2006

Image content annotation based on visual features

Lei Ye

University of Wollongong, lei@uow.edu.au

Philip Ogunbona

University of Wollongong, philipo@uow.edu.au

J. Wang

University of Wollongong

Follow this and additional works at: <https://ro.uow.edu.au/infopapers>

Part of the [Physical Sciences and Mathematics Commons](#)

Recommended Citation

Ye, Lei; Ogunbona, Philip; and Wang, J.: Image content annotation based on visual features 2006.
<https://ro.uow.edu.au/infopapers/534>

Image content annotation based on visual features

Abstract

Automatic image content annotation techniques attempt to explore structural visual features of images that describe image content and associate them with image semantics. In this paper, two types of concept spaces, atomic concept and collective concept spaces, are defined and the annotation problems in those spaces are formulated as feature classification and Bayesian inference, respectively. A scheme of image content annotation in this framework is presented and evaluated as an application of photo categorization using MPEG-7 VCE2 dataset and its ground truth. The experimental results show a promising performance.

Disciplines

Physical Sciences and Mathematics

Publication Details

This paper was originally published as: Ye, L, Ogunbona, P & Wang, J, Image Content Annotation Based on Visual Features, 8th IEEE International Symposium on Multimedia 2006 (ISM'06), San Diego, USA, 11-13 December 2006, 62-69. Copyright IEEE 2006.

Image Content Annotation Based on Visual Features

Lei Ye, Philip Ogunbona and Jianqiang Wang
School of Information Technology and Computer Science
University of Wollongong, Australia
{lei, philipo, jw69}@uow.edu.au

Abstract

Automatic image content annotation techniques attempt to explore structural visual features of images that describe image content and associate them with image semantics. In this paper, two types of concept spaces, atomic concept and collective concept spaces, are defined and the annotation problems in those spaces are formulated as feature classification and Bayesian inference, respectively. A scheme of image content annotation in this framework is presented and evaluated as an application of photo categorisation using MPEG-7 VCE2 dataset and its ground truth. The experimental results show a promising performance.

1. Introduction

Annotated images are useful for image retrieval based on keywords and image content management. Manual annotation is not only tedious but also not practical in many cases. Most images are therefore available without adequate annotation. Automatic image content annotation becomes a recent research interest.

Automatic image content annotation techniques attempt to explore the visual characteristics of images and associate them with image semantics. In recent years, most reported work has been conducted in the context of image classification or categorisation that labels images with a few high-level concepts that are defined in an application domain.

In this paper, it is recognized that some concepts can be characterised by structural visual features of images, some concepts are not directly characterised by structural visual features of images but can be described by certain concepts that can be characterised by structural visual features of images, and some concepts can not even be derived from image visual features without further information or domain knowledges. We propose two categories of concepts, named atomic concepts and collective concepts, and formulate the annotation problems as feature classification and concept inference problems. An image content scheme based on

this formulation is presented and evaluated with MPEG-7 VCE2 dataset and its ground truth [21].

The rest of the paper is organised as follows. Section 2 reviews prior work in the area. Section 3 discuss the general annotation problem and concepts for annotation. Sections 4 and 5 provide the formulation of the annotation problems. Section 6 presents an annotation scheme and experimental results.

2. Prior Work

Saber *et al.* [23] proposed a colour classification method based on image pixels to annotate local areas of images.

To label photos of city scenes and landscape scenes, Gorkani *et al.* [12] used a multiscale steerable pyramid to find dominant orientations in 4×4 subblocks of the image. The image is classified as a city scene if enough subblocks have strong dominant vertical orientation, or alternatively medium-strong vertical orientation and also horizontal orientation. The performance was tested with a very limited number of 98 test images. Yiu [32] used the same dominant orientation features as [12] and together colour features to classify indoor and outdoor scenes. A nearest neighbour classifier was used for the colour feature and an SVM classifier for dominant orientation.

