Hindawi Publishing Corporation
EURASIP Journal on Applied Signal Processing
Volume 2006, Article ID 76093, Pages 1–11
DOI 10.1155/ASP/2006/76093

Discovering Recurrent Image Semantics from Class Discrimination

Joo-Hwee Lim[1] and Jesse S. Jin[2]

[1] *Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613*
[2] *School of Design, Communication and Information Technology, Faculty of Science and Information Technology, University of Newcastle, Callaghan, NSW 2308, Australia*

Received 17 August 2004; Revised 1 March 2005; Accepted 5 April 2005

Supervised statistical learning has become a critical means to design and learn visual concepts (e.g., faces, foliage, buildings, etc.) in content-based indexing systems. The drawback of this approach is the need of manual labeling of regions. While several automatic image annotation methods proposed recently are very promising, they usually rely on the availability and analysis of associated text descriptions. In this paper, we propose a hybrid learning framework to discover local semantic regions and generate their samples for training of local detectors with minimal human intervention. A multiscale segmentation-free framework is proposed to embed the soft presence of discovered semantic regions and local class patterns in an image independently for indexing and matching. Based on 2400 heterogeneous consumer images with 16 semantic queries, both similarity matching based on individual index and integrated similarity matching have outperformed a feature fusion approach by 26% and 37% in average precisions, respectively.

1. INTRODUCTION

Using supervised pattern classifiers to learn image semantics and ensemble of pattern classifiers to enhance system performance have become an active trend in content-based analysis research [1–4]. One of the most notable efforts by the IBM Research Group [4, 5] deployed numerous SVM classifiers in multistage optimization for the learning and detection of visual concepts in the TRECVID news video corpus. While the semantics design process and the computation involved to train and validate the SVM classifiers are certainly nontrivial, they are relatively insignificant when compared to the several months of manual annotation effort for the training, validation, and test samples by the TREC participants, with the comprehensive VideoAnnEx annotation tool [6] developed by the IBM Team.

In short, supervised learning requires labeled data. Ensemble learning with multiple classifiers demands more data for feature and classifier selection. In particular, probabilistic generative models usually require more data than discriminative models to estimate parameters reliably [7]. Hence, the bottleneck for a supervised learning approach to multimedia semantic analysis is the manual effort of data labeling.

On the other hand, supervised learning of multimedia semantics is primarily design-oriented. The designers must possess knowledge about the content domain (e.g., sports, news, medical, etc.) in order to design the ontology and relevant features and classifiers for the domain before data annotation can take place. While this design framework is useful for many applications, there are situations (e.g., images from planet Mars, unmanned robots and vehicles in unexplored areas, unexpected behaviors in open surveillance applications) whereby limited prior knowledge is available about the multimedia data source and a complete design approach is infeasible or ineffective.

Hence, an alternative semantics discovery approach is desired, for alleviating the manual annotation effort and for dealing with exploratory content domains. In this paper, we focus on image semantics discovery and use image indexing and retrieval for evaluation. The framework proposed can be extended to other modality in future.

We define the problem of image semantics discovery (ISD) (Figure 1) as follows. Given a number of classes of images, the task is to discover the local semantic regions (e.g., faces and foliage in bounding boxes as shown in Figure 1) that are recurrent within each class and discriminative against other classes. The recurrent visual patterns depend on a given image collection. For instance, while greenery image regions are recurrent in foliage images, the recurrent patterns for X-ray lung images would be very different. However, the technique must be transferable (i.e., generic)

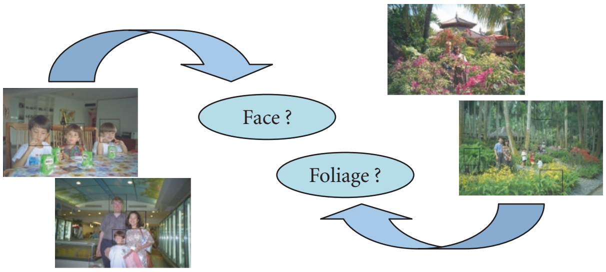


Figure 1: The problem of image semantics discovery.

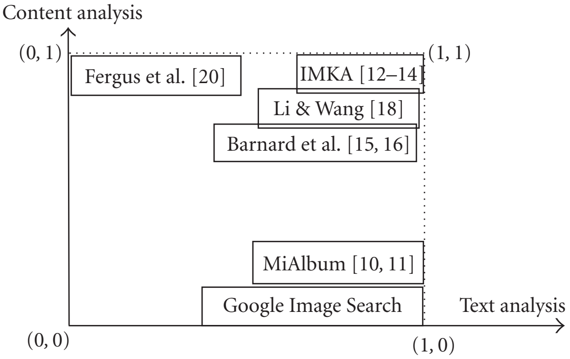


Figure 2: Automatic image annotation approaches.

for various image domains to discover different recurrent visual patterns. Note that the only prior knowledge we have here is the prior groupings of the image collection, that is, some form of global knowledge about the images. The emphasis here is on local image semantics discovery based on global image grouping information.

The problem of ISD is a relatively new one. However, we can position ISD in the context of automatic image annotation (AIA) and review existing works related to AIA. In general, the several AIA approaches discussed here can be placed on a two-dimensional grid (Figure 2). The x-axis denotes the extent of the exploitation of text information associated with the images (if they are available) and the y-axis indicates the extent of content-based analysis on the images. Note that manual effort is required at some point in time to produce the associated text information though the text might be generated for other purpose and is treated as free information source to aid image annotation.

