdot v = \frac{1}{m} \sum_j \frac{\sum_i T_i(Y_j) T_i(Z_j)}{\sqrt{\sum_k T_k(Y_j)^2} \sqrt{\sum_k T_k(Z_j)^2}} \tag{6}$$

The category-based similarity μ between a query q and an image x is computed as

$$\mu(q, x) = 1 - \frac{1}{2} \sum_k |R_k(q) - R_k(x)| \tag{7}$$

3 Semantics Discovery Approach

To address the issue of minimal supervision, we propose a framework to discover the local semantics that distinguish image classes and use these Discovered Semantic Regions (DSR) to span a semantic space for image indexing. Fig. 6 depicts the steps in the framework which can be divided into three learning phases as described below.

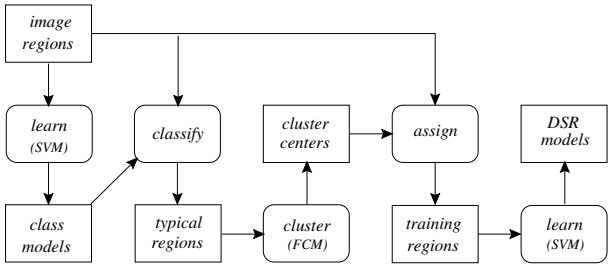

Figure 6: A schematic digram of local semantics discovery

First support vector machines (SVM) are trained on local image blocks from a small number of images labeled as several semantic categories. Then to bootstrap the local semantics, *typical* image blocks that produce high SVM outputs are grouped into Discovered Semantic Regions (DSR) using fuzzy c-means clustering. The training samples for these DSR are automatically induced from cluster memberships and subject to local support vector machine learning to form local semantic detectors for DSR. An image is indexed as a tessellation of DSR histograms (similar to the SSR indexing shown in Fig. 3) and matched using histogram intersection.

At the same time, the support vector image classifiers trained on local image blocks (i.e. Local Support Classes (LSC)) can also be used to form detection-based image indexes in terms of local class patterns. During retrieval, similarities based on DSR and LSC indexes are combined to rank images. Query-by-example experiments on 2400 heterogeneous consumer photos with 16 semantic queries show that the combined matching aproach is better than matching with single index. The proposed approach also achieved a very significant improvement of 37% in average precision than using combined matching of color and texture features.

3.1 Local Semantics Discovery

Given a content or application domain, some distinctive classes C_k with their image samples are identified. For consumer images used in our experiments, a taxonomy as shown in Fig. 5 is adopted. This hierarchy of 11 categories is more comprehensive than the 8 categories (plus 3 "others") addressed in [30]. We select the 7 disjoint categories represented by the leaf nodes (except the `miscellaneous` category) in Fig. 5 and their samples to train 7 binary SVM. The training samples are tessellated image blocks z from the class samples. After learning, the class models would have captured the local class semantics and a high SVM output (i.e. $\mathcal{C}_k(z) \gg 0$) would suggest that the local region z is typical to the semantics of class k.

The same color and texture features as well as the modified dot product similarity measure used in the supervised learning framework (Eq. (2)) are adopted for the support vector classifier training with polynomial kernels. With the help of the learned class models C_k, we can generate sets of local image regions that characterize the class semantics (which in turn captures the semantic of the content domain) $\mathcal{X}_k$ as

$$\mathcal{X}_k = \{z|\mathcal{C}_k(z) > \rho\} \quad (\rho \geq 0) \tag{8}$$

However, the local semantics hidden in each $\mathcal{X}_k$ is opague and possibly multi-mode. We would like to discover the multiple groupings in each class by unsupervised learning such as Gaussian mixture modeling and fuzzy c-means clustering. The result of the clustering is a collection of partitions m_{kj}, $j = 1, 2, \cdots, N_k$ in the space of local semantics for each class, where m_{kj} are usually represented as cluster centers and N_k are the numbers of partitions for each class.

