

Home Photo Indexing using Learned Visual Keywords

Joo-Hwee Lim

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

Jesse S. Jin

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

Abstract

With rapid advances in sensor, storage, processor, and communication technologies, consumers can now afford to create, store, process, and share large digital photo collections. With more and more digital photos accumulated, consumers need effective and efficient tools to index and retrieve relevant photos. In this paper, we propose a novel image representation called *Visual Keyword Histogram* (VKH) for content-based indexing and retrieval. Visual keywords are domain-relevant visual prototypes (e.g. faces, foliage, buildings etc) with both perceptual appearance and textual semantics. Collectively, VKHs are computed over spatial tessellation to represent the distribution of visual keywords in various parts of an image. To construct a vocabulary of visual keywords, an incremental neural network is adopted to learn visual keywords from examples. This allows us to build domain-specific visual vocabularies rapidly and incrementally. We demonstrate our approach on 2400 home photos with 15 semantic queries.

keywords: Image Indexing, Image Retrieval

1 Introduction

With recent advances and affordability in digital imaging devices, storage, processors, and communications, consumers can now easily create, store, process, and share large digital photo collections. In particular, with the ease of use (e.g. viewfinder) and flexibility (e.g. deletion) of digital cameras, consumers tend to take and accumulate more and more digital photos. Hence they need effective and efficient tools to index and retrieve relevant photos. As a matter of fact, research on content-based image retrieval (CBIR) in the last decade [22] was focused on general CBIR (e.g. on Corel images). As a consequence, key efforts have been concentrated on using low-level features such as color, texture, and shape to describe and compare image contents. CBIR is yet to bridge the semantic gap between feature-based indexes computed automatically and human expectation on retrieval outcome.

In this paper, we propose a novel image representation called *Visual Keyword Histogram* (VKH) for content-based indexing and retrieval. Visual keywords are domain-relevant visual prototypes (e.g. faces, foliage, buildings etc) with both perceptual appearance and textual semantics. Instead of representing and thus indexing an image content as an aggregate of primitive feature measures (e.g. color his-

togram, edge histogram) or as a set of imperfectly segmented regions that do not necessarily possess clear object semantics, we represent the results of multi-scale detection of visual keywords on local receptive fields as histograms of visual keywords which are further aggregated spatially as image index.

In essence, we transform a local image region from low-level feature space to a new pattern space spanned by visual keywords [11]. Collectively, VKHs are computed over spatial tessellation to represent the distribution of visual keywords in various parts of an image. To compute these VKHs, a 3-layer feed-forward visual information processing architecture is proposed. The first layer presents raw image pixels. The next layer captures a tessellation of fuzzy object distribution patterns. Each pattern is derived from comparing a local region with a vocabulary of visual keywords. No image segmentation is required and thus object detection is not dependent on pre-segmented regions. Moreover, no object classification decision is made to label the content. The last layer further tessellates over the second layer to summarize the image content according to a spatial configuration. To construct a vocabulary of visual keywords, an incremental neural network [8, 9] is adopted to learn visual keywords from examples. This allows us to build domain-specific visual vocabularies rapidly and incrementally. We demonstrate our approach on 2400 home photos with 15 semantic queries.

2 Related Works

In the past, global measures of primitive features such as color, texture, and shape are exploited to index and retrieve visual documents (e.g. [1, 14, 17]). However, this approach often produce results incongruent with human expectations [12] because it does not consider spatial localities and higher-level perceptive cues. For example, images sharing similar overall color distribution can differ greatly in semantic content. This paradigm roughly corresponds to pre-attentive similarity matching which is a low-level function in human visual perception. Nevertheless, new low-level features such as banded color correlograms [5], joint histograms [16] etc are still being proposed to improve the approach on global measures of low-level features.

In contrast, recent region-based methods (e.g. [3, 20, 23]) pre-segment an image by color (or both color and texture) into regions and compute the similarity between two images in terms of the features (and spatial relationships [20]) of these regions. But image segmentation is generally unreliable. A poor segmentation can result in incongruent regions for further similarity matching.

The approach proposed here strives to go beyond primitive features and segmented regions. Based on

the principles of generic view-based object detection [10, 15] with spatial aggregation, the proposed description scheme captures intuitive visual semantics (i.e. visual keywords) specified for a given visual content domain.