On the x-axis of Figure 2, the coordinate $(1, 0)$ represents an AIA approach that index an image based on the text that describes a given image (e.g., filename, URL, web page, etc.) and possibly other non-content-based information (e.g., citation-based). This approach is exemplified by the Google Image Search engine on the Web (http://www.google.com/imghp). Since it does not analyze the image content, it is not surprising that the images returned by this approach may have content irrelevant to the intended query. For instance, a search with the keyword "Paris" to look for images of the French capital Paris may return portrait images of people with the name "Paris." On 25 March 2004, the 39th image returned by Google Image Search using keyword "Paris" shows a man Jon Paris plays "*Born to Be Wild*" to a crowd that understands (http://www.jsonline.com/general/harley95/images/paris.jpg).

In the context of relevance feedback, unlabeled images have also been used to boost the learning from very limited labeled examples (e.g., [8, 9]). In particular, the MiAlbum system uses relevance feedback [10] to automatically produce annotation for consumer photos [11]. The text keywords in a query are assigned to positive feedback examples (i.e., retrieved images that are considered relevant by the user who issues the query). This would require constant user intervention (in the form of relevance feedback) and the keywords issued in a query might not necessarily correspond to what is considered relevant in the positive examples.

Moving upwards from the x-axis, the regions towards the $(1, 1)$ coordinate in Figure 2 cover AIA approaches that exploit both image content and text information. Several methods have emerged in the past few years.

The intelligent multimedia knowledge application (IMKA) project proposes a framework for representing and discovering knowledge from multimedia content to enhance the classification, navigation, and retrieval of multimedia [12]. The MediaNet knowledge representation unifies both perceptual and semantic concepts and relationships exemplified by media [13]. Using a collection of 3624 annotated nature and news images, perceptual and semantic knowledge are automatically discovered by integrating both the processing of images and text. Perceptual knowledge is constructed by clustering the images based on both visual and text feature descriptors, and by discovering statistical and similarity relationships between the clusters. Using WordNet and the image clusters, semantic knowledge is further constructed by disambiguating the senses of words in annotations, and by finding semantic relations between the detected senses in WordNet. More recently, interdependence among discovered concepts are used to construct Bayesian networks for probabilistic inferencing in image classification task with promising results [14].

Motivated from a machine-translation perspective, object recognition is posed as a lexicon learning problem to translate image regions to corresponding words [15]. 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 [16–18]. The lexicon learning metaphor offers a new way of looking at object recognition [15] 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. [18]). While the results for the annotation problem on entire images look promising [18], the correspondence problem of associating words with segmented image regions remains very challenging [16] as segmentation, feature selection, and shape representation are critical and nontrivial choices [19].

Without assuming the availability of associated text information (i.e., represented by the $(0, 1)$ coordinate in Figure 2), researchers in the field of computer vision have been pushing the limit of learning by developing object recognition systems from unlabeled and unsegmented images [20–22]. For the purpose of image retrieval, unsupervised models based on "generic" texture-like descriptors without explicit object semantics can also be earned from images without manual extraction of objects or features [23]. As a representative of the state of the art, sophisticated 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 [20].

In this paper, we address the issue of minimal supervision differently. We do not assume availability of text descriptions for image or image classes as in [12, 16, 18]. Neither do we know the object classes to be recognized as in [20]. A novel semisupervised framework is proposed to discover and associate local unsegmented regions with semantics and generate their samples so as to construct semantic models for content-based image retrieval, all with minimal manual intervention. The contribution of the paper is as follows:

(i) a hybrid learning framework to discover intraclass recurrent local semantics using interclass discriminative class boundaries (Section 2);
(ii) a segmentation-free multiscale detection-based method for image indexing and retrieval (Section 3);
(iii) a similarity integration scheme to combine local and global class patterns that outperforms individual indexing scheme (Section 4);
(iv) an empirical evaluation using 16 semantic queries on 2400 unconstrained consumer images shows that image indexing and matching based on the proposed discovered image semantics with or without similarity integration with index based on local class patterns have attained better average precisions than a feature-fusion approach (Section 5).

2. DISCOVERING IMAGE SEMANTICS

The proposed generic framework of image semantics discovery (ISD) consists of three learning steps:

(i) supervised learning of class discrimination;
(ii) unsupervised learning of recurrent patterns;
(iii) supervised learning of discovered semantics regions.

In this paper, support vector machines (SVMs) [24] and fuzzy c-means clustering (FCM) [25] were used for the supervised and unsupervised learning steps, respectively.

We first describe the key ideas of the ISD framework (Figure 3) as follows before presenting the technical details. We assume that a set of representative images, grouped into K distinct classes, of a content domain is available. Each image is tessellated into possibly overlapping small image blocks with features appropriate for the domain extracted. That is, each image class is now represented by the collective local image blocks of the images from the same class.

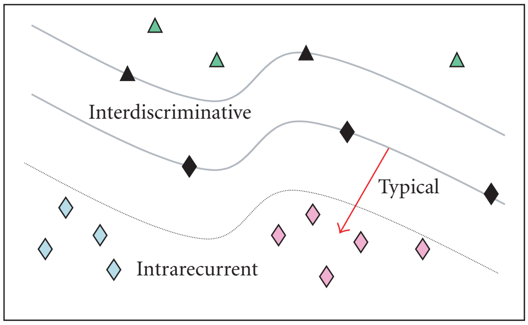


FIGURE 3: Discovering typical local patterns.

In the first supervised learning step, the class boundaries are computed based on the feature vectors of the tessellated blocks. Using binary SVMs in this paper, this step is performed K times, each time using one of the classes against all the other classes. Figure 3 depicts an example of interdiscriminative class boundaries separating two classes of local patterns, denoted as diamonds and triangles, respectively. The darken diamond and triangle shapes on the boundaries represent the support vectors derived from support vector learning [24].