Once we have obtained the typical semantic partitions for each class, we can learn the models of Discovered Semantic Regions (DSR) S_i $i = 1, 2, \cdots, N$ where $N = \sum_k N_k$ (i.e. we linearize the ordering of m_{kj} as m_i). We label a local image block ($x \in \cup_k \mathcal{X}_k$) as positive example for S_i if it is closest to m_i and as negative example for S_j $j \neq i$,

$$X_i^+ = \{x|i = \arg\min_t |x - m_t|\} \tag{9}$$

$$X_i^- = \{x|i \neq \arg\min_t |x - m_t|\} \tag{10}$$

where $|.|$ is some distance measure. Now we can perform supervised learning again on X_i^+ and X_i^- using say support vector machines $\mathcal{S}_i(x)$ as DSR models.

To visualize a DSR S_i, we can display the image block s_i that is most typical among those assigned to cluster m_i that belonged to class k,

$$\mathcal{C}_k(s_i) = \max_{x \in X_i^+} \mathcal{C}_k(x) \tag{11}$$

In our experiments, we trained 7 SVMs with polynomial kernels (degree 2, constant 1, $C = 100$ [9]) for the leaf-node categories (except `miscellaneous`) on color and texture features (Eq. (2)) of 60×60 image blocks (tessellated with 20 pixels in both directions) from 105 sample images. Hence each SVM was trained on 16,800 image blocks. After training, the samples from each class k is fed into classifier C_k to test their typicalities. Those samples with SVM output $\mathcal{C}_k(z) > 2$ (Eq. (8)) are subject to fuzzy c-means clustering. The number of clusters assigned to each class is roughly proportional to the number of training images in each class. We have 26 DSR in total.

To build the DSR models, we trained 26 binary SVM with polynomial kernels (degree 2, constant 1, $C = 100$ [9]), each on 7467 positive and negative examples (Eq. (9) and (10)). To visualize the 26 DSR that have been learned, we compute the most typical image block for each cluster (Eq. (11)) and concatenate their appearances in Fig. 7.

As mentioned, detection-based image indexing is carried out based on the steps as explained in the previous section (i.e. as in Fig. 3 with SSR replaced by DSR).

For query by examples, the intra-image similarity λ between a query q and an image x can be computed in terms of the similarities between their corresponding local regions as given in Eq. (4).

Figure 7: Most typical image blocks of the DSR learned (left to right): china utensils and cupboard top (first four) for the `inob` class; faces with different background and body close-up (next five) for the `inpp` class; rocky textures (next two) for the `mtrk` class; green foliage and flowers (next four) for the `park` class; pool side and water (next two) for the `pool` class; roof top, building structures, and roadside (next five) for the `strt` class; and beach, river, pond, far mountain (next four) for the `wtsd` class.

3.2 Local Class Patterns

The classifiers C_k trained on local image blocks in order to derive DSR can also be used to form image indexes based on local class patterns. In [28], classification decisions on image blocks have been used as binary patterns for indoor and outdoor image classification. Our aim here is not image classification but image indexes based on local class patterns. Moreover, we preserve the soft classification decision vectors and allow fine-grain tessellated blocks. That is, detection-based image indexing is carried out as in Fig. 3 with SSR replaced by Local Support Classes C_k. The inter-class similarity μ between a query q and an image x is computed as given in Eq. (7).

4 Integrated Similarity Matching

We believe that both the intra-image content-based similarity and inter-class context-based similarity are important and complementary. They can be combined into a single similarity for ranking images relevant to a query example. A simple linear combination ($\omega \in [0, 1]$) is

$$\rho(q, x) = \omega \cdot \lambda(q, x) + (1 - \omega) \cdot \mu(q, x) \tag{12}$$

When a query has multiple examples, $q = \{q_1, q_2, \cdots, q_K\}$, the similarity $\rho(q, x)$ for any database image is computed as

$$\rho(q, x) = max_i \rho(q_i, x) \tag{13}$$

5 Empirical Evaluation

In this paper, we evaluate our proposed image indexing approach on 2400 genuine consumer photos, taken over 5 years in several countries with both indoor and outdoor settings. The images are those of the smallest resolution (i.e. 256×384) from Kodak PhotoCDs, in both portrait and landscape layouts. After removing possibly noisy marginal pixels, the images are of size 240×360. The indexing process automatically detects the layout and applies the corresponding tessellation template. Fig. 8 displays typical photos in this collection. Photos of bad quality (e.g. faded, over-exposed, blurred, dark etc) (not shown here) are retained in order to reflect the complexity of the original data.