Vailaya *et al.* [28] attempted to capture high-level concepts from low-level image features by using binary Bayesian classifiers. Their work focused on hierarchical classification of vacation images. Images are first classified into indoor and outdoor images. Outdoor images are further classified as city or landscape. Finally, a subset of landscape images is classified into sunset, forest, and mountain images. A vector quantizer is used and class-conditional densities of the features are estimated for Bayesian classifiers.

Chang *et al.* [5] proposed a solution to the annotation problem by using Bayesian Point Machines (BPM). In their method, each training image is manually assigned a concept term from the lexicon, and the visual characteristics of the whole image are modelled using a colour and texture

feature vector (144-dimension). BPM is then used to train a classifier for each concept to determine the confidence score of assigning the concept to the image. For a new image, the system chooses those concept terms with high confidence scores.

Li *et al.* [15] pointed out that the fixed set of semantics, set of low-level features and set of training instances (3 common assumptions in training-based methods) result in a static classifier, and thereby are likely to suffer from the followings: (1) inability to recognise an instance of new semantics; (2) inability to provide a way to realize a potential misprediction automatically; (3) inability to comprehend the causes of a misprediction. Therefore, they proposed a Confidence-based Dynamic Ensemble (CDE). CDE uses multi-level indicators to assert the class prediction confidence at the SVM binary-class level, the ensemble multi-class level, and the bag (multiple multi-class classifiers) level.

Methods based on global features of images can classify images into certain classes that can be characterised by global features. Some methods have been proposed to classify images based on local features of regions or partitions of images.

A method proposed by Mori *et al.* [20] assumed that each image in the training set associates with several keywords. The image is divided into fixed-size blocks (from 3×3 to 7×7 pixels) and each block inherits the whole set of keywords associated with the image. Blocks are then clustered using vector quantization, and the accumulation of frequencies of keywords in each cluster are used to predict the keywords for new images.

Wang *et al.* [29] proposed a method that assigns a textual description of concepts for an image collection and employs a 2-D multi-resolution Hidden Markov Model (HMM) to capture the cross blocks and cross resolution dependencies between blocks for the entire image collection. Given a new image, the feature vector of the image is compared with the trained models, and statistically significant terms are extracted to annotate the image.

Tsai *et al.* [27] proposed a new approach to classify images based on the combination of image processing techniques and hybrid neural networks. In their approach, images are divided into a number of blocks using the quadtree decomposition algorithm. Five biggest blocks are chosen from 5 sub-regions (4 quadrant regions and 1 centre region) as the representative blocks to represent the main regions/objects of the image. An unsupervised neural network, Self-Organizing Maps (SOMs), is then applied to cluster the images based on colour and texture features extracted from the representative blocks. Finally, Support Vector Machines (SVMs) are trained by the representative images generated by SOMs to classify images into semantic categories.

Barnard *et al.* [2] used Blobworld [4] to produce the segmented regions within an image. The method uses statistics of word and region feature occurrence and co-occurrence. The region-word probabilities are then used to associate words with regions in new images.

Jeon *et al.* [14] assumed regions in an image can be described using a small vocabulary of blobs. Blobs are generated from image features using clustering based on Blobworld [4] segmentation method. Given a set of training image with annotations, the relationship between the set of keywords and the set of blobs in each image is derived by using the cross-media relevance models. Given an input image, the maximum-likelihood estimator is used to obtain a set of keywords for the image.

Fan *et al.* [10] proposed a multi-level approach to annotate images of natural scenes by using both the salient objects and the relevant semantic concepts. To detect the salient objects automatically, it firstly segments images using the mean shift technique [6]. Then the image regions are classified using an SVM classifier with an optimal model parameter search scheme. To exploit the contextual relationships between the concepts and the relevant salient objects, the Finite Mixture Model (FMM) is used to approximate the class distribution of the salient objects that are relevant to a certain concept.