While the support vectors are important parameters in the classification decision function for discrimination [24], they may not refer to local visual patterns unique to a class of images. Conversely, input patterns that result in high SVM classification outputs, denoted by diamond shapes further away from the class decision boundary, may refer to local visual patterns that are *typical* in that image class, hence capturing intraclass recurrent patterns. The second learning step in the ISD framework identifies these typical training patterns in each class by examing the SVM output for each training pattern. Unsupervised learning algorithm such as FCM is applied to these identified typical patterns in each of the K classes in turn to discover their multimode groupings shown using different colors for two groups of diamond shapes in Figure 3. The clusters of local patterns are called *discovered semantic regions* (DSRs).

The last step of the ISD framework is to generate the positive and negative training samples from the clusters in the previous unsupervised step for the learning of DSRs. In this paper, we also adopt binary SVM classifiers to model the DSRs. That is, using Figure 3 as illustration, the task is to discriminate the diamond shapes of the same color from diamond shapes of different colors and from triangle shapes. The local patterns that are nearest to the respective cluster centers can be computed and their visual appearances in the images can be extracted as a means to visualize the DSRs.

The flow of learning in the proposed ISD framework is summarized in Figure 4. Note that while each learning technique such as SVM and FCM used in the steps is not new by itself, the proposed integrated flow is novel and powerful for learning local semantics without region segmentation and without the need for local region labeling. We now describe the steps in more detail.

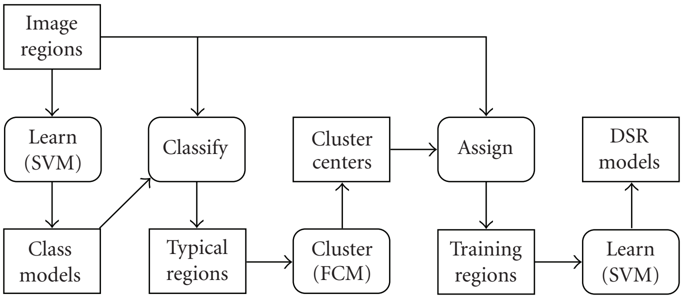


Figure 4: Flow of image semantics discovery.

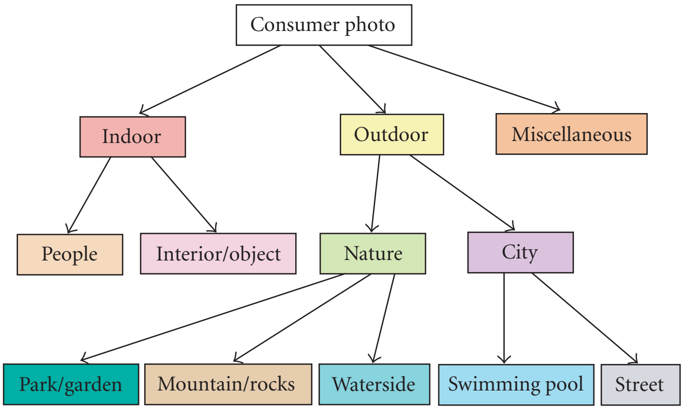


Figure 5: Proposed consumer image's taxonomy.

In this paper, we have decided to use unconstrained consumer images as the test data. We refer to consumer images as family photographs taken by average home users as opposed to professional photographs.

As a broad domain, unconstrained consumer images pose great technical challenges for content-based image retrieval research. Unlike professional images, which are well defined, carefully taken, and clearly layered, or domain-specific images such as medical images, which have a clear classification and are usually attached with semantic annotation, consumer images vary significantly 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.

Given an application domain, some typical classes C_k with their image samples are identified. For consumer images used in our experiments, a taxonomy as shown in Figure 5 has been designed. This hierarchy of 11 categories is more comprehensive than the 8 categories addressed in [26]. We trained 7 binary SVMs on the following categories (leaf nodes of Figure 5 except miscellaneous): interior or objects indoor (inob), people indoor (inpp), mountain and rocky area (mtrk), parks or gardens (park), swimming pool (pool), street scene (strt), and waterside (wtsd). 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 C_k.

In this paper, as our test data are heterogeneous consumer images, we extract color and textures features for a local image block and denote this feature vector as z. Hence, 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, as the image patch for training and detection is relatively small, the mean and standard deviation of each color channel are deemed sufficient (i.e., z^c has 6 dimensions). In our experiments, we use the YIQ color space over other color spaces (e.g., RGB, HSV, LUV) as it performed better in our experiments. For the texture feature, we adopted the Gabor coefficients which have been shown to provide excellent pattern retrieval results [27]. Similarly, the mean and standard deviation of the Gabor coefficients (5 scales and 6 orientations) in an image block are computed as z^t which has 60 dimensions. To normalize both the color and texture features, we use the Gaussian (i.e., zero-mean) normalization.

The power of SVM lies in the kernel trick: the kernel function $K(y,z)$ represents the desired notion of similarity between two data points y and z without actual computation of the inner product on the transformed feature space $\Phi(y) \cdot \Phi(z)$. The Mercer condition for the kernel ensures the convergence of the SVM algorithm towards a unique optimum because the SVM problem will be convex whenever a Mercer kernel is used. The Mercer condition requires that if and only if, for any $g(z)$ such that $\int g(y)^2 dy$ is finite, then $\int K(y,z)g(y)g(z)dy\,dz \geq 0$.

Research on different kernel functions in different application areas is very active, especially on defining new kernel functions to deal with nontraditional data such as strings, sets, trees, and so forth. However, defining or proving a kernel that satisfies the Mercer condition is not an easy task (in fact, the popular sigmoid kernel does not satisfy the Mercer condition on all parameter values). This difficulty has not stopped researchers from experimenting with non-Mercer kernels with practical values (e.g., [28]).

