Automatically Annotating Images with Keywords: A Review of Image Annotation Systems

Chih-Fong Tsai[1,*] and Chihli Hung[2]

[1]*Department of Accounting and Information Technology, National Chung Cheng University, Taiwan,* [2]*Department of Management Information Systems, Chung Yuan Christian University, Taiwan*

Received: August 1, 2007; Accepted: November 7, 2007; Revised: November 20, 2007

Abstract: The explosive growth of image data leads to the research and development of Content-Based Image Retrieval (CBIR) systems. CBIR systems extract and retrieve images by their low-level features, such as color, texture, and shape. However, these visual contents do not allow users to query images by semantic meanings. Image annotation systems, a solution to solve the inadequacy of CBIR systems, aim at automatically annotating image with some controlled keywords. Machine learning techniques are used to develop the image annotation systems to map the low-level (visual) features to high-level concepts or semantics. This paper reviews 50 image annotation systems using supervised machine learning techniques to annotate images for image retrieval. Future research issues are provided.

Keywords: Image annotation, machine learning, content-based image retrieval.

1. INTRODUCTION

The advances of computing and multimedia technologies allow the construction and archive of images with low cost. As a result, the size of image collections is increasing. Content-based image retrieval (CBIR) was proposed to allow users retrieve relevant images in an effective and/or efficient manner [1, 2].

Some representative CBIR systems, such as QBIC [3], VisualSEEk [4], Photobook [5], etc. support content-based retrieval by color, texture, and shape. These visual features can be automatically extracted by some image processing techniques. However, retrieval results of CBIR systems are not satisfactory. This is because humans recognize images based on high-level concepts. That is, users are familiar with natural language-like queries, such as text and typically query images by semantics [6-8].

Image annotation or automatically annotate images with keywords is a solution to this problem. It is based on some machine learning techniques, which learn the correspondence between visual features and semantics of images. That is, image annotation systems can recognize or classify visual features into some pre-defined classes [9]. Figure **1** shows a general architecture of image annotation systems.

The segmentation component partitions images into local contents via either some block or region based method (c.f. Section 2). Then, the feature extraction component extracts low-level features from the segmented images (c.f. Section 3). That is, each segmented block or region is represented by feature vectors. Next, the annotation component assigns the (low-level) feature vectors to some pre-defined categories. This performs like the pattern classification task. Finally, the post-processing component (dependent on the application) uses the output of the annotation component to decide on some recommended action for the final decision.

*Address correspondence to this author at the Department of Accounting and Information Technology, National Chung Cheng University, Taiwan; Tel: +88652720411; Ext: 34519; Fax: +88652721197;
E-mail: actcft@ccu.edu.tw

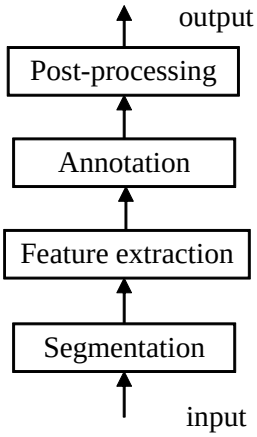

Fig. (1). Block diagram of an image annotation system.

This paper reviews 50 image annotation systems published from 1997 to 2006 in terms of their image processing and learning modules, numbers of keywords assigned per image, scalability (numbers of keywords that the systems learn for annotation), and dataset used.

This paper is organized as follows. Section 2 briefly describes commonly used image segmentation methods including global and local feature extraction. Section 3 covers the low-level features which are generally extracted from images. Section 4 overviews the mostly used supervised learning methods for the task of image annotation. Some representative image annotation systems are also described. Section 5 presents advanced techniques by combining multiple classifiers and hybrid classifiers for image annotation. Section 6 provides a comparison of related image annotation systems in terms of their feature representation, classifier used, classification scale, etc. for image annotation. Discussion of some future research issues for image annotation and conclusion of the survey are given in Section 7 and 8 respectively.

2. IMAGE SEGMENTATION

In general, visual content of an image can be represented by either global or local features. Global features take all the

1874- 4796 /08 $100.00+.00

pixels of an image into account. Color histograms [10], for examples, can be extracted to represent or describe the global color content of images. In this case, an image can be described that it contains 40% of *blue*, 37% of *yellow* and so on. However, as global features consider the visual features of the whole image, they cannot completely describe different parts of an image. On the other hand, image segmentation into local contents (i.e. different regions or areas) is able to provide more detailed information of images.

In general, there are two strategies for extracting local features [11-18]. The first one is to partition a set of fixed sized blocks or tiles (see Fig. **2** for some examples) and the second for a number of variable shaped regions of interest (see Fig. **3**) [19]. After performing block and/or region based segmen-tation, low-level feature(s) can be extracted from the tiles or regions for local feature representation [19].

3. IMAGE LOW-LEVEL FEATURES

In general, low-level features such as color, texture, shape, and spatial relationship are extracted to represent image features.

3.1. Color

Color is the most used visual feature for image retrieval due to the computational efficiency of its extraction. All colors can be represented variable combinations of the three so-called primary colors: red (R), green (G), and blue (B). There are some other color spaces for representing the color feature, such as HSV, L*u*v*, YIQ, etc. [13]. In particular, color histogram [10] is one common method used to represent color contents for indexing and retrieval. It shows the proportion of pixels of each color within the image, which is represented by the distribution of the number of pixels for each quantized bin.

3.2. Texture

Texture is an important element of images for surface, object identification, and region distinctions. In addition to colors, it has been extracted to classify and recognize objects and scenes. Texture can be regular or random. Most natural textures are random. Regular textures are composed of textures that have a regular or almost regular arrangement of identical, or at least similar, components. Irregular textures are composed of irregular and random arrangements of components related some statistical properties [20].

3.3 Shape

Shape is one of the most important features for describing the content or object(s) of an image. Compared with colour and texture features, shape features are usually