To overcome segmentation problem, a novel statistical learning-based approach is proposed by Feng *et al.* [11]. In this method, the Blobworld [4] and UCSB [7] segmentation methods are used to segment the image into two independent sets of regions. The association between each set of regions and the semantic concepts is separately learned by a classifier combined with traditional binary SVM and soft decision binary SVM. Given a new image, a greedy strategy is used to annotate the image by using the trained classifier to associate one or more concepts to each region.

Rui *et al.* [22] proposed a method for clustering of regions into region clusters by incorporating pair-wise constraints and a greedy selection and joining algorithm to find the independent sub-sets of region clusters.

Yang *et al.* [30] used a Bayesian classifier for image regions with class conditional probabilities constructed by Complement Components Analysis (CCA).

Izquierdo *et al.* [13] proposed a method to reduce feature space for pattern classification using MPEG-7 descriptors of image regions.

In another direction of research, some methods have been proposed based on learning from user feedback [31, 26].

In addition, Lu *et al.* [17] proposed a method to map the feature vector to a model vector.

Our approach recognised that not all concepts used to annotate images can be visually perceived and therefore they can not be characterised by structural visual features. In this

paper, distinction is made between two types of concepts. Images are first annotated with atomic concepts that can be characterised by visual features and the collective concepts that can not be directly characterised by visual features but can be characterised by atomic concepts are then inferred from the atomic concepts.

3. Image Content Annotation Based on Visual Features

Image content annotation is a mapping A from an image I to a set C of concepts, known as the annotation space. In an image content annotation system based on visual features, visual features are first extracted from images, denoted as E_f , then features are mapped to concepts, denoted as C_f . Visual features are represented as a set F of feature vectors $\mathbf{f}_i$, ($i=1,2,...,n$), where $\mathbf{f}_i \in F_i \subset R^{k_i}$. Generally, the image content annotation based on visual features can be expressed as operations as follows.

$$I \xrightarrow{A} C = I \xrightarrow{E_f} F \xrightarrow{C_f} C. \quad (1)$$

3.1. Visual Feature Extraction

A great variety of image visual features have been explored for image retrieval and, recently, annotation, including features describing colours, textures and shapes etc. ISO/IEC MPEG-7 standardised a set of visual feature tools, called visual descriptors, for image and video content description [25, 19, 1, 18].

The following MPEG-7 descriptors are used in the proposed scheme in Section 6.

- Dominant colour Descriptor (DCD);
- Scalable colour Descriptor (SCD);
- colour Structure Descriptor (CSD);
- colour Layout Descriptor (CLD);
- Edge Histogram Descriptor (EHD).

In general the best features to characterise a concept are unknown and different from one concept to another. Concepts are not normally characterised by one visual feature. Eidenberger [9] analysed MPEG-7 visual descriptors and found that most of them are highly redundant. In our work, a selected set of visual features will be used for image grouping and image atomic concept annotation.

3.2. Concepts in Images

Normally, an image can be described by a few keywords. The description can be very subjective. For example, a photograph taken at a water side of Hawaii in an evening can

be described as a photography of *beach, holiday, seascape, sunset* even *Hawaii beach* etc. It can also be described as a photography of *sea, water, sand, sky, dark sky, reddish sky* etc. An image contains a great varieties of concepts. Some of them related to visual perceptions and some not. Some of concepts can be described by other concepts. For example, *beach* can be described as *water, sand* etc., and *sea* as a lot of *water*. The photography can be only possibly annotated as *Hawaii beach* with human feedback. In a practical automatic image content system, the concept set needs to be defined properly.

In an automatic image content annotation system, the concept set C are often predefined. We recognise that there are two types of concepts: atomic concepts and collective concepts. Atomic concepts are those concepts that describe a portion of an image and can be characterized by visual features; collective concepts are those concepts that describe the whole image and can be described by a few atomic concepts. For example, *beach* is a collective concept that describe the whole image content and can be described by *water, sand, sky* etc.