For the experimental results reported in this paper, we have adopted the polynomial kernels with a hybrid cosine similarity measure that balances the influence from both color and texture feature vectors and that performs better than traditional dot product on concatenated feature vectors,

$$K(y,z) = \left(\frac{1}{2}\left(\frac{y^c \cdot z^c}{|y^c||z^c|} + \frac{y^t \cdot z^t}{|y^t||z^t|}\right) + 1\right)^d. \qquad (1)$$

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

$$\mathcal{X}_k = \{z \mid \mathcal{C}_k(z) > \rho\} \quad (\rho \geq 0). \qquad (2)$$

However, the local semantics hidden in each $\mathcal{X}_k$ is opague and possibly multimode. 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}, $j = 1, 2, \ldots, N_k$, 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 (DSRs) S_i, $i = 1, 2, \ldots, N$, where $N = \Sigma_k N_k$ (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 the closest to m_i and as negative example for S_j, $j \neq i$,

$$\begin{aligned} X_i^+ &= \{x \mid i = \operatorname{argmin}_t |x - m_t|\}, \\ X_i^- &= \{x \mid i \neq \operatorname{argmin}_t |x - m_t|\}, \end{aligned} \qquad (3)$$

where $|\cdot|$ is some distance measure. Now we can perform supervised learning again on X_i^+ and X_i^- using SVMs $\mathcal{S}_i(x)$ 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 belongs to class C_k,

$$\mathcal{C}_k(s_i) = \max_{x \in X_i^+} \mathcal{C}_k(x). \qquad (4)$$

As mentioned, we trained the 7 SVMs with polynomial kernels (degree 2, $C = 100$ [29]) for the leaf-node categories (except miscellaneous) on color and texture features (1) of 60×60 image blocks (tessellated with 20 pixels in both directions) from 105 sample images. Hence, each SVM $\mathcal{C}_k$ was trained on 16, 800 image blocks z.

Table 1: Training statistics for ISD.

Class	Size	#trg.	#SV	#data	#clus.
inob	134	15	1905	1429	4
inpp	840	20	2249	936	5
mtrk	67	10	1090	1550	2
park	304	15	955	728	4
pool	52	10	1138	1357	2
strt	645	20	2424	735	5
wtsd	150	15	2454	732	4

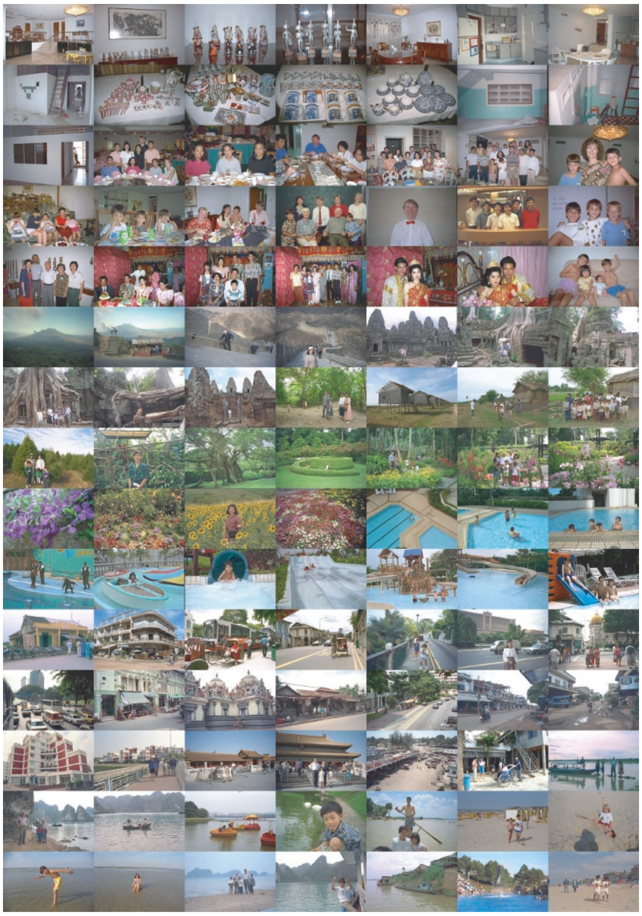


Figure 6: Training set of 105 images.

Table 1 lists the training statistics of the semantic classes C_k for bootstrapping local semantics. The columns (from left to right) list the class labels, the number of images of each class in the 2400 collection, the number of training images, the number of support vectors learned, the number of typical image blocks subject to clustering ($\mathcal{C}_k(z) > 2$), and the number of clusters assigned. The 105 training images are shown in Figure 6. Their top-down, left-to-right order (and the number of images in each class) corresponds to the classes (and #trg.) as listed in Table 1.

The number of training images is roughly proportional to the class distribution. A minimum of 10 images is considered necessary for classes of size 100 or less and an addition

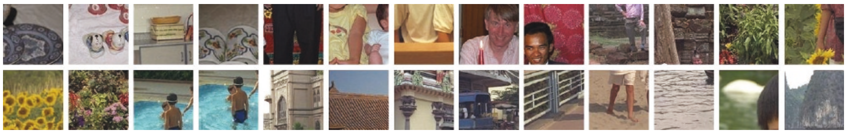


Figure 7: Most typical image blocks of the DSRs.

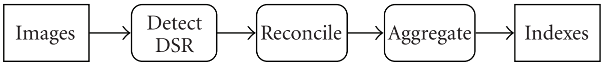


Figure 8: Flow of segmentation-free image indexing.

of 5 images is allocated for each 400 increase in class size, that is, 15 training images for class size 101 to 500 and 20 training images for class size 501 to 900. After training, the samples from each class are fed into classifier C_k to test their typicalities. Those samples with SVM output $\mathcal{C}_k(z) > 2$ (2) 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 as shown in Table 1. Hence, we have 26 DSRs in total.