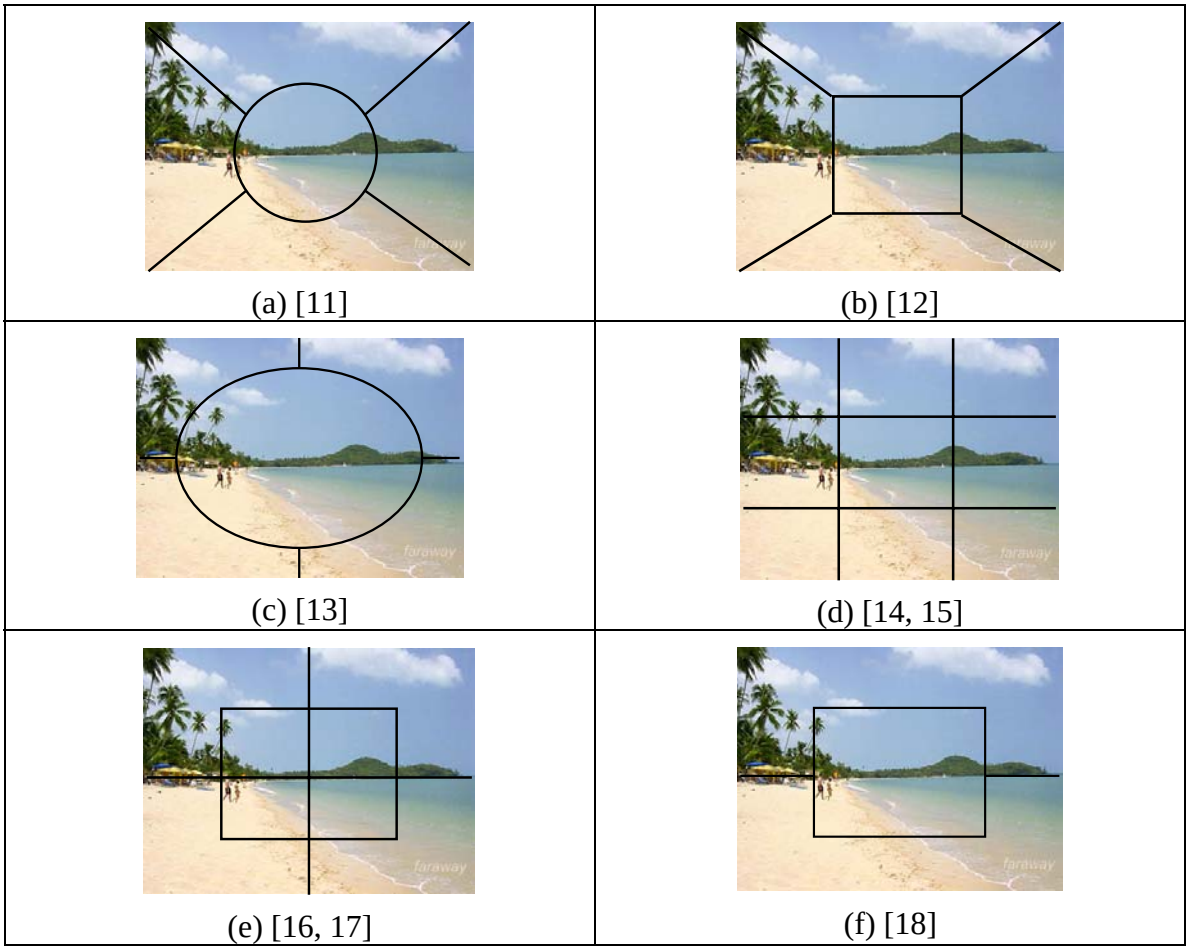

Fig. (2). Examples of block-based segmentation.

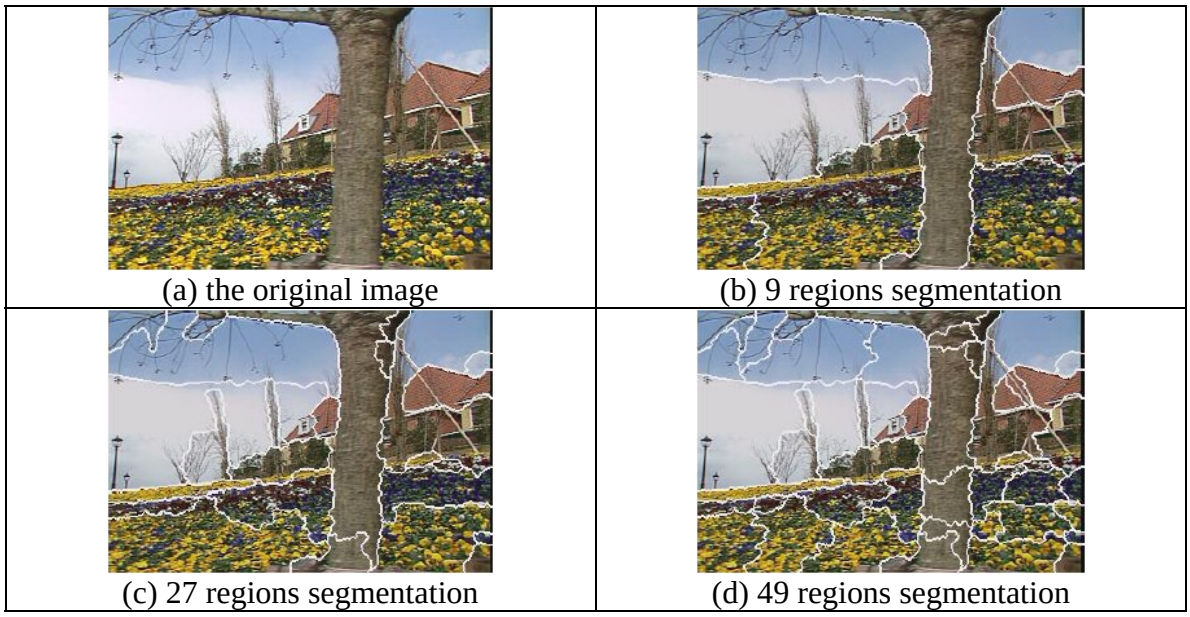

Fig. (3). Region-based segmentation [19].

described after images have been segmented into regions or objects. The shape representations can be divided into two categories, boundary-based (or edge detection) and region-based. The former uses only the outer boundary of the shape, such as the chain code method, while the latter uses the entire shape region [1]. However, to effectively extract shape features depends on segmentation methods.

3.4. Spatial Relationship

Objects and the spatial relationships (such as *left of*, *inside*, and *above*) among objects in an image are used to represent the image content [21]. That is, an image can be divided into a number of sub-blocks as described in Section 2 and colour, texture, and/or shape features are extracted from each of the sub-blocks. Then, we can project them along the *x* and *y* axes, such as 'left/right', 'below/above' relationships between them. Ko *et al*. [14] consider spatial colour histograms which show better performances than the traditional one, i.e. global colour histogram.

4. SUPERVISED LEARNING MODELS

The goal of pattern classification is to allocate an object represented by a number of measurements (i.e. feature vectors) into one of a finite set of classes. Supervised learning can be thought as learning by examples or learning with a teacher [22]. The teacher has knowledge of the environment which is represented by a set of input-output examples. In order to classify unknown patterns, a certain number of training samples are available for each class, and they are used to train the classifier [23].

The learning task is to compute a classifier or model that approximates the mapping between the input-output examples and correctly labels the training set with some level of accuracy. This can be called the *training* or *model generation* stage. After the model is generated or trained, it is able to classify an unknown instance, into one of the learned class labels in the training set. More specifically, the classifier calculates the similarity of all trained classes and assigns the unlabeled instance to the class with the highest similarity measure.

Therefore, image annotation can be approached by the model/classifier to bridge the gap between low-level feature vectors and high-level concepts, i.e. a function is learned which can directly correspond the low-level feature sets to high-level conceptual classes.

4.1. Probabilistic Classifiers

The naïve Bayes classifier requires all assumptions be explicitly built into models which are then used to derive 'optimal' decision/classification rules. It can be used to represent the dependence between random variables (features) and to give a concise and tractable specification of the joint probability distribution for a domain. It is constructed by using the training data to estimate the probability of each class given the feature vectors of a new instance. Given an example represented by the feature vector *X*, the Bayes' theorem provides a method to compute the probability that *X* belongs to class C_i, denoted as $p(C_i|X)$:

$$P(C_i|X) = \prod_{j=1}^{N} P(x_j|C_i) \qquad (1)$$

That is, the naïve Bayes classifier learns the conditional probability of each attribute x_j (j = 1, 2, …, N) of X given the class label C_i. Therefore, the (image) classification problem can be stated as given a set of observed features x_j, from an image *X*, classify *X* into one of the classes C_i.

Vailaya *et al*. [24] proposed a hierarchical classification scheme to first classify images into indoor or outdoor

categories, then, outdoor images are further classified as city or landscape; finally, landscape images are classified into sunset, forest, and mountain classes. In other words, three Bayes classifiers are used for the three-stage classification.

Ghoshal *et al.* [25] use a hidden Markov model for image annotation based on two datasets individually, which are COREL and TRECVID. Various model parameters and parameter estimations are examined to form the 'best' image annotation model. Similarly, Wang and Li [26] propose a two-dimensional hidden Markov model for image annotation.

Blei and Jordan [27] propose a correspondence latent Dirichl *et al.* location (LDA) model which finds conditional relationships between latent variable representations of image regions and words. This model is compared with a Gaussian-multinomial mixture model and a Gaussian-multinomial LDA model.

4.2. Artificial Neural Networks

Neural networks (or artificial neural networks) learn by experience, generalize from previous experiences to new ones, and can make decisions. A neural network can be thought of as a *black box* non-parametric classifier [23]. That is, different from naïve Bayes, we do not need to make assumptions about the distribution densities. Neural networks are therefore more flexible.

A multilayer perceptron (MLP) network consists of an input layer including a set of sensory nodes as input nodes, one or more hidden layers of computation nodes, and an output layer of computation nodes. The input nodes/neurons are the feature values of an instance, and the output nodes/neurons (usually lying in the range [1]) represent a discriminator between its class and all of the other classes. That is, each output value is a measure of the network's confidence that the class corresponding to the highest output value is returned as the prediction for an instance. Each interconnection has associated with it a scalar weight which is adjusted during the training phase [28]. Figure **4** shows an example of a three-layer feed-forward network having input, output, and one hidden layers.

