COMBINING LOCAL CLASS PATTERNS AND DISCOVERED SEMANTICS FOR IMAGE RETRIEVAL

Joo-Hwee Lim

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

Jesse S. Jin

University of New South Wales
Sydney 2052, Australia
jesse@cse.unsw.edu.au

ABSTRACT

Detecting meaningful visual entities (e.g. faces, sky, foliage, buildings etc) based on supervised pattern classifiers has become a trend in content-based image retrieval. However, a drawback of the supervised learning approach is the need for manually labeled regions as training samples. In this paper, we propose a semi-supervised framework to discover local semantic patterns and generate their samples for training with minimal human intervention. 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 approach outperformed the fusion of color and texture features significantly in average precision by 37%.

1. INTRODUCTION

In order to bridge the semantic gap [6] that exists between computed perceptual visual features and conceptual query expectation, detecting semantic objects (e.g. faces, sky, foliage, buildings etc) based on trained pattern classifiers has received serious attention (e.g. [5]). 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.

In the field of computer vision, researchers have developed object recognition systems from unlabeled and unsegmented images. 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 [2]. On the other hand, object recognition has been posed as a lexicon learning problem to translate image regions to corresponding words. 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]. While the results for the annotation problem on entire images look promising [3], the correspondence problem of associating words with segmented image regions remains very challenging [1] as segmentation, feature selection, and shape representation are critical and non-trivial choices.

We address the issue of minimal supervision differently. We do not assume availability of text descriptions for image or image classes as in [1, 3]. Neither do we know the object classes to be recognized as in [2]. We discover and associate local unsegmented regions with semantics and generate their training samples to construct models for content-based image retrieval, all with minimal manual intervention. This is realized as a hybrid framework that interleave supervised and unsupervised learnings (Fig. 1).

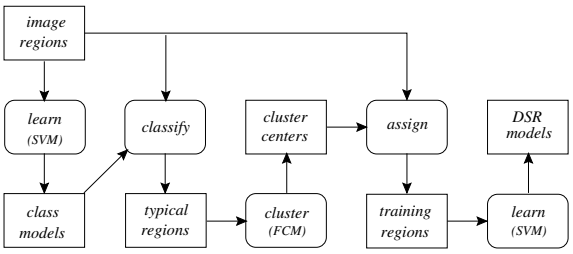

Fig. 1. A schematic digram of local semantics discovery

First support vector machines (SVMs) 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 (DSRs) using fuzzy c-means clustering. The training samples for these DSR are automatically induced from cluster member-

0-7803-8554-3/04/$20.00 ©2004 IEEE.

ships and subject to local support vector machine learning to form local semantic detectors for DSRs. 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 approach outperformed the fusion of color and texture features significantly in average precision by 37%.

2. LOCAL SEMANTICS DISCOVERY

Given an application domain, some typical classes C_k with their image samples are identified. For consumer images used in our experiments, we trained 7 binary SVMs on the following categories: 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 k.

As we are dealing with heterogeneous consumer photos, we adopt color and texture features to characterize z. 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 [4]. 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,

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

With the help of the learned class models C_k, we can generate sets of local image blocks that characterize the class semantics (which in turn captures the semantic of the content domain) $\mathcal{X}_k$ as

$$\mathcal{X}_k = \{z | \mathcal{C}_k(z) > \rho\} \quad (\rho \geq 0) \tag{2}$$

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 $j \neq i$,

$$X_i^+ = \{x | i = \arg\min_t |x - m_t|\} \tag{3}$$

$$X_i^- = \{x | i \neq \arg\min_t |x - m_t|\} \tag{4}$$

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,

$$\mathcal{C}_k(s_i) = \max_{x \in X_i^+} \mathcal{C}_k(x) \tag{5}$$

In our experiments, we trained 7 SVMs with polynomial kernels (degree 2, constant 1) on 60×60 image blocks (tessellated with 20 pixels in both $X - Y$ 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. (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. We have 26 DSR in total.