3 Visual Keyword Histograms

Our objective is to represent the content of an image beyond simple aggregate measures of primitive features. Furthermore, we want to preserve local object cues without premature decision dictated by segmentation and classification.

We propose to represent an image content against a set of visual prototypes called Visual Keywords [10, 11] that are relevant to a content domain. These visual keywords span a new high-level pattern space into which any arbitrary local image region (say indexed as coordinate p, q) can be projected (Fig. 1) by comparing its perceptual features such as color, texture, and shapes with those of the visual keywords. In this new high-level pattern space (right-hand-side of Fig. 1), each axis is a measure of the fuzziness of the local region resembling the visual keyword for that axis. These fuzzy memberships to the visual keywords are computed based on the relative distances of the local instance to the visual keywords in the perceptual feature space (left-hand-side of Fig. 1). As these fuzzy memberships sum to unity, the projected manifold is a constrained hyperplane shown as shaded area in Fig. 1.

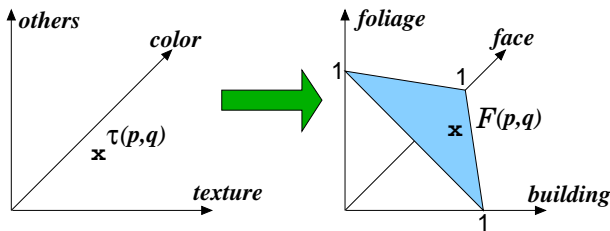


Figure 1: Projection from feature space to visual keyword space

In essence, visual keywords are visual prototypes obtained by supervised learning from examples. A visual content is represented as flexible spatial aggregation of soft presence of visual keywords, upon synchronized results of multi-scale view-based detection. The projection of local image content into the visual keyword space is carried out in two stages as realized by our proposed 3-layer feedforward visual information processing architecture described below. Robust object segmentation is not required.

3.1 Learning Visual Keywords

In general, visual keywords are organized as a hierarchy of visual concepts. That is, they are visual object classes (or subclasses) whose instances have concrete visual appearance in the images of a domain. In this paper, we are only concerned with home photographs and we design a simple two level concept hierarchy for practical usage, namely visual object classes and their prototypes. Fig. 2 illustrates some of the examples used to train the visual keywords in our experiments.

Visual keywords of the same class form an equivalence class of visual synonyms. This concept-oriented visual thesaurus is different from the visual relations proposed by R.W.Picard [18], which are founded on similarities between low-level visual features. Thus visual keywords are highly flexible visual knowledge

that can be customized according to a content domain.

In our experiments, we use color and texture to characterize a visual keyword v and a training example τ . For color, we have chosen the YIQ color space over other color space (e.g. RGB, HSV, LUV) and means plus standard deviations over other color feature measures (e.g. local color histograms) as they have better performance in our experiments. Similar reason has been reported [6]. For texture, we adopted the Gabor features which have been shown to provide excellent pattern retrieval results [13]. The feature vector v (or τ) constitutes two parts, namely, a color feature vector v^c (or τ^c) and a texture feature vector v^t (or τ^t). For v^c (or τ^c), the means and standard deviations of the Y, I, Q color channels within the scanned window are computed. For v^t (or τ^t), the means and standard deviations of the coefficients, which are the outcome of convolution with Gabor filters of 5 scales and 6 orientations, within the scanned window are adopted [13]. Thus each feature vector v (τ) has 66 (6+60) dimensions and they are subjected to zero-mean normalization.

In this paper, we adopted a Supervised Incremental Clustering Architecture (SICA) [8, 9] to learn visual keywords from examples.

SICA is a 3-layer feedforward neural network with dynamic node creation capability (Fig.3). Each input node corresponds to a feature and each output node is a class. The only hidden layer, which grows prototypes from scratch, captures the regularity of input examples through learning. Each hidden node (or prototype) receives full connections from the input layer, with a weight vector representing the position of the prototype in the input space. Prototypes of the same class are joined to the output node denoting their class with weight values '1', thus giving an 'OR' (union) operation. Learning involves the modification of the weight vectors to the prototypes as well as the recruitment and initialization of new prototypes.