To build the DSR models, we trained 26 binary SVMs with polynomial kernels (degree 2, $C = 100$ [29]), each on 7467 positive and negative examples (3) (i.e., sum of column 5 of Table 1). To visualize the 26 DSRs that have been learned, we compute the most typical image block for each cluster (4) and concatenate their appearances in Figure 7 (from 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. INDEXING AND MATCHING

Image indexing based on DSRs consists of three steps (Figure 8), namely detection, reconciliation, and aggregation. Once the SVMs $\mathcal{S}_i$ have been trained, the detection vector T of a local image block z can be computed via the softmax function [30] as

$$T_i(z) = \frac{\exp^{\mathcal{S}_i(z)}}{\Sigma_j \exp^{\mathcal{S}_j(z)}}. \tag{5}$$

To detect DSRs 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 [31]. 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 DSR detection.

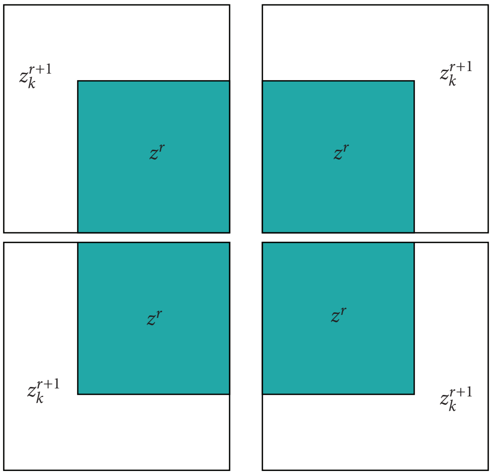


Figure 9: Reconciling multiscale SSR detection maps.

To reconcile the detection maps across different resolutions onto a common basis, we adopt the following principle. If the detection value of the most confident class of a region at resolution r is less than that of a larger region (at resolution $r+1$) that subsumes the region, then the detection vector of the region should be replaced by that of the larger region at resolution $r + 1$. For instance, if the detection of a face is more confident than that of a building at the nose region (assuming that nose is not part of DSRs), then the entire region covered by the face, which subsumes the nose region, should be labeled as face.

To illustrate the point, suppose a region at resolution r is covered by 4 larger regions at resolution $r + 1$ as shown in Figure 9. Let $\rho = \max_k \max_i T_i(z_k^{r+1})$, where k refers to one of the 4 larger regions in the case of the example shown in Figure 9. Then the principle of reconciliation says that if $\max_i T_i(z^r) < \rho$, the classification vector $T_i(z^r)$ for all i should be replaced by the classification vector $T_i(z_m^{r+1})$ for all i, where $\max_i T_i(z_m^{r+1}) = \rho$.

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 that a region Z comprises of n small equal regions with feature vectors $z_1, z_2, \ldots, z_n$, respectively. To account for the size of detected DSRs in the area Z, the DSR detection vectors of the reconciled detection map are aggregated as

$$T_i(Z) = \frac{1}{n}\Sigma_k T_i(z_k). \tag{6}$$

Note that this indexing framework is independent of the image features. As long as appropriate features can be defined and extracted for building the SVMs for an image domain, the indexing framework is applicable.

For query by examples, the content-based similarity λ between a query q and an image x can be computed in terms of the similarity 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

$$\lambda(q,x) = 1 - \frac{1}{2m}\Sigma_j\Sigma_i \,|\, T_i(Y_j) - T_i(Z_j) \,|. \tag{7}$$

This is equivalent to histogram intersection [32] except that the bins have semantic interpretation. In general, we can attach different weights to the regions (i.e., Y_j, Z_j) to emphasize the focus of attention (e.g., center). In this paper, we report experimental results based on even weights as grid tessellation is used. Also we have attempted various similarity and distance measures (e.g., cosine similarity, L_2 distance, Kullback-Leibler distance, etc.) and the city-block distance in (7) has the best performance.

4. LOCAL CLASS PATTERNS AND INTEGRATED SIMILARITY MATCHING

The classifiers $\mathcal{C}_k$ trained on local image blocks in order to derive DSRs can also be used to form image indexes based on local class patterns (LCPs). In [33], classification decisions on image blocks have been used as binary patterns for indoor and outdoor image classifications. Our aim here is not image classification but image indexes based on LCPs. Moreover, we preserve the soft classification decision vectors and allow more fine-grained tessellated blocks. That is, detection-based image indexing is carried out as in Figure 8 with DSRs $\mathcal{S}_i$ replaced by LCPs $\mathcal{C}_k$,

$$R_k(z) = \frac{\exp^{\mathcal{C}_k(z)}}{\Sigma_j \exp^{\mathcal{C}_j(z)}}. \tag{8}$$

The similarity μ between a query q with m local regions Y_j and an image x with m local regions Z_j is computed as

$$\mu(q,x) = 1 - \frac{1}{2m}\Sigma_j\Sigma_k \,|\, R_k(Y_j) - R_k(Z_j) \,|. \tag{9}$$

Both the DSR-based and LCP-based similarities can be combined into a single similarity for ranking images relevant to a query example. A simple linear combination ($\omega \in [0,1]$) is computed as

$$\rho(q,x) = \omega \cdot \lambda(q,x) + (1-\omega) \cdot \mu(q,x). \tag{10}$$

When a query has multiple examples, $Q = \{q_1, q_2, \ldots, q_K\}$, the similarity $\rho(Q,x)$ for database image x becomes

$$\rho(Q,x) = \max_i \rho(q_i, x). \tag{11}$$

This similarity-matching scheme for query by multiple examples also applies to λ and μ independently when either DSR-based or LCP-based index is used for retrieval.

