Semantics Discovery for Image Indexing

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

[1] Institute for Infocomm Research
21 Heng Mui Keng Terrace, Singapore 119613
joohwee@i2r.a-star.edu.sg
[2] University of New South Wales, Sydney 2052, Australia
jesse@cse.unsw.edu.au

Abstract. To bridge the gap between low-level features and high-level semantic queries in image retrieval, detecting meaningful visual entities (e.g. faces, sky, foliage, buildings etc) based on trained pattern classifiers has become an active research trend. However, a drawback of the supervised learning approach is the human effort to provide labeled regions as training samples. In this paper, we propose a new three-stage hybrid framework to discover local semantic patterns and generate their samples for training with minimal human intervention. Support vector machines (SVM) are first trained on local image blocks from a small number of images labeled as several semantic categories. Then to bootstrap the local semantics, image blocks that produce high SVM outputs are grouped into Discovered Semantic Regions (DSRs) using fuzzy c-means clustering. The training samples for these DSRs are automatically induced from cluster memberships and subject to support vector machine learning to form local semantic detectors for DSRs. An image is then indexed as a tessellation of DSR histograms and matched using histogram intersection. We evaluate our method against the linear fusion of color and texture features using 16 semantic queries on 2400 heterogeneous consumer photos. The DSR models achieved a promising 26% improvement in average precision over that of the feature fusion approach.

1 Introduction

Content-based image retrieval research has progressed from the feature-based approach (e.g. [9]) to the region-based approach (e.g. [5]). In order to bridge the semantic gap [20] that exists between computed perceptual visual features and conceptual user query expectation, detecting semantic objects (e.g. faces, sky, foliage, buildings etc) based on trained pattern classifiers has received serious attention (e.g. [15,16,22]. 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 [8,19,25]. In the context

T. Pajdla and J. Matas (Eds.): ECCV 2004, LNCS 3021, pp. 270–281, 2004.

of relevance feedback, unlabeled images have also been used to bootstrap the learning from very limited labeled examples (e.g. [24,26]). 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 [18]. 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 [1,3,11,12]. 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 [12]). While the annotation results on entire images look promising [12], 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].

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 [3, 12]. 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. 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 (DSRs) using fuzzy c-means clustering. The training samples for these DSRs are automatically induced from cluster memberships and subject to local support vector machine learning to form local semantic detectors for DSRs. An image is indexed as a tessellation of DSR histograms and matched using histogram intersection.

We evaluate our method against the linear fusion of color and texture features using 16 semantic queries on 2400 heterogeneous consumer photos with many cluttered scenes. The DSR implementation achieved a promising 26% improvement in average precision over that of the feature fusion approach.

The rest of the paper is presented as follows. We explain our local semantics discovery framework followed by the mechanisms for image indexing and matching in the next two sections respectively. Then we report and compare the results on the query-by-example experiments. Last but not least, we discuss the relevant aspects of our approach with other promising works in unsupervised semantics learning and issues for future research.

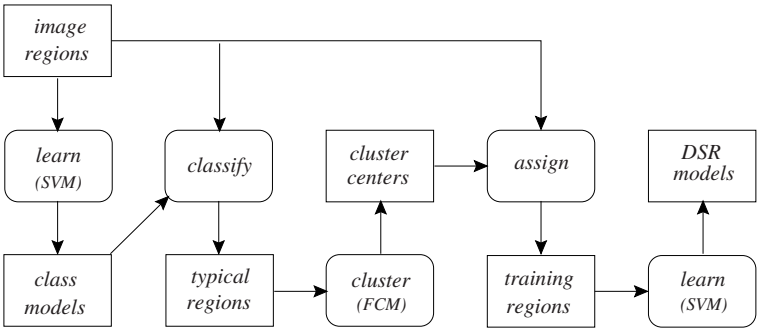

Fig. 1. A schematic digram of local semantics discovery

2 Local Semantics Discovery

Image 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 more effective and efficient way by focusing only at relevant classes and subclasses. Moreover the classes provide context for other tasks. For example, for medical images, the context could be the pathological classes for diagnostic purpose [4] or imaging modalities for visualization purpose [14]. In this paper, we propose a framework to discover the local semantics that distinguish image classes and use these Discovered Semantic Regions (DSRs) to span a semantic space for image indexing. Fig. 1 depicts the steps in the framework which can be divided into three learning phases as described below.

2.1 Learning of Local Class Semantics

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. 2 has been designed. This hierarchy of 11 categories is more comprehensive than the 8 categories addressed in [23]. We select the 7 disjoint categories represented by the leaf nodes (except the `miscellaneous` category) in Fig. 2 and their samples to train 7 binary support vector machines (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.