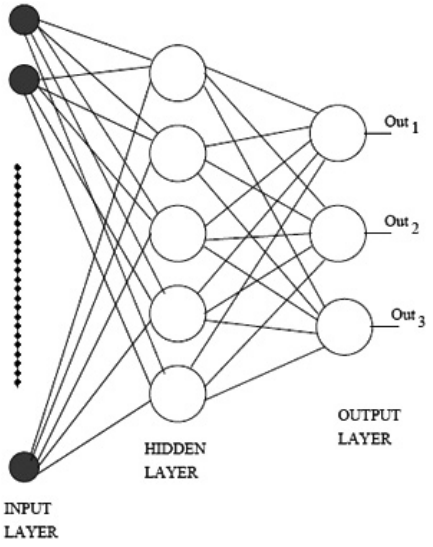

Fig. (4). The three-layer neural network.

For image annotation, low-level feature vectors are fed into the input layer of an MLP where each of the input neurons corresponds to each of the feature vectors, and the output neurons of the MLP represent the class labels of images to be classified.

In Kuroda and Hagiwara [29], an MLP neural network is used to first roughly classify scenery images into *sky*, *earth*, and *water* categories based on color and texture features of image regions. Then, an impression word estimation neural network further estimates ambiguous impression words and concrete nouns of the classified image regions, such as *mountain* from the *earth* category or *cloudy sky* from the *sky* category. These hierarchical neural networks classify image regions into one of 15 categories.

Lim *et al.* [30] propose a typical event taxonomy for home photos. A three-layer feed-forward neural network is designed to learn 26 keywords by using color and texture features of each training patch/block (each image is partitioned into non-overlapping left, right, top, bottom, and centre areas) (like Fig. **2b** in Section 2.2.1). In Park *et al.* [31], each image is first pre-processed to remove background (region) for object segmentation in order to minimize misclassification. Then, a three-layer MLP neural network classifier is designed for image classification for 30 categories. Kim *et al.* [32] combine region-based segmentation using color features and a neural network classifier for object and non-object classifications.

4.3. Support Vector Machines

Support Vector Machines (SVMs) [33] are designed for binary classification. That is, to separate a set of training vectors which belong to two different classes, (x_1, y_1), (x_2, y_2),…, (x_m, y_m) where $x_i \in R^d$ denotes vectors in a d-dimensional feature space and $y_i \in \{-1, +1\}$ is a class label. During the SVM model generation, the input vectors, i.e. low-level feature vectors, such as color and/or texture for image retrieval, are mapped into a new higher dimensional feature space denoted as $\Phi : R^d \rightarrow H^f$ where $d < f$. Then, an optimal separating hyperplane in the new feature space is constructed by a kernel function, $K(x_i, x_j)$ which products between input vectors x_i and x_j where $K(x_i, x_j) = \Phi(x_i) \cdot \Phi(x_j)$. Figure **5** [34] illustrates this procedure. Two most used kernel functions are Polynomial and Gaussian Radial Basis Function (RBF) kernel functions which are $K_{poly}(x_i, x_j) = (x_i \cdot x_j + 1)^p$ (p is the degree of polynomial) and $K_{Gaussian}(x_i, x_j) = e^{\frac{\|x_i - x_j\|^2}{2\delta}}$ (σ is Gaussian sigma) respectively.

All vectors lying on one side of the hyperplane are labelled as -1, and all vectors lying on another side are labeled as +1. The training instances that lie closest to the hyperplane in the transformed space are called support vectors. The number of these support vectors is usually small compared to the size of the training set and they determine the margin of the hyperplane, and thus the decision surface.

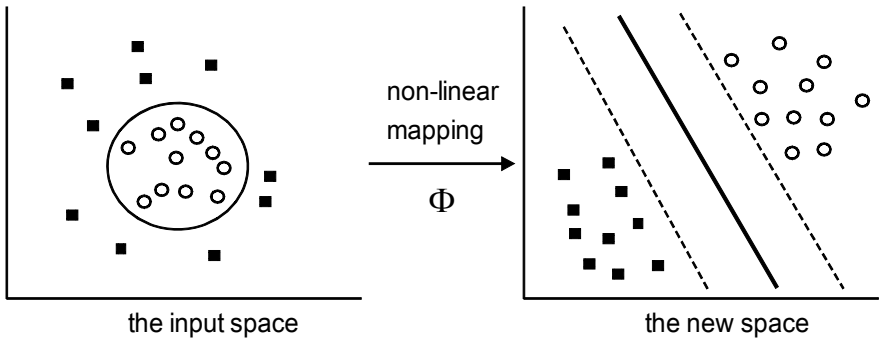

Fig. (5). SVM model generation [34].

In order to produce good generalization, the SVM maximizes the margin of the hyperplane [35] and diminishes the number of support vectors for it [36]. For a detailed survey of SVMs for different pattern recognition applications refer to Byun and Lee [37].

Chapelle *et al*. [38] compare SVMs based on polynomial and Gaussian RBF kernels learned by global RGB and HSV color histograms individually for 7-category classification. In addition, they report that SVMs outperform the *k*-NN classifier. In Brank [39], HSV color auto-correlogram [40] is extracted to train SVMs for 14-category classification.

4.4. Decision Trees

A decision tree classifies an instance by sorting it through the tree to the appropriate leaf node, i.e. each leaf node represents a classification. Each node represents some attribute of the instance, and each branch corresponds to one of the possible values for this attribute [22]. ID3 and C4.5 are the algorithms to construct a decision tree classifier [41]. Figure **6** shows an example of a decision tree.

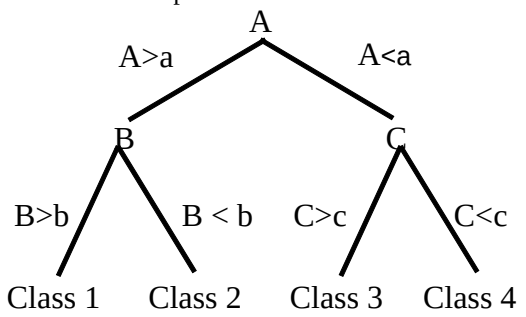

Fig. (6). Example of a decision tree.

Another well known program for constructing decision trees is CART (Classification and Regressing Tree) [42]. A decision tree with a range of discrete (symbolic) class labels is called a classification tree, whereas a decision tree with a range of continuous (numeric) values is called a regression tree.

Huang *et al*. [43] construct a classification tree for hierarchical 11-cateogry classification. They report that using the color correlogram performs better than the general histogram, and the classification tree outperforms the traditional nearest neighbour classifier.

4.5. k-Nearest Neighbor

In pattern classification, the *k*-Nearest Neighbor (*k*-NN) classifier is a conventional non-parametric classifier [28]. To classify an unknown image represented by some feature vectors as a point in the feature space, the *k*-NN classifier calculates the distances between the point and points in the training data set. Then, it assigns the point to the class among its *k* nearest neighbors (where *k* is an integer).

k-NN is different from the *inductive* learning approach described previously. It has thus been called instance-based learning [22] or *lazy learners*. That is, without off-line training (i.e. model generation) the *k*-NN algorithm only needs *searching* through all the examples of the given training set for classifying the new instance. Therefore, the main computation of *k*-NN is the on-line scoring of training examples to find the *k* nearest neighbors of the new instance.

Ladret and Guérin-Dugué [44] use DCT (Discrete Cosinus Transform) as the feature representation for a *k*-NN classifier for outdoor urban scenes, indoor scenes, 'closed' landscapes (e.g. valleys, mountains, forests, etc.) and 'open' landscapes (e.g. deserts, fields, beaches, etc.) classifications. Iqbal and Aggarwal [45] propose structure-based features (shape-like) to represent images which contain manmade objects, such as buildings, towers, bridges, and other architectural objects. Then, a *k*-NN classifier is used to classify images into structure, non-structure, and intermediate classes. In Cheng and Chen [46], images are segmented based on color and texture features and the *k*-NN classifier is used for 20-category classification.