5. EVALUATION ON IMAGE RETRIEVAL

In this paper, we evaluate our proposed approach on 2400 heterogeneous consumer photos from a single family. These genuine consumer photos are 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. On one hand, the small size of images allows for more efficient processing. On the other hand, it poses greater challenge for feature extraction and DSR detection.

To have a feel for the content diversity in our 2400 collections, we show 48 (2%) of them in Figure 10. For outdoor images, the content varies from natural landscape (beach, lakeside, river, pond, park, forest, garden, mountain, rocky area, etc.) to city scenes (urban area, rural area, crowded street, market, road with vehicles, swimming pool, temple, mosque, castle, etc.) from different countries and cultures (Singapore, France, China, Cambodia, Malaysia, Indonesia, etc.). The indoor images are taken with different focus (portrait of single person or a few people, groups of different sizes, people having meal, cultural performance, wedding ceremony, interior layout, display of objects like painting, toys, antique collection, etc.). In both outdoor and indoor images, the subject of focus could be people (or faces in photo frame), statues, animals, flowers, buildings (or their miniature in theme park), and so forth and their mixture with occlusion, taken with different posture, during day or night, from different viewpoints, and at different distances. Figure 11 illustrates some of the photos of bad quality (e.g., faded, overexposed, blurred, dark, etc.). We did not remove these bad quality photos from our test collection in order to reflect the complexity of the original data.

We defined 16 semantic queries and their ground truths (GT) among the 2400 images (Table 2). In fact, Figure 10 shows, in top-down left-to-right order, 3 relevant images for queries Q01–Q16, respectively. As these images have highly varied and complex contents, we represent each query with 3 relevant images as examples in query-by-examples (QBEs) experiments. The precisions and recalls were computed without the query images themselves in the lists of retrieved images.

We compare our indexing and matching schemes based on λ, μ, and ρ (denoted as "DSR," "LCP," and "Dscv," resp.) with the feature-based approach that combines color and texture in a linearly optimal way (denoted as "CTO"). We did not compare it with region-based approach here as our initial

Figure 10: Sample consumer photos associated with queries 01 to 16.

Figure 11: Some consumer photos of bad quality.

Table 2: Semantic queries used in QBE experiments.

Query	Description	GT
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

Table 3: Average precisions at top retrieved images.

Avg. prec.	CTO	DSR	LCP	Dscv
At 20	0.64	0.71	0.70	0.80
At 30	0.59	0.68	0.69	0.76
At 50	0.52	0.63	0.63	0.70
At 100	0.46	0.57	0.58	0.62
Overall	*0.38*	*0.48*	*0.48*	*0.52*

experiments with image segmenatation on unconstrained consumer images are unsatisfactory. All indexing is carried out with a 4 × 4 grid on each image.

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. For the texture-based signature, we adopted the mean and standard deviation of Gabor coefficients and the associated distance measure as reported in [27]. 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 (10). 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).

Tables 3 shows the average precisions (over 16 queries) 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 approach Dscv achieved average precision (over 16 queries) of 0.52, a significant 37% improvement over that of the CTO method (last row of Table 3). In practice, 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 cellular phone with limited display area. Our approach can sustain a high precision value that shows many relevant images in the first few pages before the user loses his or her patience. Lastly, the combined approach is also better than the individual DSR and LCP indexing schemes.

6. DISCUSSION

For the current implementation of our DSR framework, there are still several issues to be addressed. We can improve the sampling of image blocks for semantic class learning by randomly selecting, say, 20% of the ground truth images in each class as positive samples (and as negative samples for all other classes) as well as by tessellating image blocks with different sizes (e.g., $20 \times 20, 30 \times 30$, etc.) and displacements (e.g., 10 pixels) to generate a more complete and denser coverage of the local semantic space. But these attempts turned out to be too ambitious for practical training.

Another doubt is the usefulness of the semantic class learning in the first place. Can we perform clustering of image blocks in each class directly (i.e., without worrying about $\mathcal{C}_k(z) > \rho$)? The result was indeed inferior (with average precision of 0.39) for the QBE experiments. Hence, the typicality criterion is important to pick up the relevant hidden local semantics for discovery.

Cluster validity is a tricky issue. We have tried fixed number of clusters (e.g., $3, 4, 5, 7$) and retained large clusters as DSRs. Alternatively, we relied on human inspection to select perceptually distinctive clusters (visualized using (4)) as DSRs. However, the current way of assigning number of clusters roughly proportional to the number of training images has produced the best performance in our experiments. In future, we would explore other ways to model DSRs (e.g., Gaussian mixture) and to determine the value of ρ. We would also like to verify our approach on other content domains such as art images, medical images, and so forth to see if the DSRs make sense to the domain experts.

Although our attempt to alleviate the supervised learning requirement of labeled images and regions differs from the current trends of unsupervised object recognition and matching words with pictures, the methods do share some common techniques. For instance, similar to those of Schmid [23] and Fergus et al. [20], our approach computes local region features based on tessellation instead of segmentation though [20] used an interest detector and kept the number of features below 30 for practical implementation. While Schmid focused on "Gabor-like" features [23] and Fergus et al. worked on monochrome information only [20], we have incorporated both color and texture information. As the clusters in [23] were generated by unsupervised learning only, they may not correspond to well-perceived semantics when compared to our DSRs. As we are dealing with cluttered and heterogeneous scenes, we did not model object parts as in the comprehensive case of [20]. On the other hand, we handle scale invariance with multiscale detection and reconciliation of DSRs during image indexing. Last but not least, while the generative and probabilistic approaches [18, 20] may enjoy modularity and scalability in learning, they do not exploit interclass discrimination to compute features unique to classes as in our case.