To build the DSR models, we trained 26 binary SVM with polynomial kernels (degree 2, constant 1), each on 7467 positive and negative examples (Eq. (3) and (4)). To visualize the 26 DSR that have been learned, we compute the most typical image block for each cluster (Eq. (5)) and concatenate their appearances in Fig. 2.

Fig. 2. 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. INTEGRATED INDEXING AND MATCHING

Given a local image block with feature vector z, a support vector classifier $\mathcal{S}_i$ is a detector for DSR i on z. The classification vector T for region z can be computed via the softmax function as

$$T_i(z) = \frac{\exp^{\mathcal{S}_i(z)}}{\sum_j \exp^{\mathcal{S}_j(z)}} \tag{6}$$

To detect DSR with translation and scale invariance in an image to be indexed, the image is scanned with windows of different scales. 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 DSR in the spatial area Z, the DSR detection vectors of the reconciled detection map is aggregated as

$$T_i(Z) = \frac{1}{n} \sum_k T_i(z_k). \tag{7}$$

For query by examples, the 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

$$\lambda(q, x) = 1 - \frac{1}{2m} \sum_j \sum_i |T_i(Y_j) - T_i(Z_j)| \tag{8}$$

The classifiers C_k trained on local image blocks to derive DSRs can also be used to form indexes based on local class patterns. In [7], classification decisions on image blocks were 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. That is, detection-based image indexing (including reconciliation and aggregation Eq. (7)) is carried out as described above with DSR replaced by Local Support Classes (LSC) C_k,

$$R_k(z) = \frac{\exp^{\mathcal{C}_k(z)}}{\sum_j \exp^{\mathcal{C}_j(z)}}. \tag{9}$$

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

$$\mu(q, x) = 1 - \frac{1}{2m} \sum_j \sum_k |R_k(Y_j) - R_k(Z_j)| \tag{10}$$

Both the DSR-based and LSC-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

$$\rho(q, x) = \omega \cdot \lambda(q, x) + (1 - \omega) \cdot \mu(q, x) \tag{11}$$

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

$$\rho(q, x) = max_i \rho(q_i, x) \tag{12}$$

Fig. 3. Sample consumer photos from the 2400 collection

4. EXPERIMENTAL RESULTS

We evaluate our proposed image indexing approach on 2400 genuine consumer photos. The images, in both portrait and landscape layouts, are size-normalized to 240×360. The indexing process automatically detects the layout and applies the corresponding tessellation template. Fig. 3 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. 3 shows, in top-down left-to-right order, 2 relevant images for queries Q01-Q16 respectively. As these images have highly varied and complex contents, we represent each query with 3 relevant photos as examples in our experiments. The precisions and recalls were computed without the query images themselves in the lists of retrieved images.

We compare our proposed approach ($\omega = 0.5$ in Eq. (11), denoted as "Dscv") with the feature-based approach

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

that combines color and texture in a linearly optimal way (denoted as "CTO"). We do not compared with region-based approach here as our initial experiments with image segmenatation on unconstrained consumer images are unsatisfactory. 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. For the texture-based signature, we adopted the means and standard deviations of Gabor coeffients and the associated distance measure as reported in [4]. 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. (11). 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).

Table 2. Average precisions at top retrieved images

Avg.Prec.	CTO	DSR	LSC	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

Tables 2 shows 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 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 2). 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 cellphone with limited display area. 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. Lastly the combined approach is also better than the individual DSR and LSC indexing schemes.

5. CONCLUSION

In this paper, we have presented a new method to discover local semantics from image classes. Segmentation-free indexes based on discovered semantics and local class patterns are combined to rank images. An empirical evaluation has been carried out using 16 semantic queries on 2400 unconstrained consumer images to verify the usefulness of the proposed framework against a feature-fusion approach.

6. REFERENCES

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

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

[3] 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.

[4] 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.

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

[6] 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.

[7] 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.