4.6. Template Matching

Template matching is one of the simplest and earliest approaches to pattern recognition. Matching is used to determine the similarity between two entities (such as points, curves or shapes) of the same type. The pattern/feature to be classified is matched against the stored template (typically, a 2-D shape). The template itself is often learned from the training set [47].

Smith and Li [48] propose composite region templates to classify regions of images into ten categories. Lipson *et al*. [49] present the configural recognition approach for encoding scene class structure, and construct class templates

for snowy mountains, snowy mountains with lakes, fields, and waterfalls by using local color features.

5. ADVANCED LEARNING TECHNIQUES

5.1. Combination of Multiple Classifiers

In pattern recognition and machine learning, the combination of a number of classifiers has recently been an active research area [50, 51]. It can be called ensemble classifiers or modular classifiers. Ensemble classifiers aim at obtaining highly accurate classifiers by combining less accurate ones. They are proposed to improve the classification performance of a single classifier [52, 53]. That is, the combination is able to complement the errors made by the individual classifiers on different parts of the input space. Therefore, the performance of modular classifiers is likely better than the one of the best single classifier used in isolation [28].

5.1.1. Majority Voting

The simplest method to combine classifiers is majority voting. The binary outputs of the *k* individual classifiers are pooled together. Then, the class which receives the largest number of votes is selected as the final classification decision [53]. In general, the final classification decision that reaches the majority of $\frac{k+1}{2}$ votes is taken. Figure **7** shows the general architecture of a classifier ensemble [23]. A number of differently trained neural networks (i.e. experts) share the input and whose outputs are combined to produce an overall output. Note that the experts can be trained by different examples (or different features) of a given training set or different learning models trained by the same training set.

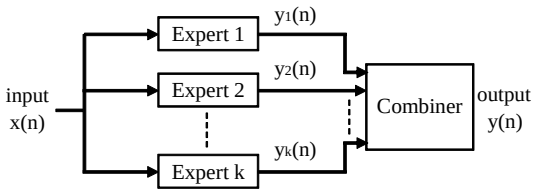

Fig. (7). Architecture of a classifier ensemble.

The earliest related work using ensemble classifiers for image classification is by Szummer and Picard [54]. First of all, each image is segmented by a fixed number of blocks, and color and texture features of each block are extracted. Then, a color and texture *k*-NN classifier is designed to classify the color and texture features of each block into indoor and outdoor categories individually. Finally, the final output is based on the blocks of an image which have the highest vote for one of the indoor and outdoor categories.

In Chang *et al.* [55], an ensemble of Bayes networks is trained to give multiple soft labels, i.e. class membership to an image and then, the most *correct* label(s) which have higher confidence values can be decided from these labels. Schettini *et al.* [56] design ensemble CARTs to classify color, texture, and shape features of images into indoor, outdoor and close-up categories. For each image the final label is assigned by majority vote.

5.1.2. Bagging

In bagging, several networks are trained independently by different training sets via the bootstrap method [57]. Bootstrapping builds *k* replicate training data sets to construct *k* independent networks by randomly re-sampling the original given training dataset, but with replacement. That is, each training example may appear repeated but not at all in any particular replicate training data set of *k*. Then, the *k* networks are aggregated via an appropriate combination method, such as majority voting [58]. Figure **8** shows the block diagram of the SVM ensemble.

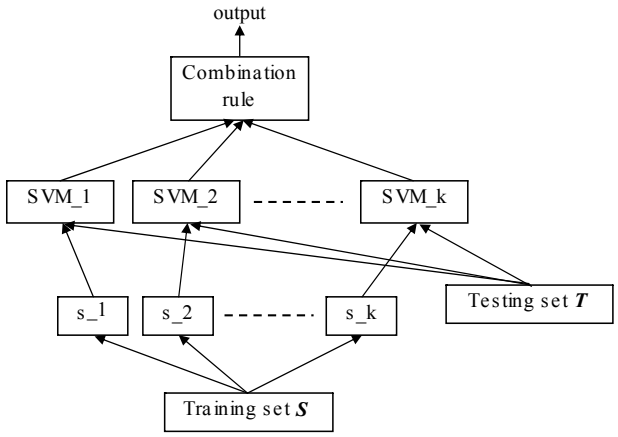

Fig. (8). Architecture of the SVM ensemble.

Li *et al.* [59] propose a confidence-based dynamic ensemble by *bagging* SVMs to improve the performance of image annotation based on traditional *static* classifiers. Color and texture image features are used to train the ensemble SVMs for 116-category classification. In Feng and Chua [60], the bootstrapping approach is used to co-train two different features extracted from two segmentation methods by using SVMs. However, their classification scale and averaged classification accuracy are not reported.

5.1.3. Boosting

In boosting, like bagging, each network is trained using a different training set. However, the *k* networks are trained not in a parallel and independent way, but sequentially. The original boosting approach, *boosting by filtering*, was proposed by Schapire [61]. It attempts to 'boost' the accuracy of any given classifier (or learning algorithm).

AdaBoost is a combination of the ideas behind boosting and bagging and does not demand a large training data set. Initially, each training example of a given training set has the same weight. For training the *k*-th network as a *weak learning model*, *n* sets of training samples ($n<m$) among *S* are used to train the *k*-th network. Then, the trained network is evaluated by *S* to identify those training examples which cannot be classified correctly. The *k*+1 network is then trained by a modified training set which boosts the importance of those incorrectly classified examples. This

sampling procedure will be repeated until *K* training samples is built for constructing the *K*th network. Therefore, the final decision is based on the weighted vote of the individual classifiers [62].

In related literature, Howe [63] compares feature and vector based boosting methods using *k*-NN to classify different color histogram representations into one of five categories.

5.1.4. Stacked Generalization

Stacked generalization (or stacking) was proposed by Wolpert [64]. The outputs of individual classifiers as the level-0 generalizers are used to train the 'stacked' classifier as the level-1 generalizer. The final decision is made based on the output of the stacked classifier. While combiners like majority voting, are *static* combiners, the stacked classifier is a *trainable* combiner. Figure **9** shows the architecture of two-level stacked generalization [28]. The level-0 generalizers use partitioning of the training set for training (e.g. *m* networks trained by *m* sets of the training data). Then, the level-1 generalizer is trained based on the outputs of level-0 generalizer tested by the original training set for target outputs. It is a scheme for estimating the errors of a generalizer/classifier when working on a particular learning set, and then correcting those errors.

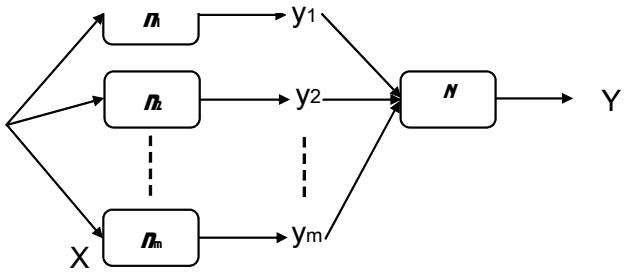

Fig. (9). Two-level stacked generalization.

Serrano *et al*. [65] use color and texture SVMs to classify color and texture features of 16 blocks per image into indoor and outdoor classes individually. Then, two decision values associated with each block of an image assigned by the color and texture SVMs are further used to train a new SVM. The classification result is better than majority voting suggested in Szummer and Picard [54].

In Tsai *et al*. [34], the CLAIRE image annotation system based on the idea of stacked generalization is developed. The first-level classifiers focus on classifying color and texture features respectively and the outputs of the classifiers are inputted into the second-level classifier for the final classification.

5.2. Hybrid Learning Models

A hybrid learning model can be based on the combination of clustering and classification techniques. Clustering can be used as a pre-processing stage to identify pattern classes for subsequent supervised classification [66]. That is, clustering is used for pre-classification of unlabelled collections. During clustering, the unlabelled data (or sometimes called training data) are clustered, and the centre of each cluster (or some chosen data in each cluster) can be used as training examples to train a classification model.