When an input vector X is presented, the closest prototype M_k from among the existing prototypes, M_i , is first determined as follows

$$\Omega(X, M_k) \geq \Omega(X, M_i) \quad \forall i, \quad (1)$$

where $\Omega(X, Y) \in [\Omega_{min}, \Omega_{max}]$, $\Omega_{min}, \Omega_{max} \in R$, is some similarity function between vectors X and Y .

If the following conditions are fulfilled

$$class(M_k) = class(X) \vee \Omega(X, M_k) > \alpha, \quad (2)$$

where $class(X)$ gives the class label of X and α is a *Prototype Creation Threshold (PCT)*, we adapt M_k towards X

$$M_k \leftarrow \frac{n_k \cdot M_k + X}{n_k + 1}, \quad (3)$$

$$n_k \leftarrow n_k + 1, \quad (4)$$

where n_k is the number of examples that have been 'won' by (i.e. assigned to) M_k . This update rule ensures that the prototypes are indeed the means of all examples that have been assigned to them. In this way, similar cases are generalized to their statistical average (i.e. *local generalization*). When n_k goes to infinity, the movement of winners will diminish asymptotically. Therefore, it implements some form of decaying learning rate automatically.

Otherwise (i.e. if Equation (2) is not satisfied), we have a wrong classification. We memorize X as a new prototype

$$M_{new} \leftarrow X, \quad (5)$$

$$n_{new} \leftarrow 1. \quad (6)$$

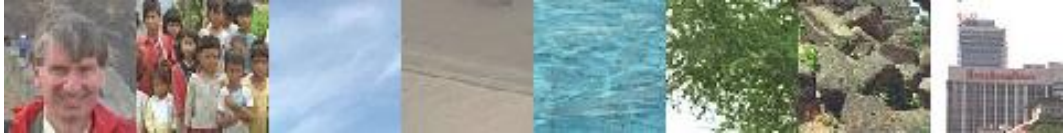


Figure 2: One example from each visual object class: face, crowd, sky, ground, water, foliage, mountain/rock, building

where M_{new} is a dynamically created prototype.

The similarity function between a training example τ and a visual keyword v is defined as

$$\Omega(\tau, v) = 1 - \delta(\tau, v), \quad (7)$$

where $\delta(\tau, v)$ is a distance measure between τ and v . In our experiments, it is defined as

$$\delta(\tau, v) = \frac{|\tau^c - v^c|}{N_c} + \frac{|\tau^t - v^t|}{N_t} \quad (8)$$

where $|\cdot|$ is the city block distance, N_c and N_t are the dimensions of the color and texture feature vectors respectively.

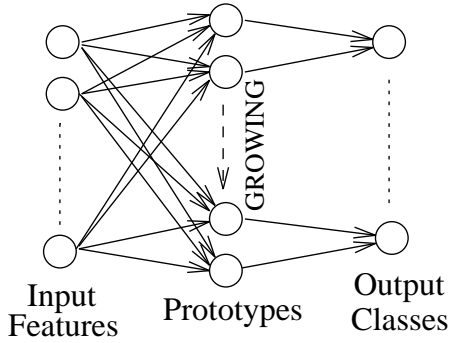


Figure 3: Supervised Incremental Clustering Architecture

3.2 Image Indexing

We propose a 3-layer feedforward visual information processing architecture to automatically compute a tessellation of visual keyword histograms as image index. Essentially an image is scanned with windows of different scales. Each scanned window is a *visual token* reduced to a feature vector τ compatible (i.e. same feature types and dimension) to those of the visual keywords v . The fuzzy memberships to these visual keywords are registered in a *Fuzzy Object Map (FOM)* (Fig. 4).

More precisely, given an image I with resolution $M \times N$, a *FOM* \mathcal{F} has a lower resolution of $P \times Q$, $P \leq M$, $Q \leq N$. Each pixel (p, q) corresponds to a two-dimensional region of size $r_x \times r_y$ in I . We further allow tessellation displacements $d_x, d_y > 0$ in X, Y directions respectively such that adjacent pixels in \mathcal{F} along X direction (along Y direction) have receptive fields in I which are displaced by d_x pixels along X direction (d_y pixels along Y direction) in I .