In this paper, as our test data are heterogeneous consumer photos, 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 is 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,

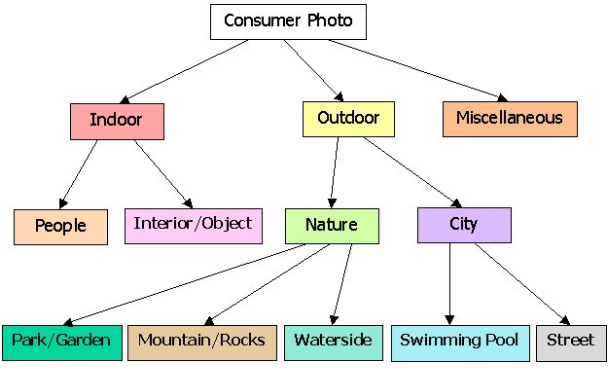

Fig. 2. Proposed hierarchy of consumer photo categories

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 [13]. Similarly, the means and standard deviations 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 distance or similarity measure depends on the kernel adopted for the support vector machines. For the experimental results reported in this paper, we adopted polynomial kernels with the following modified dot product similarity measure between feature vectors y and z,

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

2.2 Learning of Typical Semantic Partitions

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

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

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.

2.3 Learning of Discovered Semantic Regions

After obtaining the typical semantic partitions for each class, we can learn the models of DSRs S_i $i = 1, 2, \cdots, N$ where $N = \sum_k N_k$ (i.e. linearize m_{kj} subscript

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,

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

3 Image Indexing and Matching

Image indexing based on DSRs consists of three steps, namely detection, reconciliation, and aggregation. Once the support vector machines $\mathcal{S}_i$ have been trained, the detection vector T of a local image block z can be computed via the softmax function [6] as

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

As each binary SVM is regarded as an expert on a DSR, the outputs of $\mathcal{S}_i$ $\forall i$ is set to 0 if there exist some $\mathcal{S}_j, j \neq i$ has a positive output. That is, T_j is close to 1 and $T_i = 0$ $\forall i \neq j$.

To detect DSRs with translation and scale invariance in an image, the image is scanned with multi-scale windows, following the strategy in view-based object detection [17]. 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 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 $r + 1$. 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 $z_1, z_2, \cdots, z_n$ 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

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

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

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

This is equivalent to histogram intersection [21] 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 (KL) distance etc) and the simple city block distance in Equation (8) has the best performance. When a query has multiple examples, $Q = \{q_1, q_2, \cdots, q_K\}$, the similarity is computed as

$$\lambda(Q, x) = max_i \lambda(q_i, x) \tag{9}$$

4 Experimental Results

In this paper, we evaluate our DSR-based image indexing approach on 2400 genuine consumer photos, taken over 5 years in several countries with both indoor and outdoor settings. After removing possibly noisy margins, the images are size-normalized to 240×360. The indexing process automatically detects the layout and applies the corresponding tessellation template. In our experiments, the tessellation for detection of DSRs is a 4×4 grid of rectangular regions. 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.

Fig. 3. 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. Training statistics of the semantic classes C_k for bootstrapping local semantics. The columns (left to right) list the class labels, the size of ground truth, 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.

Class	G.T.	#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

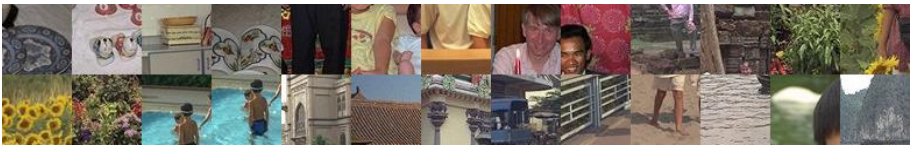

Fig. 4. Most typical image blocks of the DSRs learned (left to right): china utensils and cupboard top (first four) for the `inob` class; faces with different background and body close-up (next five) for the `inpp` class; rocky textures (next two) for the `mtrk` class; green foliage and flowers (next 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.

We trained 7 SVMs with polynomial kernels (degree 2, $C = 100$ [10]) for the leaf-node categories (except `miscellaneous`) on color and texture features (Equation (1)) 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 $\mathcal{C}_k(z) > 2$ (Equation (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. Table 1 lists training statistics for these semantic classes: `inob` (indoor interior/objects), `inpp` (indoor people), `mtrk` (mountain/rocks), `park` (park/garden), `pool` (swimming pool), `strt` (street), and `wtsd` (waterside). We have 26 DSRs in total.

To build the DSR models, we trained 26 binary SVM with polynomial kernels (degree 2, $C = 100$ [10]), each on 7467 positive and negative examples (Equations (3) and (4)) (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 (Equation (5)) and concatenate their appearances in Fig. 4. Image indexing was based on the steps as explained in Section 3.

Table 2. Results of QBE experiments for 16 semantic queries (left to right): query id, query description, size of ground truth, average precisions based on random retrieval (RAND), linear fusion of color and texture features (CTO), and discovered semantic regions (DSRs) (the indexing for the last two methods are based on 4×4 grid).