In Barnard *et al*. [8], a well-known and widely referenced work, proposes image annotation as a machine translation problem. First of all, a number of segmented blobs as regions of images in the training set are clustered based on color, texture, shape, and position features. At the same time, the words associated with their images in the training set are tokenized. Only those clusters which contain a certain large number of patterns (i.e. training examples of regions) are considered to represent potential classes. That is, at the beginning there are 500 clusters defined as a priori, and the most common words (above 20 times) correspond to the clusters are 155. Then, the frequencies of words of all regions in each cluster are accumulated and the likelihood for every word corresponding to the cluster is then calculated to produce a probability table based on the Expectation-Maximization (EM) clustering algorithm. It is used to learn a lexicon to link blob tokens with work tokens. On a given test image, words are predicted for each blob using the probability table, which acts like template matching.

In Tsai *et al*. [17], a clustering method was used to cluster the given feature vectors. Then each *winner* (i.e. centre) of the clusters was used to train SVMs for 25-category classification. Some other related work using clustering for pre-classification, such as Jeon *et al*. [67], is proposed for further improvement following Barnard *et al*. [8]. Instead of using the EM algorithm to estimate the probability of image regions and their labels, they assume that the image annotation problem can be viewed as analogous to the cross-lingual retrieval problem. However, Jeon and Manmatha [68] and Lavrenko *et al*. [69] extend the work of Jeon *et al*. [67] to show that using supervised learning classification alone provides much better performance than previous work of using the clustering stage for pre-classification. In addition, Monay and Gatica-Perez [18] show that supervised learning provides better image annotation performances than unsupervised learning.

6. COMPARISONS OF RELATED WORK

This section provides some comparisons of related image annotation systems. These comparisons shown in Table 1 and 2 focus on using image (low-level) processing and supervised learning models respectively [70-93].

These comparisons do not consider the system performance, i.e. classification or annotation accuracy. This is because using different features, classifiers, and/or datasets within different numbers of training and test sets usually causes different results. Regarding the same approaches/ systems published in different years, we only selects the newest one for comparisons.

According to Table **1**, 50 image annotation systems have been compared and there are some issues which can be identified as follows:

- 21 systems use block-based image features for indexing and each image is only classified into one specific category, and 15 systems uses object-based segmentation. There are 34 systems which assign one keyword

Table 1. Comparisons of Image Processing

Work	Segmentation	Low-Level Features	Dimensions
2006			
Tsai *et al*. [34]	B	C + T	19
Wong and Hsu [70]	R	S	14
2005			
Carneiro and Vasconcelos [71]	G	C	32
Gao and Fan [72]	R	C + T	83
Ghoshal *et al.* [25]	B	C + T + E	30
Goh *et al*. [73]	G	C + T	144
2004			
Boutell *et al.* [74]	B	C	294
Chen and Wang [75]	R	C + T + S	9
Cusano *et al.* [76]	B	C	117
Fan *et al.* [77]	R	C + T + S	56
Jeon and Manmatha [68]	B	C + T	N/A
Kim *et al*. [32]	R	C	N/A
Le Saux and Amato [78]	R	C	N/A
Li *et al*. [79]	R	C + T	N/A
Park *et al.* [31]	R	C + T	49
Qiu *et al.* [80]	G	C	256
Schettini *et al.* [56]	B	C + T + S	68
Shi *et al.* [81]	R	C + T	19
Vogel and Schiele [82]	B	C + T	156
2003			
Blei and Jordan [27]	R	C + T + S + SR	N/A
Boutell *et al.* [83]	B	C + SR	294
Chang *et al.* [55]	G	C + T	144
Cheng and Chen [46]	B	C + T	512
Feng and Chua [60]	R	N/A	N/A
Jin *et al.* [84]	B	C	80
Lavrenko *et al*. [69]	R	C + T + S	36
Li *et al*. [59]	G	C + T	144
Lim *et al*. [30]	B	C + T	N/A
Yanai [85]	B	C	75
2002			
Andrew *et al*. [86]	R	C + T + S	N/A
Brank [39]	G	C	1024
Howe [63]	G	C	512

(Table 1) Contd…

Work	Segmentation	Low-Level Features	Dimensions
Iqbal and Aggarwal [45]	G	S	3
Kuroda and Hagiwara [29]	R	C + T	37
Monadjemi *et al*. [87]	B	T	8
Serrano *et al*. [65]	B	C + T	55
Wang and Li [26]	B	C + T	6
2001			
Goh *et al*. [88]	G	C + T	144
Ladret and Guérin-Dugué [44]	B	T	64
Luo and Savakis [89]	B	C + T	512
Vailaya *et al*. [24]	B	C + S	640
Teytaud and Sarrut [90]	G	C	N/A
2000			
Aksoy and Haralick [91]	N/A	T	60
Bradshaw [92]	B	C + T	26
1999			
Chapelle *et al*. [38]	G	C	4096
Setchell and Campbell [93]	R	C + T	30
Smith and Li [48]	B	C	166
1998			
Huang *et al*. [43]	G	C	512
Szummer and Picard [54]	B	C + T	512
1997			
Lipson *et al*. [49]	G	C + SR	N/A

(Note: Segmentation: 'G' means global descriptors (i.e. non-segmentation); 'B' for block-based segmentation; 'R' for region-based segmentation;
Features: 'C' means color; 'T' texture; 'S' shape; 'SR' spatial relationship; 'E' for edge)

Table 2. Comparisons of Classifier Design

Work	Classifier	No. of classes	Keywords per image	DataSet
2006				
Tsai *et al*. [34]	Ensemble SVMs	60	5	Corel
Wong and Hsu [70]	SVMs	40	1	Others
2005				
Carneiro and Vasconcelos [71]	Bayes	371	1~5	Corel
Gao and Fan [72]	SVMs	19	N/A	Corel
Ghoshal *et al*. [25]	Hidden Markov Model	375 (Corel); 75 (TRECVID)	24	Corel + TRECVID
Goh *et al*. [73]	Ensemble SVMs	116	1	Corel

(Table 2) Contd…

Work	Classifier	No. of classes	Keywords per image	DataSet
2004				
Boutell *et al.* [74]	SVMs	6	1~3	Corel + others
Chen and Wang [75]	SVMs	20	1	Corel
Cusano *et al.* [76]	SVMs	7	1	WWW
Fan *et al.* [77]	SVMs	32	N/A	Corel + WWW
Jeon and Manmatha [68]	Bayes	125	1~5	Corel
Kim *et al.* [32]	MLP	2	1	Corel
Le Saux and Amato [78]	SVMs	5	1	Others
Li *et al*. [79]	Bayes	18	N/A	Others
Park *et al*. [31]	MLP	30	1	WWW
Qiu *et al.* [80]	SVMs	10	1	Corel
Schettini *et al.* [56]	Ensemble CART	3	1	WWW
Shi *et al.* [81]	SVMs	23	N/A	Corel + others
Vogel and Schiele [82]	k-NN	9	N/A	Others
2003				
Blei and Jordan [27]	Latent Dirchl *et al.* location	N/A	2~4	Corel
Boutell *et al.* [83]	SVMs	2	1	Corel + others
Chang *et al*. [55]	Ensemble Bayes	116	1	Corel + WWW
Cheng and Chen [46]	k-NN	20	1	Corel
Feng and Chua [60]	Ensemble SVMs	N/A	N/A	Corel
Jin *et al*. [84]	SVMs	6	1	Corel
Lavrenko *et al.* [69]	Bayes	153	1~5	Corel
Li *et al*. [59]	Ensemble SVMs	116	1	Corel + WWW
Lim *et al*. [30]	MLP	26	1	Kodak Photo
Yanai [85]	k-NN	50	1	Corel + WWW
2002				
Andrew *et al.* [86]	SVMs	3	1	Corel
Brank [39]	SVMs	14	1	Others
Howe [63]	Ensemble k-NNs	5	1	Corel
Iqbal and Aggarwal [45]	k-NN	3	1	Others
Kuroda and Hagiwara [29]	MLP	15	2	Others
Monadjemi *et al.* [87]	Ensemble MLP	4	1	Others
Serrano *et al.* [65]	Ensemble SVMs	2	1	Others
Wang and Li [26]	2-D Hidden Markov Model	600	3.6	Corel