Let $C_a = \{c_i, i=1,2,...,M_a\}$ denote the atomic concept set, also called atomic concept space and $C_c = \{C_i, i=1,2,...,M_c\}$ the collective concept set, also called collective concept space. We have

$$C = C_c \cup C_a = \{C_1, C_2, ..., C_{M_c}\} \cup \{c_1, c_2, ..., c_{M_a}\}. \quad (2)$$

We assume that a well specified annotation system requires all collective concepts be described in the atomic concept space. If a collective concept C_i is described by m atomic concept, we denote it as

$$C_i = \{c_j; j = 1, 2, ..., m; c_j \in C_a\}. \quad (3)$$

In our proposed scheme, the feature mapping C_f from F to C can be decomposed as

$$I \xrightarrow{A} C = I \xrightarrow{A_a} C_a \vee I \xrightarrow{A_c} C_c \xrightarrow{A_c} C_c, \quad (4)$$

where $\vee$ means either one mapping.

Image categorisation classifies images into a few categories that can usually be described by collective concepts. With our notation, image categorisation is a special case of annotation that can be defined as the mapping from I to C_c , as

$$I \xrightarrow{A} C_c = I \xrightarrow{A_c \cdot A_a} C_c. \quad (5)$$

3.3. From Features to Concepts

In our proposed framework, features are first mapped to atomic concepts by C_{fa} and atomic concepts are then mapped to collective concepts by C_{fc} . In this paper, the mapping C_{fa} is treated as a feature classification problem and the mapping C_{fc} a statistical inference problem, which are discussed in detail in following sections.

4. Feature Classification

Visual features can be extracted from the whole image or from a partition of an image. In order to obtain atomic concepts that describe portions of an image, the features are extracted from partitions of an image. In our work, an image is partitioned into small blocks. Let $I_{b,i}$ denote the i^{th} block of the image and the same notation A_a is used for the atomic concept annotation of image blocks for simplicity.

$$A_a(I) = A_a(\{I_{b,i}, i = 1, 2, \dots, N\}) = \bigcup_{i=1}^N A_a(I_{b,i}), \quad (6)$$

where A_a annotates the image block by mapping it to an atomic concept, as

$$A_a(I_{b,i}) = c_j; c_j \in C_a. \quad (7)$$

The size of $I_{b,i}$ is sufficiently small so that it is from a region that can be described by an atomic concept. The advantage of using small image blocks is that there is not region or object segmentation required.

Atomic annotation is based on visual features of the image block.

$$E_f(I_{b,i}) = \mathbf{f}. \quad (8)$$

The problem of image annotation with atomic concepts becomes the classification of features of image blocks according to a given atomic concept set [8]. An atomic concept c_j is represented by a prototype feature vector $\mathbf{f}_j$. The problem here is to find a prediction for $\mathbf{f}_j$ given the values of $\mathbf{f}$, with squared error as the loss function. The problem is solved by nearest-neighbour methods using the training data as

$$\text{Average}(\mathbf{f} | \mathbf{f} \in N(\mathbf{f}_j)), \quad (9)$$

where N is the neighbourhood of $\mathbf{f}_j$.

The semantic meaning of each block is determined by the atomic concept associated with the image region where the block is located. The association is subjective in terms of use of words for the concept and created manually. In practice, a training set for each predefined atomic concept is created to train a classifier.

5. Collective Concept Inference

Once all image blocks are classified with the atomic concepts, the collective concept can be inferred from the resulting atomic concepts.