Query	Description	G.T.	RAND	CTO	DSR
Q01	indoor	994	0.41	0.62	0.79
Q02	outdoor	1218	0.51	0.78	0.78
Q03	people close-up	277	0.12	0.16	0.33
Q04	people indoor	840	0.35	0.59	0.76
Q05	interior or object	134	0.06	0.18	0.32
Q06	city scene	697	0.29	0.49	0.59
Q07	nature scene	521	0.22	0.35	0.46
Q08	at a swimming pool	52	0.02	0.18	0.62
Q09	street or roadside	645	0.27	0.50	0.53
Q10	along waterside	150	0.06	0.17	0.32
Q11	in a park or garden	304	0.13	0.71	0.51
Q12	at mountain area	67	0.03	0.28	0.31
Q13	buildings close-up	239	0.10	0.35	0.30
Q14	close up, indoor	73	0.03	0.15	0.30
Q15	small group, indoor	491	0.20	0.32	0.45
Q16	large group, indoor	45	0.02	0.29	0.29

We defined 16 semantic queries and their ground truths (G.T.) among the 2400 photos (Table 2). In fact, Fig. 3 shows, in top-down left-to-right order, 2 relevant images for queries Q01-Q16 respectively. As we can see from these sample images, the relevant images for any query considered here exhibit highly varied and complex visual appearance. There is usually no dominant homogeneous color or texture region and they pose great difficulty for image segmentation. Hence to represent each query, we selected 3 (i.e. $K = 3$ in Equation (9)) relevant photos as query examples for Query By Example (QBE) experiments since a single query image is far from satisfactory to capture the semantic of any query and single query images have indeed resulted in poor precisions and recalls in our initial experiments. The precisions and recalls were computed without the query images themselves in the lists of retrieved images.

In our experiments, we compare our local semantic discovery approach (denoted as "DSR") with the feature-based approach that combines color and texture in a linearly optimal way (denoted as "CTO"). All indexing are carried out with a 4×4 grid on the images.

For the color-based signature, color histograms of b^3 ($b = 4, 5, \cdots, 17$) number of bins in the RGB color space were computed on an image. The performance peaked at 2197 ($b = 13$) bins with average precision (over all recall points) $P_{avg} = 0.38$. Histogram intersection [21] was used to compare two color histograms. For the texture-based signature, we adopted the means and standard deviations of Gabor coeffients and the associated distance measure as reported in [13]. The Gabor coefficients were computed with 5 scales and 6 orientations. Convolution windows of $20 \times 20, 30 \times 30, \cdots, 60 \times 60$ were attempted. The best performance

Table 3. Comparison of average precisions at top numbers of retrieved images. The last row compares the precisions averaged over all 16 queries. The last column shows the relative improvement in percentage.

Avg.Prec.	CTO	DSR	%
At 20	0.64	0.71	10
At 30	0.59	0.68	15
At 50	0.52	0.63	21
At 100	0.46	0.57	24
overall	*0.38*	*0.48*	*26*

was obtained when 20×20 windows were used with $P_{avg} = 0.24$. The distance measures between a query and an image for the color and texture methods were normalized within $[0, 1]$ and combined linearly. Among the relative weights attempted at 0.1 intervals, the best fusion was obtained at $P_{avg} = 0.38$ with a dominant influence of 0.9 from the color feature.

As shown in Table 2, the DSR approach outperformed or matched the average precisions of the CTO method in all queries except Q11 and Q13. The random retrieval method (i.e. $G.T./2400$) (denoted as "RAND") was used as a baseline comparison. In particular, the DSR approach more than doubled the performance of RAND and surpassed the average precisions of CTO by at least 0.1 in more than half of the queries (Q03-08, Q10, Q14-15). Averaged over all queries, the DSR approach achieved a 26% improvement in precision over that of CTO (Table 3). As depicted in the same table, DSR is also consistently better than CTO in returning more relevant images at top numbers of images for practical applications. As an illustration, Fig. 5 and Fig. 6 show the query examples and top 18 retrieved images for query Q08 respectively. All retrieved images except image 18 are considered relevant.

Fig. 5. Query images for Q08.

Fig. 6. Top 18 retrieved images by DSR for query Q08.

5 Discussion

For the current implementation of our DSR approach, 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 (as visualized using Equation (5)) 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 etc 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 [18] and Fergus et al. [8], our approach computes local region features based on tessellation instead of segmentation though [8] used an interest detector and kept the number of features below 30 for practical implementation. While Schmid focused on "Gabor-like" features [18] and Fergus et al. worked on monochrome information only [8], we have incorporated both color and texture information. As the clusters in [18] 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 [8]. On the other hand, we handle scale invariance with multi-scale detection and reconciliation of DSRs during image indexing. Last but not least, while the generative and probabilistic approaches [8,12] may enjoy modularity and scalability in learning, they do not exploit inter-class discrimination to compute features unique to classes as in our case.