For image retrieval task, the image signatures based on DSRs and LCPs realize semantic abstraction via prior learning and detection of visual classes when compared to direct indexing based on low-level features. The compact representation that accommodates imperfection and uncertainty in detection also resulted in better performance than the fusion of very high dimension of color and texture features in our QBE experiments. Hence, we feel that the computational resources devoted to prior learning of local patterns and their detection during indexing are good trade-off for concise semantic representation and effective retrieval performance. Moreover, the small footprint of the signatures has an added advantage in storage space and retrieval efficiency.

7. CONCLUSION

In this paper, a hybrid learning framework that only requires small image set with class labels to discover local semantic regions is proposed. The crux of the proposed framework lies in the novel synergy of supervised and unsupervised learning techniques (i.e., SVM-FCM-SVM as illustrated in Section 2) that exploits minimum supervision information to discriminate (across classes) visual patterns that are recurrent within each class.

The algorithms operate in the space of segmentation-free local patterns rather than traditional primitive feature space. A multiscale view-based detection and indexing method based on the segmentation-free local patterns is designed to represent an image as soft presence of either discovered semantic patterns or local class patterns.

When compared to a feature-fusion approach for the indexing and retrieval of 2400 heterogeneous consumer images based on 16 semantic queries, both the proposed indexes have outperformed the feature-fusion approach by 26% in average precision. Moreover, integrated similarity matching based on both the proposed indexes has raised the average precision further to achieve 37% improvement in average precision over the feature-fusion approach. In future, we would like to solve the cluster validity issue and experiment with other application domains.

REFERENCES

[1] W. H.-M. Hsu and S.-F. Chang, "Generative, discriminative, and ensemble learning on multi-modal perceptual fusion toward news video story segmentation," in *Proceedings of IEEE International Conference on Multimedia and Expo (ICME '04)*, vol. 2, pp. 1091–1094, Taipei, Taiwan, June 2004.

[2] B. Li, K. Goh, and E. Y. Chang, "Confidence-based dynamic ensemble for image annotation and semantics discovery," in *Proceedings of 11th ACM International Conference on Multimedia (MM '03)*, pp. 195–206, Berkeley, Calif, USA, November 2003.

[3] C. G. M. Snoek, M. Worring, and A. G. Hauptmann, "Detection of TV news monologues by style analysis," in *Proceedings of IEEE International Conference on Multimedia and Expo (ICME '04)*, vol. 2, pp. 1103–1106, Taipei, Taiwan, June 2004.

[4] B. L. Tseng, C.-Y. Lin, M. R. Naphade, A. Natsev, and J. R. Smith, "Normalized classifier fusion for semantic visual concept detection," in *Proceedings of IEEE International Conference on Image Processing (ICIP '03)*, vol. 2, pp. 535–538, Barcelona, Spain, September 2003.

[5] A. Amir, G. Iyengar, C.-Y. Lin, et al., "The IBM semantic concept detection framework," 2003, http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html.

[6] C.-Y. Lin, B. L. Tseng, and J. R. Smith, "VideoAnnEx: IBM MPEG-7 annotation tool for multimedia indexing and concept learning," in *Proceedings of IEEE International Conference*

on Multimedia and Expo (ICME '03), Baltimore, Md, USA, July 2003.

[7] W. H. Adams, G. Iyengar, C.-Y. Lin, et al., "Semantic indexing of multimedia content using visual, audio, and text cues," *EURASIP Journal on Applied Signal Processing*, vol. 2003, no. 2, pp. 170–185, 2003.

[8] L. Wang, K. L. Chan, and Z. Zhang, "Bootstrapping SVM active learning by incorporating unlabelled images for image retrieval," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR '03)*, vol. 1, pp. 629–634, Madison, Wis, USA, June 2003.

[9] Y. Wu, Q. Tian, and T. S. Huang, "Discriminant-EM algorithm with application to image retrieval," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR '00)*, vol. 1, pp. 222–227, Hilton Head Island, SC, USA, June 2000.

[10] Y. L. Lu, C. Hu, X. Zhu, H. J. Zhang, and Q. Yang, "A unified framework for semantics and feature based relevance feedback in image retrieval systems," in *Proceedings of 8th ACM International Conference on Multimedia (MM '00)*, pp. 31–37, Los Angeles, Calif, USA, October–November 2000.

[11] W. Liu, Y. Sun, and H. Zhang, "MiAlbum—a system for home photo management using the semi-automatic image annotation approach," in *Proceedings of 8th ACM International Conference on Multimedia (MM '00)*, pp. 479–480, Los Angeles, Calif, USA, October–November 2000.

[12] A. B. Benitez and S.-F. Chang, "Automatic multimedia knowledge discovery, summarization and evaluation," to appear in *IEEE Trans. Multimedia*.

[13] A. B. Benitez, J. R. Smith, and S.-F. Chang, "MediaNet: a multimedia information network for knowledge representation," in *Internet Multimedia Management Systems*, vol. 4210 of *Proceedings of SPIE*, pp. 1–12, Boston, Mass, USA, November 2000.

[14] A. B. Benitez and S.-F. Chang, "Image classification using multimedia knowledge networks," in *Proceedings of IEEE International Confererence on Image Processing (ICIP '03)*, vol. 3, pp. 613–616, Barcelona, Spain, September 2003.

[15] P. Duygulu, K. Barnard, N. de Freitas, and D. Forsyth, "Object recognition as machine translation: learning a lexicon for a fixed image vocabulary," in *Proceedings of 7th European Conference on Computer Vision (ECCV '02)*, vol. 4, pp. 97–112, Copenhagen, Denmark, May 2002.

[16] K. Barnard, P. Duygulu, D. Forsyth, N. de Freitas, D. M. Blei, and M. I. Jordan, "Matching words and pictures," *Journal of Machine Learning Research*, vol. 3, no. 6, pp. 1107–1135, 2003.