The problem of image content annotation by the collective concept is to find what is the the most possible collective concept given a set of atomic concepts. As discussed in the SubSection 3.2, a collective concept can be described by a set of atomic concepts and expressed as Eq.(3). Generally,

a collective concept C_i can be parameterised by the probability distribution of atomic concepts, which is represented by a vector,

$$\mathbf{p}_i = (p_1, p_2, \dots, p_j, \dots, p_{M_a}), \quad (10)$$

where p_j is the probability that the atomic concept c_j appears in the set of atomic concepts that describes the collective concept C_i . For an image I , if N_j is the total number of image blocks that are described by the atomic concept c_j , the probability p_j is estimated by the frequency N_j/N . Let $\mathbf{p}$ denote the atomic concept distribution of the image I , that is

$$\mathbf{p} = (N_1/N, N_2/N, \dots, N_j/N, \dots, N_{M_a}/N). \quad (11)$$

The collective concept annotation becomes the problem of finding the most probable collective concept C_i among the given collective concept space C_c for the atomic concept distribution vector $\mathbf{p}$ of an image I . That is, for a distribution $\Pr(\mathbf{p}_i | C_i)$ for the atomic concept distribution (parameters) of each collective concept C_i , the posterior probability of a given collective concept, the probability of the image I has the collective concept C_i , is

$$\Pr(C_i | I) = \Pr(C_i | \mathbf{p}) = \Pr(\mathbf{p} | C_i). \quad (12)$$

According to Bayesian rules,

$$\Pr(C_i | \mathbf{p}) \propto \Pr(\mathbf{p} | C_i) \cdot \Pr(C_i). \quad (13)$$

Using the maximum a posterior criterion,

$$C_i = \max_{C_i \in C_c}^{-1} \Pr(\mathbf{p} | C_i) \cdot \Pr(C_i). \quad (14)$$

It is assumed that the atomic concept distribution vector $\mathbf{p}$ for a collective concept follows a Gaussian distribution, the conditional density is given as

$$\Pr(\mathbf{p} | C_i) = \frac{1}{\sqrt{2^{M_a} \cdot \pi^{M_a} \cdot \Sigma}} e^{-\frac{(\mathbf{p} - \bar{\mathbf{p}})^T \Sigma^{-1} (\mathbf{p} - \bar{\mathbf{p}})}{2}} \quad (15)$$

where $\bar{\mathbf{p}}$ is the mean of atomic concept distribution for the collective concept C_i , Σ is the covariance matrix. It is further assumed that each collective concept is independent. The prior probability $\Pr(C_i)$ can be estimated from the training set. Therefore, the collective concept annotation can be solved as Bayesian-Gaussian inference [3].

6. An Image Content Annotation Scheme and Experimental Results

An image content annotation scheme is designed based on the models of atomic concept and collective concept spaces.

Figure 1. Block Diagram: An Image Content Annotation Scheme

It is found that the atomic concepts have very different visual features under different environmental illumination such as daylight, night-time or evening etc. A grouping is therefore employed using a colour feature of the whole image to group images into three groups. Then the images are partitioned into fixed-size blocks. Each image block is annotated with an atomic concept. The atomic concept histogram is calculated to estimate the probability distribution of atomic concepts of the image. Then Bayesian-Gaussian inference is used to determine the collective concept of the image. Fig.1 shows the block diagram of the scheme.

6.1. Experiment Dataset

Experiments are carried out to evaluate the performance of the scheme. We used the MPEG-7 VCE2 image dataset and its ground truth (GT) [21]. VCE2 dataset contains 3828 photographic images including consumer photos and photos from Corel Image Collection. This dataset was created for photo categorisation tests. The images in the VCE2 dataset are labelled with 7 categories including *architecture*, *interior*, *terrain*, *night-scene*, *snowscape*, *sunset* and *water-side*. These categories form the collective concept space in our experiments. The performance will be evaluated on the VCE2 dataset as a photo categorisation application against the VCE2 GT, which is a special case of collective concept annotation in our notation.