We defined 16 semantic queries and their ground truths (G.T.) among the 2400 photos (Table 1). In fact, Fig. 8 shows, in top-down left-to-right order, 2 relevant images for queries Q01-Q16 respectively. As these unconstrained consumer images have highly varied and complex contents, we represent each query with 3 relevant photos as query examples in our experiments. The precisions and recalls were computed

Figure 8: Sample consumer photos from the 2400 collection. They also represent 2 relevant images (top-down, left-right) for each of the 16 queries used in our experiments.

Table 1: Semantic queries used in QBE experiments

Query	Description	G.T.
Q01	indoor	994
Q02	outdoor	1218
Q03	people close-up	277
Q04	people indoor	840
Q05	interior or object	134
Q06	city scene	697
Q07	nature scene	521
Q08	at a swimming pool	52
Q09	street or roadside	645
Q10	along waterside	150
Q11	in a park or garden	304
Q12	at mountain area	67
Q13	buildings close-up	239
Q14	close up, indoor	73
Q15	small group, indoor	491
Q16	large group, indoor	45

without the query images themselves in the lists of retrieved images.

We compare our proposed cascading approaches ("Dsgn" for the semantics design framework and "Dscv" for the semantics discovery framework) with the feature-based approach that combines color and texture in a linearly optimal way (denoted as "CTO"). We have not compared with region-based approach as we believe that current image segmentation algorithms will not be robust enough for unconstrained consumer images such as those used in our experiments. All indexing are carried out with a 4×4 grid on the images. For the color-based signature, local color histograms of b^3 ($b = 4$ to 17) number of bins in the RGB color space were computed and compared using histogram intersection [27]. For the texture-based signature, we adopted the means and standard deviations of Gabor coeffients and the associated distance measure as reported in [19]. The Gabor coefficients were computed with 5 scales and 6 orientations. Convolution windows of 20×20 to 60×60 were attempted. The distance measures between a query and an image for the color and texture methods were normalized within $[0, 1]$ and combined linearly similar to Eq. (12). Among the relative weights attempted at 0.1 intervals, the best overall average precision of 0.38 was obtained with a dominant influence of 0.9 from the color feature (2197 bins) and 0.1 influence from the texture feature (20×20 windows).

While Table 2 compares the average precisions generated by different methods for each query, Tables 4 and 3show the average precisions among the top $20, 30, 50$ and 100 retrieved images as well as the overall average precisions for the methods compared.

In a nutshell, our proposed approaches Dsgn and Dscv achieved high average precisions of 0.59 and 0.52 respectively, which are significant improvements of 55% and 37% over that of the CTO method (last row of Table 4). Indeed both Dsgn and Dscv outperformed CTO in all except two queries (Q12 and Q16 for Dsgn; Q11 and Q16 for Dscv) in average precisions as seen in Table 2. The random retrieval method (i.e. $G.T./2400$) (denoted as "RAND") was used as a baseline comparison. The integrated matching (Eq. (12)) has also shown to be effective in combining the complementary indexes to produce better average precisions (Table 3).

Table 2: Average precisions for the 16 queries (left to right): query id, average precisions based on random retrieval (RAND), linear fusion of color and texture features (CTO), semantics design approach (Dsgn), and semantics discovery approach (Dscv).

Query	RAND	CTO	Dsgn	Dscv
Q01	0.41	0.62	0.91	0.86
Q02	0.51	0.78	0.91	0.79
Q03	0.12	0.16	0.36	0.37
Q04	0.35	0.59	0.90	0.83
Q05	0.06	0.18	0.43	0.36
Q06	0.29	0.49	0.79	0.67
Q07	0.22	0.35	0.80	0.52
Q08	0.02	0.18	0.57	0.59
Q09	0.27	0.50	0.81	0.65
Q10	0.06	0.17	0.37	0.34
Q11	0.13	0.71	0.81	0.62
Q12	0.03	0.28	0.24	0.39
Q13	0.10	0.35	0.40	0.37
Q14	0.03	0.15	0.31	0.31
Q15	0.20	0.32	0.56	0.46
Q16	0.02	0.29	0.26	0.20

From a practical point of view (c.f. Table 4), a user is able to locate at least 25% more relevant images retrieved at first 1 to 3 pages of image thumbnails displayed on a computer screen. This is especially crucial when the client terminal is a mobile device such as PDA and cellphone with limited display area (say 4 to 6 thumbnails per screen). Our approach can sustain a high precision value that shows many relevant photos in the first few pages before the user loses his or her patience. Retrieval is also very efficient as similarity matching involves simple arithmetic operations only. Learning and indexing require more computation but they are carried out off-line and the algorithms are inherently parallel, hence concurrent and parallel implementation are straight forward.