The feature vector τ_{pq} computed for this region is compared against the feature vectors v_{ij} of all visual keywords (class i , prototype j) to derive a fuzzy membership vector $\mu(\tau_{pq}, v_{ij})$ ($\sum_{ij} \mu(\tau_{pq}, v_{ij}) = 1$),

$$\mu(\tau_{pq}, v_{ij}) = \frac{\frac{1}{\delta(\tau_{pq}, v_{ij})^2}}{\sum_{ij} \frac{1}{\delta(\tau_{pq}, v_{ij})^2}}, \quad (9)$$

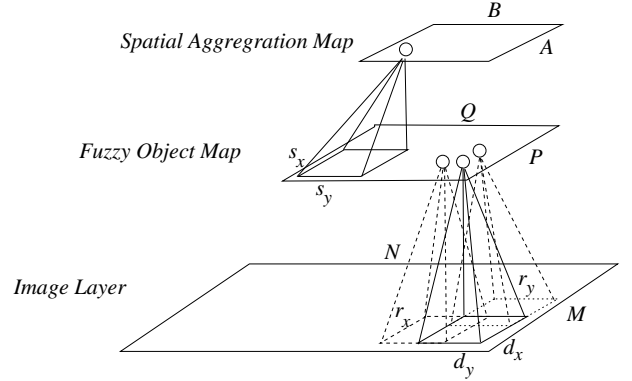


Figure 4: Visual keyword detection and spatial aggregation

where $\delta(\tau_{pq}, v_{ij})$ is a distance measure between τ_{pq} and v_{ij} . In our experiments, it is defined as

$$\delta(\tau_{pq}, v_{ij}) = \frac{|\tau_{pq}^c - v_{ij}^c|}{N_c} + \frac{|\tau_{pq}^t - v_{ij}^t|}{N_t} \quad (10)$$

where $|\cdot|$ is the city block distance, N_c and N_t are the dimensions of the color and texture feature vectors respectively.

Other alternatives to derive a soft membership vector includes the *softmax* function [2] which has been adopted for modelling conditional probabilities by neural networks researchers,

$$\mu(\tau_{pq}, v_{ij}) = \frac{\exp^{\delta(\tau_{pq}, v_{ij})}}{\sum_{ij} \exp^{\delta(\tau_{pq}, v_{ij})}}. \quad (11)$$

In a nutshell, the *FOM* records the outcome of visual keyword detection across different locations (i.e. translation invariance) in image I at scale k as specified by $r_x \times r_y$,

$$\mathcal{F}^k(p, q, i, j) = \mu_k(\tau_{pq}, v_{ij}). \quad (12)$$

To achieve scale invariance, we have multiple *FOMs* with different scan window sizes $r_x \times r_y$ (e.g. $20 \times 20, 30 \times 30, \dots, 60 \times 60$). We designed a simple algorithm to synchronize \mathcal{F}^k into a single *FOM* \mathcal{F} of the lowest resolution ($r_x \times r_y$ as 20×20). Last but not least, appropriate feature measures can be designed to cater for at least some degree of rotational or skew invariance in visual keyword detection. For simplicity, we only show one *FOM* in Fig. 4.

Likewise, *SAM* \mathcal{S} tessellates over *FOM* \mathcal{F} with $A \times B$, $A \leq P$, $B \leq Q$ pixels. Each *SAM* pixel (a, b) aggregates the fuzzy memberships for visual keyword ij over those *FOM* pixels (p, q) (which in turn cover a tessellated set of regions in I) covered by (a, b) ,

$$\mathcal{S}(a, b, i, j) = \sum_{(p, q) \in (a, b)} \mathcal{F}(p, q, i, j), \quad (13)$$

which is normalized to sum to unity by dividing it with $\sum_{ij} \mathcal{S}(a, b, i, j)$. That is, the *SAM* is a spatial tessellation of fuzzy distributions of visual keywords to characterize the visual content of an image. These visual information processing steps are depicted in Fig. 4.