6.2. Grouping

The same atomic concepts under various environmental illumination have dramatically different perceptual characteristics and visual features. A image grouping is used as a pre-processing stage. The groups are dependent on the application domain. For example, the *water* in daylight and night-time is visually dramatically different. As the difference as such is a result of illumination conditions, three groups are determined according to the time, includ-

Descriptor	Group	Precision (%)	Recall (%)	Overall Precision	Overall Recall
DCD	Night-Scene	91.2	100.0	91.0	91.5
	Daylight	92.7	92.1		
	Sunset	89.1	82.3		
CLD	Night-Scene	83.9	95.7	88.6	88.3
	Daylight	90.7	96.2		
	Sunset	91.3	72.9		
CSD	Night-Scene	78.2	75.1	79.4	76.4
	Daylight	92.0	96.3		
	Sunset	67.9	57.9		
SCD	Night-Scene	90.5	95.5	89.7	89.7
	Daylight	91.2	90.4		
	Sunset	87.3	83.1		

Table 1. colour Descriptor Evaluation for Image Grouping

ing night-scene, daylight and sunset for the VCE2 dataset. This grouping is basically reflected by the colour characteristics of images so that a colour feature of the whole image is to be selected for this purpose.

Four MPEG-7 colour descriptors, including DCD, CLD, CSD and SCD, are evaluated. A simple 3-class clustering classifier is trained using a colour feature. Table 1 shows the comparative performance of the four colour descriptors. It is found that DCD outperforms other colour descriptors for image grouping into the defined 3 groups. This finding is the same as our intuition since the dominant colour of photography is usually determined by the environmental illumination.

6.3. Feature Extraction and Classification

Images are partitioned into 64 small blocks and each block is annotated with an atomic concept. The atomic concept space is domain dependent. For VCE2 dataset and the defined 7 collective concepts, an atomic concept space containing 20 atomic concepts is defined. The atomic concept space contains concepts such as *sky*, *water*, *sand*, *tree*, *rock*, *grass* and *building* etc. These concepts describe various local areas of an image. Two MPEG-7 visual descriptors, CSD and EHD, are used to characterise image blocks. As formulated in the section 4, clustering classifiers are used for feature classification. A training set is created for each atomic concept to train the classifier using the GLA algorithm [8, 16].

Multi-feature description of images are studied in the context of content-based image retrieval and a linear combination of multiple features with a weighted Euclidean distance are commonly used [24]. However, the linear combination is not a good model for multi-feature description of visual characteristics of images for semantic clustering [24]. In our scheme, the 2 visual descriptors are used in tandem.

6.4. Parameter Estimation

As formulated in Section 5, collective concepts are parameterised with the atomic concept distributions as Eq. (10). The histogram of atomic concepts of images are calculated using Eq. (11) to estimate the distributions of atomic concepts.

6.5. Collective Concept Annotation - Bayesian Inference

The model is trained with a training set based on the VCE2 GT. We used 100 images randomly selected from each category as a training set except *snowscape* and *sunset* from which 50 images are selected.

Table 2 shows the experimental results with EHD followed by CSD. Table 3 shows the experimental results with CSD followed by EHD. There is no significant difference between the orders of applications of the 2 descriptors because they are independent in describing the images. In both cases, the annotation performances are promising.

Direct comparisons with other published approaches are not possible because different datasets and concepts or categories are used in experiments published in literatures. Furthermore, various performance measures have been used in image annotation research. In this paper, the precision and recall is used to measure the annotation performance against the ground truth of the dataset, which is the performance measure adopted by the ISO/IEC MPEG committee for the photo categorisation core experiments.