(Table 2) Contd…

Work	Classifier	No. of classes	Keywords per image	DataSet
2001				
Goh *et al*. [88]	SVMs	15	1	Corel
Ladret and Guérin-Dugué [44]	k-NN	4	1	Corel
Luo and Savakis [89]	Bayes	2	1	Kodak Photo
Vailaya *et al*. [24]	Bayes	3	1	Corel + others
Teytaud and Sarrut [90]	SVMs	14	1	Corel
2000				
Aksoy and Haralick [91]	Bayes	18	1	Corel
Bradshaw [92]	Bayes	2	1	Corel
1999				
Chapelle *et al*. [38]	SVMs	7	1	Corel
Setchell and Campbell [93]	MLP	11	4	Others
Smith and Li [48]	Template	10	N/A	Others
1998				
Huang *et al*. [43]	Decision Trees	11	1	Corel
Szummer and Picard [54]	Ensemble k-NNs	2	1	Others
1997				
Lipson *et al*. [49]	Template	4	1	Corel

- to each image, but only 8 systems consider assigning multiple keywords to an image by using either region-based or local block-based image features.
- 14 systems only use color features for high-level concept learning and classification under the problem scale between 2 to 50 categories. The majority, i.e. 19 systems, uses color and texture features for the problem scale between 2 to 125 categories. Only 6 systems consider other features such as shape in addition to color and texture features.
- As the problem scale increases, i.e. larger numbers of categories, only Tsai *et al*. [34] reports the number of (un) predictable classes since there should be some categories which are difficult to classify.
- SVMs and ensemble classifiers have attracted much more attention recently. That is, 18 systems use SVMs and ensemble classifiers; k-NN, and naïve Bayes are used for 7 and 8 systems respectively. Only one system uses decision trees and two uses template matching method.
- 33 systems use Corel as their ground truth dataset.

7. FUTURE RESEARCH ISSUES

7.1. Large Scale Image Annotation

As most supervised classifiers are only designed for small scale problems, i.e. classifying small numbers of categories, it is still an open research problem to construct a large scale learning machine/classifier. When a certain large number of categories is considered, both concrete (*ofness*) and abstract (*aboutness*) concepts may be trained to annotate images for concept-based image retrieval. Three levels of image retrieval are distinguished by Eakins [94] and Eakins and Graham [95]. Level 1 is based on primitive features of images, such as color, texture, and shape involving lowest degree of abstraction. Level 2 is based on retrieving objects identified within an image and involves some degree of logical inference. Level 3 is based on abstract attributes which involves a high degree of reasoning about the meaning and purpose of the scenes depicted, such as named events or types of activity and/or particular mood or possessing artistic, religious or symbolic significance. Most users want to retrieve images by locating images of a particular type or individual instance of an object, phenomenon, event, i.e. retrieval at Levels 2 (mostly) and 3. However, most of existing image annotation systems described above only allow Level 2 retrieval.

In addition, the more categories need to be classified, the more ambiguous categories exist (e.g. *building*/*home*, *men*/*women*/*children*, *rural*/*park*/*garden*, etc.) and the more challenging it is. Many conceptually different categories are visually similar in the feature space which may cause *feature overlapping* and thus degrades generalization. Specifically, we can categorize images into two types of 'concrete' and 'abstract' categories, in which 'concrete' means a physical

object or entity (e.g. trees and cars) and 'abstract' means abstraction, human activity, or an assemblage of multiple physical objects/entities (e.g. festival and happy). Therefore, it would be useful if we could understand the performance of both types of categories under large scale problems.

7.2. User-Centered Evaluation

In general, image annotation evaluation is based on some chosen ground truth dataset. That is, the test images have been annotated manually like Corel in which each image is associated with a specific theme (or category). Then, image annotation systems can be evaluated by assessing whether the keywords assigned to the test images exactly match their original categories. Although the keywords associated with images are labeled by professional indexers, such as the Corel dataset, those keywords are unlikely to fully and exactly describe image contents similarly to *real world users*, especially as each image is only classified into one specific theme. That is, using the ground truth answer of Corel cannot make fair conclusion about the system performance. In addition, image retrieval systems are designed for real users. As a result, user-centered evaluation is necessary in order to fully understand the performance of an image annotation system. However, very few studies consider using user-cantered evaluation to validate image annotation systems.

User-centered evaluation can be classified into two types. First of all, keywords associated with their images can be selected as relevant or not by the judges. Then, the conclusion can be drawn from the analysis of the collected *qualitative data*. Wang *et al*. [96] propose a hierarchical scheme (e.g. *Paris* is under *Europe*) for human subjects to judge their image annotation system. However, they do not report the real testing by some chosen human judges.

The second type of evaluation is to ask people to annotate a given set of images. If the results show a certain degree of consistency, they can be treated as ground truth. Then, a system can be tested by the same dataset given to the human subjects and its outputs can be compared with the ground truth. Note that the difference between the ground truth dataset, such as Corel and the one of human annotation here is that the former is based on some selected professional indexers, but the latter naïve/real users. In other words, one obtains a different 'ground truth' depending on whether the data is derived from the viewpoint of a professional indexer, or from the viewpoints of more general classes of users. There is only one related work, i.e. Schettini *et al*. [56], which asks five human subjects to label 9000 images for indoor, outdoor, and close-up categories to assess their image classification system. The final label of each image is decided by majority vote. In Tsai *et al*. [97], two types of evaluation are considered to assess image annotation systems.

8. CURRENT AND FUTURE DEVELOPMENTS

Bridging the semantic gap for image retrieval is a very hard problem to solve. In the context of automatic image annotation, the major difficulty is to make computers understand image content in terms of high-level concepts or semantics, which is closely related to the problem of computer vision and object recognition.

To bridge the semantic gap between low-level features and high-level concepts could be approached by image classification. A learning machine or classifier is trained by learning low-level features for classifying images into some conceptual categories. The classification process can be thought of as image annotation.

Related work has shown the applicability of using machine learning techniques for automatic image annotation in a small number of conceptual categories. However, to make image annotation perform ideally still has a long way to go. One technical future research issue is to try to investigate current image annotation systems and/or to develop new image annotation systems for classifying large numbers of conceptual categories. In order to *fully* understand the performance of image annotation systems, another research issue is to consider user-centered evaluation to validate image annotation systems in addition to using some chosen ground truth data set as system-centered evaluation.

REFERENCES

[1] Rui Y, Huang TS, Chang S-F. Image retrieval: current techniques, promising directions and open issues. J Visual Comm Image Rep 1999; 10(1): 39-62.

[2] Smeulders AWM, Worring M, Santini S, Gupta A, Jain R. Content-based image retrieval at the end of the early years. IEEE Trans Patt Anal Machi Intel 2000; 22(12): 1349-1380.

[3] Flickner M, Sawhney H, Niblack W, *et al*. Query by image and video content: the QBIC system. IEEE Computer 1995; 28(9): 23-32.

[4] Smith JR, Chang S-F. VisualSEEk: a fully automated content-based image query system. *Proceedings of the ACM International Conference on Multimedia*, Boston, Massachusetts, Nov. 18-22, 1996; 87-98.

[5] Pentland A, Picard RW, Sclaroff S. Photobook: content-based manipulation of image databases. Int J Comp Vision 1996; 18(3): 233-254.

[6] Eakins JP. Towards intelligent image retrieval. Patt Recog 2002; 35: 3-14.

[7] Enser PGB. Visual image retrieval: seeking the alliance of concept-based and content-based paradigms. J Info Sci 2000; 26(4):199-210.

[8] Barnard K, Duygulu P, Forsyth D, de Freitas N, Blei D, Jordan MI. Matching words and pictures. J Mach Learn Res 2003; 3: 1107-1135.

[9] Antani S, Kasturi R, Jain R. A survey on the use of pattern recognition methods for abstraction, indexing and retrieval of images and video. Patt Recog 2002; 35: 945-965.

[10] Swain MJ, Ballard DH. Color indexing. Int J Comp Vision 1999; 7(1): 11-32.