For the purpose of image retrieval, the images signatures based on DSRs 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 query-by-example experiments. Hence we feel that the computational resources devoted to prior learning of DSRs and their detection during indexing are good trade-off for concise semantic representation and effective retrieval performance. Moreover, the small footprint of DSR signatures has an added advantage in storage space and retrieval efficiency.

6 Conclusion

In this paper, we have presented a hybrid framework that interleaves supervised and unsupervised learning to discover local semantic regions without image segmentation and with minimal human effort. The discovered semantic regions serve as new semantic axes for image indexing and matching. Experimental query-by-example results on 2400 genuine consumer photos with cluttered scenes have shown that images indexes based on the discovered local semantics are more compact and effective over linear combination of color and texture features.

References

1. Barnard, K., Forsyth, D.: Learning the semantics of words and pictures. Proc. of ICCV (2001) 408–415
2. Barnard, K., et al.: The effects of segmentation of feature choices in a translation model of object recognition. Proc. of CVPR (2003)
3. Barnard, K., et al.: Matching words and pictures. J. Machine Learning Research **3** (2003) 1107–1135
4. Brodley, C.E., et al.: Content-based retrieval from medical image databases: A synergy of human interaction, machine learning and computer vision. Proc. of AAAI (1999) 760–767
5. Carson, C., et al.: Blobworld: image segmentation using expectation- maximization and its application to image querying. IEEE Trans. on PAMI **24** (2002) 1026–1038
6. Bishop, C.M.: Neural Networks for Pattern Recognition. Clarendon Press, Oxford (1995)
7. Duygulu, P., et al.: Object recognition as machine translation: learning a lexicon for a fixed image vocabulary. Proc. of ECCV (2002) 97–112
8. Fergus, R., Perona, P., Zisserman, A.: Object class recognition by unsupervised scale-invariant learning. Proc. of IEEE CVPR (2003)
9. Flickner, M., et al.: Query by image and video content: the QBIC system. IEEE Computer **28** (1995) 23–30
10. Joachims, T.: Making large-scale SVM learning practical. Advances in Kernel Methods - Support Vector Learning. B. Scholkopf, C. Burges, and A. Smola (ed.). MIT-Press (1999)
11. Kutics, A., et al.: Linking images and keywords for semantics-based image retrieval. Proc. of ICME (2003) 777–780
12. Li, J., Wang, J.Z.: Automatic linguistic indexing of pictures by a statistical modeling approach. IEEE Trans. on PAMI **25** (2003) 1–14
13. Manjunath, B.S., Ma, W.Y.: Texture features for browsing and retrieval of image data. IEEE Trans. on PAMI **18** (1996) 837–842
14. Mojsilovic, A., Gomes, J.: Semantic based categorization, browsing and retrieval in medical image databases. Proc. of IEEE ICIP (2002)

15. Naphade, M.R., Kozintsev, I.V., Huang, T.S.: A factor graph framework for semantic video indexing. IEEE Trans. on CSVT **12** (2002) 40–52
16. Naphade, M.R., et al.: A framework for moderate vocabulary semantic visual concept detection. Proc. IEEE ICME (2003) 437–440
17. Papageorgiou, P.C., Oren, M., Poggio, T.: A general framework for object detection. Proc. of ICCV (1997) 555–562
18. Schmid, C.: Constructing models for content-based image retrieval. Proc. of CVPR (2001) 39–45
19. Selinger, A., Nelson, R.C.: Minimally supervised acquisition of 3D recognition models from cluttered images. Proc. of CVPR (2001) 213–220
20. Smeulders, A.W.M., et al.: Content-based image retrieval at the end of the early years. IEEE Trans. on PAMI **22** (2000) 1349–1380
21. Swain, M.J., Ballard, D.N.: Color indexing. Intl. J. Computer Vision **7** (1991) 11–32
22. Town, C., Sinclair, D.: Content-based image retrieval using semantic visual categories. Technical Report 2000.14. AT&T Research Cambridge (2000)
23. Vailaya, A., et al.: Bayesian framework for hierarchical semantic classification of vacation images. IEEE Trans. on Image Processing **10** (2001) 117–130
24. Wang, L., Chan, K.L., Zhang, Z.: Bootstrapping SVM active learning by incorporating unlabelled images for image retrieval. Proc. of IEEE CVPR (2003)
25. Weber, M., Welling, M., Perona, P.: Unsupervised learning of models for recognition. Proc. of ECCV (2000) 18–32
26. Wu, Y., Tian, Q., Huang, T.S.: Discriminant-EM algorithm with application to image retrieval. Proc. of CVPR (2000) 1222–1227