[17] A. Kutics, A. Nakagawa, K. Tanaka, M. Yamada, Y. Sanbe, and S. Ohtsuka, "Linking images and keywords for semantics-based image retrieval," in *Proceedings of IEEE International Conference on Multimedia and Expo (ICME '03)*, vol. 1, pp. 777–780, Baltimore, Md, USA, July 2003.

[18] J. Li and J. Z. Wang, "Automatic linguistic indexing of pictures by a statistical modeling approach," *IEEE Transactions On Pattern Analysis And Machine Intelligence*, vol. 25, no. 9, pp. 1075–1088, 2003.

[19] K. Barnard, P. Duygulu, R. Guru, P. Gabbur, and D. Forsyth, "The effects of segmentation and feature choice in a translation model of object recognition," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR '03)*, vol. 2, pp. 675–682, Madison, Wis, USA, June 2003.

[20] R. Fergus, P. Perona, and A. Zisserman, "Object class recognition by unsupervised scale-invariant learning," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR '03)*, vol. 2, pp. 264–271, Madison, Wis, USA, June 2003.

[21] A. Selinger and R. C. Nelson, "Minimally supervised acquisition of 3D recognition models from cluttered images," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR '01)*, vol. 1, pp. 213–220, Kauai, Hawaii, USA, December 2001.

[22] M. Weber, M. Welling, and P. Perona, "Unsupervised learning of models for recognition," in *Proceedings of 6th European Conference on Computer Vision (ECCV '00)*, vol. 1, pp. 18–32, Dublin, Ireland, June–July 2000.

[23] C. Schmid, "Constructing models for content-based image retrieval," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR '01)*, vol. 2, pp. 39–45, Kauai, Hawaii, USA, December 2001.

[24] V. N. Vapnik, *Statistical Learning Theory*, John Wiley & Sons, New York, NY, USA, 1998.

[25] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum Press, New York, NY, USA, 1981.

[26] A. Vailaya, M. A. T. Figueiredo, A. K. Jain, and H.-J. Zhang, "Image classification for content-based indexing," *IEEE Transactions On Image Processing*, vol. 10, no. 1, pp. 117–130, 2001.

[27] B. S. Manjunath and W. Y. Ma, "Texture features for browsing and retrieval of image data," *IEEE Transactions On Pattern Analysis And Machine Intelligence*, vol. 18, no. 8, pp. 837–842, 1996.

[28] S. Boughorbel, J.-P. Tarel, and F. Fleuret, "Non-mercer kernel for SVM object recognition," in *Proceedings of British Machine Vision Conference (BMVC '04)*, pp. 137–146, London, UK, September 2004.

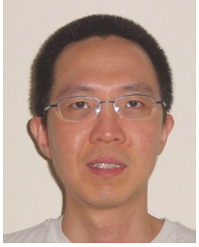
[29] T. Joachims, "Making large-scale SVM learning practical," in *Advances in Kernel Methods—Support Vector Learning*, B. Schölkopf, C. J. C. Burges, and A. Smola, Eds., pp. 169–184, MIT Press, Cambridge, Mass, USA, 1999.

[30] C. M. Bishop, *Neural Networks for Pattern Recognition*, Oxford University Press, Oxford, UK, 1995.

[31] C. P. Papageorgiou, M. Oren, and T. Poggio, "A general framework for object detection," in *Proceedings of IEEE 6th International Conference on Computer Vision (ICCV '98)*, pp. 555–562, Bombay, India, January 1998.

[32] M. J. Swain and D. H. Ballard, "Color indexing," *International Journal of Computer Vision*, vol. 7, no. 1, pp. 11–32, 1991.

[33] M. Szummer and R. W. Picard, "Indoor-outdoor image classification," in *Proceedings of IEEE International Workshop on Content-Based Access of Image and Video Databases*, pp. 42–51, Bombay, India, January 1998.



Joo-Hwee Lim received his B.S. (honors I) and M.S. (by research) degrees in computer science from the National University of Singapore and his Ph.D. degree in computer science and engineering from the University of New South Wales. He has joined the Institute for Infocomm Research, Singapore, since October 1990. He has conducted research in connectionist expert systems, neural-fuzzy systems, handwriting recognition, multiagent systems, and content-based retrieval. He was a key researcher in two international research collaborations, namely, the Real World Computing Partnership funded by METI, Japan, and the Digital Image/Video Album Project with CNRS, France, and School of Computing, National University of Singapore. He also contributed technical solutions to a few industrial projects involving pattern-based diagnostic tools for aircraft and battleship navigation systems and knowledge-based postprocessing for automatic fax/form recognition. He has published more than sixty refereed international journal and conference papers in his research areas including content-based processing, pattern recognition, and neural networks.

Jesse S. Jin graduated with a Ph.D. degree from the University of Otago, New Zealand. He worked as a Lecturer in Otago, a Lecturer, Senior Lecturer, and Associate Professor in the University of New South Wales, and an Associate Professor in the University of Sydney. He is now the Chair Professor of IT in the University of Newcastle. His areas of interest include multimedia technology, medical imaging, computer vision, and the Internet. He has published over 160 articles, and 14 books and edited books. He also has one patent and is in the process of filing 3 more patents. He has received several millions research funding from government agents (ARC, DIST, etc.), universities (UNSW, USyd, Newcastle, etc.), industry (Motorola, NewMedia, Cochlear, Silicon Graphics, Proteome Systems, etc.), and overseas organization (NZ Wool Board, UGC HK, CAS, etc.). He established a spin-off company and the company won the 1999 ATP Vice-Chancellor New Business Creation Award. He is a Consultant of many companies such as Motorola, Computer Associates, ScanWorld, Proteome Systems, HyperSoft, and so forth.