Table 3: Average precisions at top numbers of retrieved images (left to right): numbers of retrieved images, average precisions based on Semantic Support Regions (SSR), Semantic Support Classes (SSC), and their integration (Dsgn) as well as Discovered Semantic Regions (DSR), Local Support Classes (LSC), and their combination (Dscv). The last row shows precisions averaged over all 16 queries.

Avg.Prec.	SSR	SSC	Dsgn	DSR	LSC	Dscv
At 20	0.76	0.71	0.84	0.71	0.70	0.80
At 30	0.70	0.68	0.78	0.68	0.69	0.76
At 50	0.62	0.64	0.72	0.63	0.63	0.70
At 100	0.54	0.58	0.65	0.57	0.58	0.62
Overall	0.45	0.53	0.59	0.48	0.48	0.52

Table 4: Average precisions at top numbers of retrieved images (left to right): numbers of retrieved images, average precisions based on linear fusion of color and texture features (CTO), semantics design approach (Dsgn), and semantics discovery approach (Dscv). The numbers in brackets show the improvement in percentage over CTO. The last row shows precisions averaged over all 16 queries and improvement.

Avg.Prec.	CT	Dsgn (%)	Dscv (%)
At 20	0.64	0.84 (31)	0.80 (25)
At 30	0.59	0.78 (32)	0.76 (29)
At 50	0.52	0.72 (38)	0.70 (35)
At 100	0.46	0.65 (41)	0.62 (35)
Overall	0.38	0.59 (55)	0.52 (37)

6 Conclusion

In this paper, we have presented dual cascading frameworks for combining intra-content and inter-class semantics from both semantics design and discovery perspectives. More specifically, our contributions can be listed as follows.

- The supervised framework provides a structured design methodology to build semantic image indexing and retrieval systems with prototypical categories providing contextual estimation to the relevant class.
- The unsupervised framework provides an automatic approach to discover local semantics bootstrapped from semantic image classes for image indexing and retrieval.
- The proposed intra-image and inter-class image indexes are based on segmentation-free detection of SSR and DSR as well as class memberships of SSC and LSC that are all derived from statistical learning in a modular manner.
- The integrated similarity matching of both content-based and category-based semantic indexes has been shown to more effective than the individual indexes in retrieval precisions.
- A comprehensive empirical evaluation has been carried out using 16 semantic queries on 2400 real unconstrained consumer images to verify the usefulness of the proposed framework against a typical feature-fusion approach.

References

[1] K. Barnard and D. Forsyth. Learning the semantics of words and pictures. In *Proc. of ICCV*, pp. 408-415, 2001.

[2] K. Barnard et al. The effects of segmentation of feature choices in a translation model of object recognition. In *Proc. of CVPR*, 2003.

[3] K. Barnard et al. Matching words and pictures. *J. Machine Learning Research*, 3: 1107-1135, 2003.

[4] C.M. Bishop. *Neural Networks for Pattern Recognition*. Clarendon Press, Oxford, 1995.

[5] B. Bradshaw. Semantic based image retrieval: a probabilistic approach. In *Proc. of ACM Multimedia*, pp. 167-176, 2000.

[6] C.E. Brodley et al. Content-based retrieval from medical image databases: A synergy of human interaction, machine learning and computer vision. In *Proc. of AAAI*, pp. 760-767, 1999.

[7] P. Duygulu et al. Object recognition as machine translation: learning a lexicon for a fixed image vocabulary. In *Proc. of ECCV'2002*, pp.IV: 97-112, 2002.

[8] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *Proc. of IEEE CVPR*, 2003.

[9] T. Joachims. Making large-scale SVM learning practical. *Advances in Kernel Methods - Support Vector Learning*. B. Scholkopf, C. Burges, and A. Smola (ed.). MIT-Press, 1999.