[11] Laaksonen J., Koskela M., Oja E. Content-based image retrieval using self-organizing maps. *Proceedings of the International Conference on Visual Information Systems*, Amsterdam, Netherlands, June 2-4, 1999; 541-548.

[12] Meghini C, Sebastiani F, Straccia U. A model of multimedia information retrieval. J ACM, 2001; 48(5): 909-970.

[13] Schettini R, Ciocca G, Zuffi S. A Survey on methods for color image indexing and retrieval in image databases. In Color Imaging Science*:* Exploiting digital media, Luo, R. and MacDonald, L. (Eds.), John Wiley, England. 2001.

[14] Ko B, Lee H-S, Byun, H. Image retrieval using flexible image subblocks. Proceedings of the ACM Symposium on Applied Computing, Como, Italy, March 19-21, 2000; 574-578.

[15] Shyu, M.-L., Chen, S.-C., Chen, M., Zhang, C., and Sarinnapakorn, K. Image database retrieval utilizing affinity relationship. *Proceedings of the 1st ACM International Workshop on Multimedia Databases*, New Orleans, Louisiana, Nov. 7, 2003; 78-85.

[16] Zhao R., Grosky WI. From features to semantics: some preliminary results. *Proceedings of the IEEE International Conference on Multimedia & Expo*, New York City, New York, July 30-Aug. 2, 2000; 679-682.

[17] Tsai C-F, McGarry K, Tait J. Image classification using hybrid neural networks. *Proceedings of the 26th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Toronto, Canada, July 28-Aug. 1, 2003; 431-432.

[18] Monay F., Gatica-Perez D. On image auto-annotation with latent space models. *Proceedings of the ACM International Conference on Multimedia*, Berkeley, California, Nov. 2-8, 2003; 275-278.

[19] Deng Y., Manjunath BS., Shin H. Color image segmentation. *Proceeding of the IEEE International Conference on Computer Vision and Pattern Recognition*, Ft. Collins, Colorado, June 23-25 1999; 2446-2451.

[20] Jahne B. *Digital Image Processing: Concepts, Algorithms, and Scientific Applications*. Springer-Verlag, Berlin 1995.

[21] Aslandogan, YA, Yu, CT. Techniques and systems for image and video retrieval. IEEE Transactions on Knowledge and Data Engineering 1999; 11(1): 56-63.

[22] Mitchell T. Machine Learning. McGraw Hill, New York. 1997.

[23] Haykin S. Neural networks: a comprehensive foundation, 2nd Edition. Prentice Hall, New Jersey. 1999

[24] Vailaya A, Figueiredo AT, Jain AK, Zhang H-J. Image classification for content-based indexing. IEEE Trans Image Proc 2001; 10(1): 117-130.

[25] Ghoshal A, Ircing P, Khudanpur S. Hidden Markov models for automatic annotation and content-based retrieval of images and video. *Proceedings of the 28th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Salvador, Brazil, Aug 2005; 15-19: 544-551.

[26] Wang JZ, Li J. Learning-based linguistic indexing of pictures with 2-D MHMMs. *Proceedings of the ACM International Conference on Multimedia*, Juan-les-Pins, France, Dec. 1-6, 2002; 436-445.

[27] Blei, DM, Jordan, MI. Modeling annotated data. *Proceedings of the 26th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Toronto, Canada, July 28 – Aug. 1, 2003; 127-134.

[28] Bishop CM. Neural networks for pattern recognition. Oxford University Press, Oxford. 1995

[29] Kuroda K, Hagiwara M. An image retrieval system by impression words and specific object names - IRIS. Neurocomputing, 2002; 43(1-4): 259-276.

[30] Lim J-H, Tian Q, Mulhem P. Home photo content modeling for personalized event-based retrieval. IEEE Multimedia 2003; 10(4): 28-37.

[31] Park SB, Lee JW, Kim SK. Content-based image classification using a neural network. *Pattern Recognition Letters* 2004; 25: 287-300.

[32] Kim, S., Park, S., Kim, M. Image classification into object / non-object classes. *Proceedings of the International Conference on Image and Video Retrieval*, Dublin, Ireland, July 21-23, 2004; 393-400.

[33] Vapnik V. Statistical Learning Theory. John Wiley, New York, 1998.

[34] Tsai C-F, McGarry K, Tait JC. A modular support vector image indexing and classification system. ACM Transactions on Information Systems 2006a; 24(3): 353-379.

[35] Cristianini N, Shawe-Taylor J. An introduction to support vector machines. Cambridge University Press, Cambridge 2000.

[36] Cortes C, Vapnik V. Support vector networks. Mach Learn 1995; 20: 273-297.

[37] Byun H, Lee S-W. A survey on pattern recognition applications of support vector machines. Int J Pat Recog Art Intel 2003; 17(3): 459-486.

[38] Chapelle O, Haffner P, Vapnik VN. Support vector machines for histogram-based image classification. IEEE Trans Neural Networks 1999; 10(5): 1055-1064.

[39] Brank J. Using image segmentation as a basis for categorization. Informatica, 2002; 26: 4.

[40] Huang J., Kumar SR., Mitra M., Zhu W-J., Zabih R. Image indexing using color correlograms. *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, San Juan, Puerto Rico, June 17-19, 1997; 762-768.

[41] Quinlan JR. Induction of decision trees. Mach Learn 1986; 1(1): 81-106.

[42] Breiman L, Friedman JH, Olshen RA, Stone PJ. Classification and regressing trees. Wadsworth International Group, California. 1984.

[43] Huang J., Kumar SR., Zabih R. An automatic hierarchical image classification. *Proceedings of the ACM International Conference on Multimedia*, Bristol, UK, Sep. 12-16, 1998; 219-228.

[44] Ladret P, Guérin-Dugué A. Categorisation and retrieval of scene photographs from a JPEG compressed database. Pattern Analysis & Applications 2001; 4(2-3): 185-199.

[45] Iqbal Q, Aggarwal JK. Retrieval by classification of images containing large manmade objects using perceptual grouping. Patt Recog 2002; 35(7): 1463-1479.

[46] Cheng Y-C, Chen S-Y. Image classification using color, texture and regions. Image Vision Comput 2003; 21(9): 759-776.

[47] Jain AK, Duin RPW, Mao J. Statistical pattern recognition: a review. IEEE Trans Patt Anal Mach Intel 2000; 22(1): 4-37.

[48] Smith JR, Li C-S. Image classification and querying using composite region templates. Comp Vision Image Underst 1999; 75(1/2): 165-174.

[49] Lipson P., Grimson E., Sinha P. Configuration based scene classification and image indexing. *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, San Juan, Puerto Rico, June 17-19, 1997; 1007-1013.

[50] Ghosh, J. Multiclassifier systems: back to the future. *Proceedings of the 3rd International Workshop on Multiple Classifier Systems*, Cagliari, Italy, June 24-26, 2002; 1-15.

[51] Frosyniotis D, Stafylopatis A, Likas A. A divide-and-conquer method for multi-net classifiers. J Patt Anal Appl 2003; 6(1): 32-40.

[52] Ho TK, Hull JJ, Srihari SN. Decision combination in multiple classifier systems. IEEE Trans Patt Anal Mach Intel *1994;* 16(1): 66-75.

[53] Duin RPW, Kittler J, Hater M, Mates J. On combining classifiers. IEEE Trans Patt Anal Mach Intel 1998; 20(3): 226-239.

[54] Szummer M., Picard RW. Indoor-outdoor image classification. *Proceedings of the IEEE ICCV Workshop on Content-based Access of Image and Video Databases*, Bombay, India, Jan. 3, 1998; 42-51.

[55] Chang E, Kingshy G, Sychay G, Wu G. CBSA: Content-based soft annotation for multimodal image retrieval using Bayes point machines. IEEE Trans Circ Syst Video Tech, Special Issue on Conc Dynam Aspec Multimedia Content Description 2003; 13(1): 26-38.

[56] Schettini R, Brambilla C, Cusano C, Ciocca G. Automatic classification of digital photographs based on decision forests. Int J Patt Recog Art Intel 2004; 18(5): 819-845.