Categories	Number of Test Images	Precision(%)	Recall(%)
Architecture	1325	69.1	57.2
Interior	647	78.2	27.9
Terrain	1494	77.6	62.3
Night-scene	216	91.7	98.6
Snowscape	137	66.2	54.0
Sunset	77	88.6	80.1
Waterside	523	70.1	68.1
Average	2167 (total)	74.7	58.2

Table 2. Performance of the Annotation Scheme with EHD followed by CSD

Categories	Number of Test Images	Precision(%)	Recall(%)
Architecture	1325	65.2	52.2
Interior	647	71.5	26.3
Terrain	1494	75.3	59.1
Night-scene	216	86.5	96.4
Snowscape	137	64.1	52.7
Sunset	77	85.9	79.2
Waterside	523	61.9	63.7
Average	2167 (total)	70.5	54.8

Table 3. Performance of the Annotation Scheme with CSD followed by EHD

7. Conclusions

In this paper, we proposed the concept of atomic and collective concept spaces and formulated the annotation problems in these spaces as feature classification and concept inference. This work provided a framework for image content annotation. A scheme of image content annotation in this framework was presented and evaluated as an application of photo categorisation using MPEG-7 VCE2 dataset and its ground truth. The experimental results showed a promising performance.

We used the GLA algorithm in feature classification for atomic concept annotation and Bayesian-Gaussian inference in collective concept annotation. In fact, other feature classification techniques and inference models can be used.

The concept space construction is domain dependent. What concepts are atomic and what collective are an open problem. In our work, we defined 20 atomic concepts to describe 7 collective concepts for MPEG-7 VCE2 dataset. Besides these two concept spaces, there are non-visual concepts that can not be characterised by visual features nor described by atomic concepts. Image content is hardly annotated with those non-visual concepts in an automatic annotation based on visual features. They may be associated with image content by making use of other information or

domain knowledge, such as location, event and time etc. or learned from user feedback.

Future work includes semantic and linguistic analysis of vocabularies according to their perceptual meaning and characteristics that can be described by visual feature descriptors. Concept space design algorithms based on some clustering analysis will be highly desirable.