[10] A. Kutics et al. Linking images and keywords for semantics-based image retrieval. In *Proc. of ICME*, pp. 777-780, 2003.

[11] J. Li and J.Z. Wang. Automatic linguistic indexing of pictures by a statistical modeling approach. *IEEE Trans. on PAMI*, 25(10): 1-14, 2003.

[12] J.H. Lim. Explicit query formulation with visual keywords. In *Proc. of ACM Multimedia2000*, pp. 407-409, 2000.

[13] J.H. Lim. Building visual vocabulary for image indexation and query formulation. *Pattern Analysis and Applications* (Special Issue on Image Indexation), 4(2/3): 125-139, 2001.

[14] J.H. Lim and J.S. Jin. Home photo indexing using learned visual keywords. In *Proc. Pan-Sydney Workshop on Visual Information Processing (VIP2002), Adelaide, Australia*. Conferences in Research and Practice in Information Technology, 22. Jin, J. S., Eades, P., Feng, D. D. and Yan, H., Eds., ACS, pp. 69-74.

[15] J.H. Lim, Q. Tian, and P. Mulhem. Home photo content modeling for personalized event-based retrival. *IEEE Multimedia*, 10(4): 28-37.

[16] J.H. Lim and J.S. Jin. Support regions and images for photo event retrieval. *Proc. of IEEE ICIP'2003*, II pp. 515-518, 2003.

[17] J.H. Lim and J.S. Jin. Learning consumer photo categories for semantic retrieval *Proc. of IJCAI'2003*, pp. 1413-1414, 2003.

[18] P. Lipson, E. Grimson, and P. Sinha. Configuration base scene classification and image indexing. In *Proc. of CVPR'97*, pp. 1007-1013, 1997.

[19] B.S. Manjunath and W.Y. Ma. Texture features for browsing and retrieval of image data. *IEEE Trans. on PAMI*, 18(8): 837-842, 1996.

[20] A. Mojsilovic and J. Gomes. Semantic based categorization, browsing and retrieval in medical image databases. In *Proc. of IEEE ICIP*, 2002.

[21] M.R. Naphade, I.V. Kozintsev, and T.S. Huang. A factor graph framework for semantic video indexing. *IEEE Trans. on CSVT*, 12(1): 40-52, 2002.

[22] M.R. Naphade et al. A framework for moderate vocabulary semantic visual concept detection. In *Proc. IEEE ICME*, pp. 437-440, 2003.

[23] P.C. Papageorgiou, M. Oren, and T. Poggio. A general framework for object detection. In *Proc. ICCV*, pp. 555-562, 1997.

[24] C. Schmid. Constructing models for content-based image retrieval. In *Proc. of CVPR*, pp. 39-45, 2001.

[25] A. Selinger and R.C. Nelson. Minimally supervised acquisition of 3D recognition models from cluttered images. In *Proc. of CVPR*, pp. 213-220, 2001.

[26] A.W.M. Smeulders et al. Content-based image retrieval at the end of the early years. *IEEE Trans. on PAMI*, 22(12): 1349-1380. 2000.

[27] M.J. Swain and D.N. Ballard. Color indexing. *Intl. J. Computer Vision*, 7(1): 11-32, 1991.

[28] M. Szummer and R.W. Picard. Indoor-outdoor image classification. In *Proc. of IEEE Int. Work. on Content-based Access of Image and Video Databases*, pp.42-51, Jan. 1998.

[29] C. Town and D. Sinclair. Content-based image retrieval using semantic visual categories. *Technical Report 2000.14*. AT&T Research Cambridge. 2000.

[30] A. Vailaya et al. Bayesian framework for hierarchical semantic classification of vacation images. *IEEE Trans. on Image Processing*, 10(1): 117-130, 2001.

[31] L. Wang, K.L. Chan, and Z. Zhang. Bootstrapping SVM active learning by incorporating unlabelled images for image retrieval. In *Proc. of IEEE CVPR*, 2003.

[32] M. Weber, M. Welling, and P. Perona. Unsupervised learning of models for recognition. In *Proc. of ECCV*, pp. 18-32, 2000.

[33] Y. Wu, Q. Tian, and T.S. Huang. Discriminant-EM algorithm with application to image retrieval. In *Proc. of CVPR'2000*, pp. 1222-1227. 2000.