[57] Breiman, L. (1996) Bagging predictors. Machine Learning 1996; 24(2): 123-140.

[58] Kim H-C, Pang S, Je H-M, Kim D, Bang SY. Constructing support vector machine ensemble. Patt Recog 2003; 36(12): 2757-2767.

[59] Li B., Goh K., Chang EY. Confidence-based dynamic ensemble for image annotation and semantics discovery. *Proceedings of the ACM International Conference on Multimedia*, Berkeley, California, Nov. 2-8, 2003; 528-538.

[60] Feng H., Chua T-S. A bootstrapping approach to annotating large image collection. *Proceedings of the ACM Multimedia Workshop on Multimedia Information Retrieval*, Berkeley, California, Nov. 7, 2003; 55-62.

[61] Schapire RE. The strength of weak learnabilty. Mach Learn 1990; 5(2): 197-227.

[62] Freund Y., Schapire RE. Experiments with a new boosting algorithm. *Proceedings of the International Conference on Machine Learning*, Bari, Italy, July 3-6, 1996; 148-156.

[63] Howe NR. Boosted image classification: an empirical study. *Proceedings of the ICML Workshop on Machine Learning in Computer Vision*, Sydney, Australia, July, 2002; pp.

[64] Wolpert DH. Stacked generalization. Neural Networks 1992; 5(2): 241-259.

[65] Serrano N., Savakis A., Luo J. A computationally efficient approach to indoor/outdoor scene classification. *Proceedings of the IEEE International Conference on Pattern Recognition*, Quebec, Canada, Aug. 11-15, 2002; 146-149.

[66] Jain AK, Murty MN, Flynn PJ. Data clustering: a review. ACM Comp Surv 1999; 31(3): 264-323.

[67] Jeon J., Lavrenko V., Manmatha R. Automatic image annotation and retrieval using cross-media relevance models. *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Toronto, Canada, July 28-Aug. 1, 2003; 119-126.

[68] Jeon J., Manmatha R. Using maximum entropy for automatic image annotation. *Proceedings of the International Conference on Image and Video Retrieval*, Dublin, Ireland, July 21-23, 2004; 24-32.

[69] Lavrenko V., Manmatha R., Jeon J. A model for learning the semantics of pictures. *Proceedings of the Intl Conference on Neural Information Processing Systems*, Vancouver, Canada, Dec. 8-13. 2003.

[70] Wong W-T, Hsu S-H. Application of SVM and ANN for image retrieval. Eur J Operat Res 2006; 173: 938-950.

[71] Carneiro G., Vasconcelos N. A database centric view of semantic image annotation and retrieval. *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Salvador, Brazil, Aug. 15-19, 2005; 559-566.

[72] Gao Y., Fan J. Semantic image classification with hierarchical feature subset selection. *Proceedings of the ACM Multimedia Workshop on Multimedia Information Retrieval*, Hilton, Singapore, Nov. 10-11, 2005; 135-142.

[73] Goh K-S, Chang EY, Li B. Using one-class and two-class SVMs for multiclass image annotation. IEEE Trans Know Data Eng 2005; 17(10): 1333-1346.

[74] Boutell M, Luo J, Shen X, Brown CM. Learning multi-label scene classification. Pattern Recognition 2004; 37(9): 1757-1771.

[75] Chen Y, Wang JZ. Image categorization by learning and reasoning with regions. J Mach Learn Res 2004; 5: 913-939.

[76] Cusano C., Ciocca G., Schettini R. Image annotation using SVM. *Proceedings of SPIE Conference on Internet Imaging V*, San Jose, California, Jan. 19-20, 2004; 5304: 330-338.

[77] Fan J., Gao Y., Luo H., Xu, G. Automatic image annotation by using concept-sensitive salient objects for image content representation. *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Sheffield, UK, July 25-29, 2004; 361-368.

[78] Le Saux B., Amato G. Image recognition for digital libraries. *Proceedings of the ACM Multimedia Workshop on Multimedia Information Retrieval*, New York, Oct. 15-16, 2004.

[79] Li Y., Bilmes JA., Shapiro LG. Object class recognition using images of abstract regions. *Proceedings of the International Conference on Pattern Recognition*, Cambridge, UK, Aug. 23-26. 2004.

[80] Qiu G, Feng X, Fang J. Compressing histogram representations for automatic color photo categorization. Patt Recog 2004; 37(11): 2177-2193.

[81] Shi R., Feng H., Chua T-S., Lee C-H. An adaptive image content representation and segmentation approach to automatic image annotation. *Proceedings of the International Conference on Image and Video Retrieval*, Dublin, Ireland, July 21-23, 2004; 545-554.

[82] Vogel J., Schiele B. Natural scene retrieval based on a semantic modelling step. *Proceedings of the International Conference on Image and Video Retrieval*, Dublin, Ireland, July 21-23, 2004; 207-215.

[83] Boutell M., Luo J., Gray RT. Sunset scene classification using simulated image recomposition. *Proceedings of the IEEE International Conference on Multimedia & Expo*, Baltimore, Maryland, July 6-9, 2003.

[84] Jin R., Hauptmann AG., Yan R. Image classification using a bigram model. *Proceedings of AAAI Spring Symposium on Intelligent Multimedia Knowledge Management*, Palo Alto, California, March 24-26, 2003; 83-87.

[85] Yanai K. Generic image classification using visual knowledge on the Web. *Proceedings of the ACM International Conference on Multimedia*, Berkeley, California, Nov. 2-8, 2003; 167-176.

[86] Andrews S., Tsochantaridis T., Hofmann, T. Support vector machines for multiple-instance learning. *Proceedings of the International Conference on Neural Information Processing Systems*, Vancouver, Canada, Dec. 9-14, 2002

[87] Monadjemi A., Thomas BT., Mirmehdi M. Classification in high resolution images with multiple classifiers. *Proceedings of the IASTED International Conference on Visualization, Imaging, and Image Processing* 2002; 417-421.

[88] Goh K-S., Chang EY., Cheng K-T. Support vector machine pairwise classifiers with error reduction for image classification. *Proceedings of the ACM Multimedia Workshop on Multimedia Information Retrieval*, Ontario, Canada, Oct. 5, 2001; 32-37.

[89] Luo J., Savakis A. Indoor vs outdoor classification of consumer photographs using low-level and semantic features. *Proceedings of the IEEE International Conference on Image Processing*, Thessaloniki, Greece, Oct. 7-10, 2001; 745-748.

[90] Teytaud O., Sarrut D. Kernel based image classification. *Proceedings of the International Conference on Artificial Neural Networks*, Vienna, Austria, Aug. 21-25, 2001; 369-375.

[91] Aksoy S, Haralick RM. Using texture in image similarity and retrieval. In *Texture Analysis in Machine Vision*, Pietikainen, M. and Bunke, H. (Eds.), World Scientific, Singapore 2000; 20.

[92] Bradshaw B. Semantic based image retrieval: a probabilistic approach. *Proceedings of the ACM International Conference on Multimedia*, Los Angeles, California, Oct. 30-Nov.4, 2000; 167-176.

[93] Setchell CJ., Campbell NW. Using color Gabor texture features for scene understanding. *Proceedings of the IEEE International Conference on Image Processing and Its Applications*, Manchester, UK, July 13-15, 1999; 372-376.

[94] Eakins JP. Automatic image content retrieval – how usable is the technology. Int J Elect Lib Res 1997; 7(1): 63-88.

[95] Eakins JP, Graham M. Content-based image retrieval. *A Report to the JISC Technology Applications Programme*, University of Northumbria at Newcastle. [On-line] 1999. Available from: http://www.unn.ac.uk/iidr/report.html

[96] Wang JZ, Li J, Lin SC. Evaluation strategies for automatic linguistic indexing of pictures. *Proceedings of the International Conference on Image Processing*, Barcelona, Spain, Sep. 14-17, 2003; 617-620.

[97] Tsai C-F, McGarry K, Tait J. Qualitative Evaluation of Automatic Assignment of Keywords to Images. Info Proc Manag 2006b; 42(1): 136-154.