An alternative aggregation measure is to compute the cardinality of the α -cut ([7], p.17) for each visual keyword over the spatial region covered by (a, b) ,

$$\mathcal{S}_\alpha(a, b, i, j) = |\{(p, q) \in (a, b) | \mathcal{F}(p, q, i, j) \geq \alpha\}|, \quad (14)$$

which is similarly normalized to sum to unity by dividing it with $\sum_{ij} \mathcal{S}_\alpha(a, b, i, j)$. As each visual keyword corresponds to some pseudo-object instance in the visual domain, \mathcal{S}_α can be viewed as an *object histogram* of a spatial region covered by (a, b) , as opposed to the conventional aggregate measure of low-level features (e.g. color histograms).

Visual keywords v_{ij} describes a specific appearance of a visual object class i . They are visual synonyms that allow further abstraction. Aggregate measures based on this higher level abstraction can be carried out as,

$$\mathcal{C}(a, b, i) = \sum_j \mathcal{S}(a, b, i, j), \quad (15)$$

and

$$\mathcal{C}_\alpha(a, b, i) = |\{(p, q) \in (a, b) | \sum_j \mathcal{F}(p, q, i, j) \geq \alpha\}|. \quad (16)$$

From the point of pattern recognition, visual keywords span a new object-level feature space in which spatial aggregation is computed. Each visual keyword v_{ij} denotes a dimension in this new feature space with feature value $\mathcal{F}(p, q, i, j)$ in $[0, 1]$ to represent its presence in a scan window (p, q) . For any scan window (p, q) in the image, $\mathcal{F}(p, q, i, j) \forall i, j$ is a feature vector whose feature values sum to unity. Geometrically, $\mathcal{F}(p, q)$ is a point within the constrained hyperplane (i.e. $\sum_{ij} \mu(\tau_{pq}, v_{ij}) = 1$) as shown schematically in Fig. 1.

3.3 Similarity Matching

As *SAM* summarizes the visual content of an image in terms of spatial distribution of prototypical visual objects, it can be used as a signature for comparing the similarity between two images. The similarity $\lambda(x, y)$ between two images x and y is computed as a weighted average of the similarities between the corresponding parts of the images,

$$\lambda(x, y) = \frac{\sum_{(a,b)} \omega(a, b) \lambda(a, b)}{\sum_{(a,b)} \omega(a, b)}, \quad (17)$$

where $\omega(a, b)$ is the weight assigned to the region covered by (a, b) in \mathcal{S} (or \mathcal{C}), and $\lambda(a, b)$ is defined as,

$$\lambda(a, b) = 1 - \frac{1}{2} |\mathcal{S}_x(a, b) - \mathcal{S}_y(a, b)| \quad (18)$$

where $|\cdot|$ is the city block distance. This similarity measure between two fuzzy distributions of visual keywords (Equation (13)) or two object histograms (Equation (14)) is consistent with the histogram intersection metric adopted widely for color histograms [21].

Table 1: Equivalent classes of visual keywords for home photos

CLASS	Sub-Class (Num. of prototypes)
PEOPLE	Face (9), Human Figure (13),
	Crowd (5), Skin (5)
SKY	Clear (1), Cloudy (5), Blue (3)
GROUND	Floor (10), Sand-like (6), Grass (4)
WATER	Pool (3), Pond (7), River (6)
FOLIAGE	Green (7), Flowers (6), Branches (7)
MOUNTAIN	Far (3), Rocks (4)
BUILDING	Classic (15), Modern (9), Far (14)
INTERIOR	Wall (6), Wooden Furniture (1),
	China (7), Fabric (7), Light (4)

4 Empirical Evaluation

In this paper, we have designed a visual vocabulary for home photos. There are 8 classes of visual keywords, each subdivided into 2-5 subclasses. Hence there are 26 distinct labels in total (Table 1). We used the SICA learning algorithm as described above to learn these 26 classes of visual keywords from 375 labeled image regions cropped from home photos.

The experimental results shown here are based on 2400 genuine family photos provided by a colleague, Mr. Jean-Luc Lebrun, who travels frequently. We focus on actual home photos instead of the more generalized image collections like Corel images because our research aims to create useful and automatic tools for mass consumers to organize and retrieve their home photos. Figure 5 displays typical photos in this collection and Figure 6 shows some of the photos with bad quality (e.g. faded, over-exposed, 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.