References

- [1] M. Abdel-Mottaleb and S. Krishnamachari. Multimedia descriptions based on MPEG-7: Extraction and applications. *IEEE Trans. on Multimedia*, 6(3), June 2004.
- [2] K. Barnard and D. A. Forsyth. Learning the semantics of words and pictures. In *IEEE International Conference on Computer Vision*, pages 408–415, 2001.
- [3] G. E. P. Box and G. C. Tiao. *Bayesian Inference in Statistical Analysis*. Wiley-Interscience, reprint edition edition, 1992.
- [4] C. Carson, M. Thomas, S. Belongie, and J. M. Hellerstein. Blobworld: A system for region-based image indexing and retrieval. *Lecture Notes in Computer Science*, 1614:509, 1999.
- [5] E. Y. Chang, K. Goh, G. Sychay, and G. Wu. CBSA: content-based soft annotation for multimodal image retrieval using bayes point machines. *IEEE Trans. Circuits Syst. Video Techn*, 13(1):26–38, 2003.
- [6] D. Comanicu and P. Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Trans. PAMI*, 24:603–619, 2002.
- [7] Y. Deng and B. S. Manjunath. Unsupervised segmentation of color-texture regions in image and video. *IEEE Trans on Pattern Analysis and Machine Intelligence*, 23:800–810, 2001.
- [8] R. Duda, P. Hart, and D. Stork. *Pattern Classification*. John Wiley and Sons, Inc., 2nd ed. edition, 2000.
- [9] H. Eidenberger. Statistical analysis of content-based MPEG-7 descriptors for image retrieval. *Multimedia Syst*, 10(2), 2004.
- [10] J. Fan, Y. Gao, H. Luo, and G. Xu. Automatic image annotation by using concept-sensitive salient objects for image content representation. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 361–368, Jul 2004.
- [11] H. Feng and T.-S. Chua. A learning-based approach for annotating large on-line image collection. In *10th IEEE International Multimedia Modeling Conference*, Jan 2004.
- [12] M. Gorkani and R. W. Picard. Texture orientation for sorting photos at a glance. *Proc. Int. Conf. Pat. Rec.*, 1:459–464, Oct. 1994.
- [13] E. Izquierdo and A. Dorado. Climbing the semantic ladder: Towards semantic semi-automatic image annotation using MPEG-7 descriptor schemas. In *Proc. IEEE Int'l Workshop on Computer Architecture for Machine Perception*, pages 257–262, Italy, Jul 2005.
- [14] J. Jeon, V. Lavrenko, and R. Manmatha. Automatic image annotation and retrieval using cross-media relevance models. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*, Toronto, Canada, Jul 2003.
- [15] B. Li and K. Goh. Confidence-based dynamic ensemble for image annotation and semantics discovery. In *Proceedings of the 11th ACM International Conference on Multimedia (MM-03)*, pages 195–206, Nov 2003.
- [16] Y. Linde, A. Buzo, and R. Gray. An algorithm for vector quantizer design. *IEEE Trans. on Communications*, 28(1):84–95, Jan 1980.
- [17] J. Lu, S. P. Ma, and M. Zhang. Automatic image annotation based on model space. In *Proc. Natural Language Processing and Knowledge Engineering*, pages 455–460, Oct 2005.
- [18] B. S. Manjunath, J.-R. Ohm, V. V. Vasudevan, and A. Yamada. Color and texture descriptors. *IEEE Trans on Circuits and System for Video Technology*, 11(6), June 2001.
- [19] B. S. Manjunath, P. Salembier, and T. Sikora, editors. *Introduction to MPEG-7: Multimedia Content Description Interface*. John Wiley & Sons, Ltd., 2002.
- [20] Y. Mori, H. Takahashi, and R. Oka. Image-to-word transformation based on dividing and vector quantizing images with words. *First International Workshop on Multimedia Intelligent Storage and Retrieval Management*, Oct 1999.
- [21] MPEG-7 Visual Group. CE for category based photo clustering (VCE-2) (N6369). In *ISO/IEC JTC1/SC29/WG11*, Munich, Germany, 2004.
- [22] S. Rui, W. Jin, and T.-S. Chua. A novel approach to auto image annotation based on pairwise constrained clustering and semi-naïve Bayesian model. In *Proc. the 11th Intl Multimedia Modelling Conf*, pages 322–327, Jan 2005.
- [23] E. Saber, A. M. Tekalp, R. Eschbach, and K. Knox. Automatic image annotation using adaptive color classification. *Graphical Models and Image Processing*, 58(2):115–126, March 1996.
- [24] G. Sheikholsami, W. Chang, and A. Zhang, editors. *Semantic clustering and querying on heterogeneous features for visual data*, Bristol, UK, 1998. ACM Multimedia.
- [25] T. Sikora. The MPEG-7 visual standard for content description - an overview. *IEEE Trans on Circuits and System for Video Technology*, 11(6), June 2001.
- [26] S. Tong and E. Chang. Support vector machine active learning for image retrieval. *Proceedings of the ninth ACM international conference on Multimedia*, October 2001.
- [27] C.-F. Tsai, K. McGarry, and J. Tait. Image classification using hybrid neural networks. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 431–432. ACM, Jul 2003.
- [28] A. Vailaya, M. Figueiredo, A. Jain, and H.-J. Zhang. Image classification for content-based indexing. *IEEE Transactions on Image Processing*, 10(1):117–130, Jan. 2001.
- [29] J. Z. Wang and J. Li. Learning-based linguistic indexing of pictures with 2-D MHMMs. In *Proceedings of the tenth ACM international conference on Multimedia (MM-02)*, pages 436–445, New York, USA, Dec 2002. ACM Press.
- [30] C. Yang, M. Dong, and F. Fotouhi. Image content annotation using bayesian framework and complement components analysis. In *International Conference on Image Processing*, pages I: 1193–1196, 2005.

- [31] Z. Yang and C.-C. J. Kuo. Learning image similarities and categories from content analysis and relevance feedback. *Proceedings of the 2000 ACM workshops on Multimedia*, pages 175–178, November 2000.
- [32] E. C. Yiu. Image classification using color cues and texture orientation. Master's thesis, MIT, 1996.