The photos come in different orientations, namely landscape (384×256) and portrait (256×384) layouts. The indexing process automatically detects the layout and applies the corresponding tessellation template. The spatial tessellation (a, b) adopted for *SAM* is shown in Fig. 7. The relative weight $\omega(a, b)$ for the 5 tessellations used in weighted similarity matching (Equation (17)) are also given. The center tessellation is given a higher weight as it is usually the focus of a photo content. An equivalent tessellation is also designed for images with portrait layout.

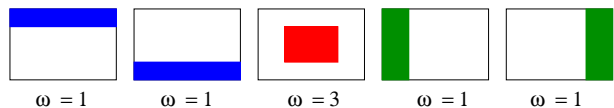


Figure 7: Tessellation and weight for query by visual example

We defined 15 semantic queries and their ground truths (G.T.) among the 2400 photos (Table 2). For each query, we selected 3 relevant photos as query examples for Query By Example (QBE) experiments.

We compare our proposed visual keyword histogram method (denoted as "VKH") with color histogram of eleven key colors (red, green, blue, black, grey, white, orange, yellow, brown, pink) in the HSV color space, as adopted by the original PicHunter system [4] (denoted as "HSV").

Table 3 lists the average precisions of retrieval for the 15 queries given in Table 2 using the two methods. Table 4 shows the average precisions (over all 15



Figure 5: Typical home photos used in our experiment



Figure 6: Some home photos of inferior quality

Table 2: The 15 queries used for benchmarking

Query	Description	G.T.
Q01	indoor	994
Q02	outdoor	1218
Q03	night or dark scene	101
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 road side	645
Q10	along waterside	150
Q11	in a park or garden	303
Q12	at mountain area	67
Q13	people close up	277
Q14	large group, indoor	45
Q15	large group, outdoor	34

Table 3: Average precisions for each and all queries

Avg. Prec.	HSV	VKH
Q01	0.60	0.72
Q02	0.61	0.63
Q03	0.08	0.33
Q04	0.57	0.77
Q05	0.12	0.24
Q06	0.46	0.41
Q07	0.25	0.39
Q08	0.13	0.23
Q09	0.43	0.31
Q10	0.15	0.19
Q11	0.45	0.47
Q12	0.05	0.28
Q13	0.13	0.26
Q14	0.10	0.37
Q15	0.12	0.22
Overall	0.28	0.39

queries) among the top 10, 20, 30, and 50 retrieved photos for the two methods compared.

From Table 3, we observe that the VKH approach outperforms the HSV approach in all queries except for queries Q06 and Q09. In particular, significant improvements (i.e. 50% or more) can be seen for queries Q03, Q05, Q07, Q08, Q12, Q13, Q14, and Q15. As a whole, the VKH method attained a 39% improvement in terms of average precision over the HSV method.

In practice, precision values at relatively small number of retrieved images are more important as a user would like to find relevant images to their queries within the first couple of pages of thumbnails of retrieved images. When we look at the average precisions of up to first 50 retrieved images as shown in Table 4, we conclude that the VKH method is able to display at least 30% (i.e. 31% to 46%) more relevant images than the HSV method to a user. In concrete terms, these translate to 2.3, 4.2, 5.1, and 6.0 more relevant images at top 10, 20, 30, and 50 retrieved images respectively on the average.

5 Conclusions and Future Works

In this paper, we have presented a novel content representation called visual keyword histogram and an associated processing architecture to transform image representation in low-level feature space to semantic space spanned by visual keywords. In our QBE experiment on 2400 home photos with 15 semantic queries,

Table 4: Average precisions at top images

Avg. Prec.	HSV	VKH	Improvement
At 10 photos	0.50	0.73	46%
At 20 photos	0.46	0.67	46%
At 30 photos	0.42	0.59	40%
At 50 photos	0.39	0.51	31%

we achieved very promising results using our proposed new scheme for image indexing and retrieval.

In this paper, we adopted SICA as the supervised pattern classifier to learn visual keywords as it allows rapid incremental learning. Visual keywords, which are reference vectors in the feature space to register local image regions in the new pattern space, can be generalized as kernels [19] for better generalization power. Currently we are experimenting with support vector machines as the pattern classifier for learning the visual keywords.

Last but not least, the spatial aggregation method (Equation (13) and Fig. 7) used is simplistic. More sophisticated pattern modeling techniques like graphical models can be useful to capture contextual information from examples.

References

- [1] Bach, J.R. et al. Virage image search engine: an open framework for image management. In *Storage and Retrieval for Image and Video Databases IV*, Proc. SPIE 2670, 1996, pp. 76-87.
- [2] Bishop, C.M. *Neural Networks for Pattern Recognition*. Clarendon Press, Oxford, 1995.
- [3] Carson, C. et al. Color- and texture-based image segmentation using EM and its application to image query and classification. Submitted to *IEEE Tran. PAMI*, 1999.
- [4] Cox, I. et al. The Bayesian image retrieval system, PicHunter: theory, implementation and psychophysical experiments. *IEEE Trans. on Image Processing*, 2000, **9**(1): 20-37.
- [5] Huang, J., Kumar, S.R., & Zabih, R. An automatic hierarchical image classification scheme. In *Proc. of ACM Multimedia '98*, 1998, pp. 219-228.
- [6] Jacobs, C.E., Finkelstein, A., & Salesin, D.H. Fast multiresolution image querying. In *Proc. SIGGRAPH'95*, 1995.
- [7] Klir, G.J., & Folger, T.A. *Fuzzy Sets, Uncertainty, and Information*. Prentice Hall, 1992.
- [8] Lim, J.H. Incremental case-based pattern classifier. In *Proc. of the International Conference on Artificial Neural Networks, Amsterdam, The Netherlands, Sep. 13-16, 1993*.
- [9] Lim, J.H. Incremental neural classifier with prototype reduction. In *Proc. of ICARCV-96, Singapore, Dec. 3-6, 1996*, pp. 933-936.
- [10] Lim, J.H. Visual keywords: From text IR to multimedia IR. In F.Crestani & G.Pasi (ed.), *Soft Computing in Information Retrieval: Techniques and Applications*, Physica-Verlag, Springer Verlag, Germany, 2000, pp. 77-101.
- [11] Lim, J.H. Building visual vocabulary for image indexation and query formulation. *Pattern Analysis and Applications* 2001, **4**(2/3): 125-139.
- [12] Lipson, P., Grimson, E., & Sinha, P. Configuration based scene classification and image indexing. In *Proc. of CVPR '97*, 1997, pp. 1007-1013.
- [13] Manjunath, B.S., & Ma, W.Y. Texture features for browsing and retrieval of image data. *IEEE Trans. PAMI* 1996, **18**(8): 837-842.
- [14] Niblack, W. et al. The QBIC project: querying images by content using color, textures and shapes. In *Storage and Retrieval for Image and Video Databases, Proc. SPIE 1908*, 1993, pp. 13-25.
- [15] Papageorgiou, P.C., Oren, M., Poggio, T. A general framework for object detection. In *Proc. ICCV*, 1997, pp. 555-562.
- [16] Pass, G. & Zabih, R. Comparing images using joint histograms. *Multimedia Systems* 1999, **7**: 234-240.
- [17] Pentland, A., Picard, R.W., & Sclaroff, S. Photobook: content-based manipulation of image databases. *International Journal of Computer Vision* 1995, **18**(3): 233-254.
- [18] Picard, R.W. Toward a Visual Thesaurus. In *Proc. of Springer-Verlag Workshops in Computing, MIRO'95, Glasgow, Sep. 1995*.
- [19] Scholkopf, B., Burges, C.J.C., & Smola, A.J. *Advances in Kernel Methods - Support Vector Learning*. MIT Press, Cambridge, MA, 1999.
- [20] Smith, J.R. & Chang, S.-F. VisualSEEK: a fully automated content-based image query system. In *Proc. ACM Multimedia 96*, Boston, MA, Nov. 20, 1996.
- [21] Swain, M.J. & Ballard, D.N. Color indexing. *International Journal of Computer Vision* 1991, **7**(1): 11-32.
- [22] Smeulders, A. et al. Content-based image retrieval at the end of the early years. *IEEE Trans. on PAMI* 2000, **22**(12): 1349-1380.
- [23] Wood, M.E.J., Campbell, N.W., & Thomas, B.T. Employing region features for searching an image database. In *Proc. 1997 British Machine Vision Conference*, 1997, pp. 620-629.