Available online at www.sciencedirect.com

ScienceDirect

Expert Systems with Applications 35 (2008) 1468–1472

Expert Systems with Applications

www.elsevier.com/locate/eswa

An EM based multiple instance learning method for image classification

H.T. Pao [a,*], S.C. Chuang [b], Y.Y. Xu [b], Hsin-Chia Fu [b]

[a] *Department of Management Science, National Chiao Tung University, Hsinchu, Taiwan, ROC*
[b] *Department of Computer Science, National Chiao Tung University, Hsinchu, Taiwan, ROC*

Abstract

In this paper, we propose an EM based learning algorithm to provide a comprehensive procedure for maximizing the measurement of diverse density on given multiple Instances. Furthermore, the new EM based learning framework converts an MI problem into a single-instance treatment by using EM to maximize the instance responsibility for the corresponding label of each bag. To learn a desired image class, a user may select a set of exemplar images and label them to be conceptual related (positive) or conceptual unrelated (negative) images. A positive image consists of at least one object that the user may be interested, and a negative image should not contain any object that the user may be interested. By using the proposed EM based learning algorithm, an image retrieval prototype system is implemented. Experimental results show that for only a few times of relearning cycles, the prototype system can retrieve user's favor images from WWW over Internet.

Keywords: Multiple-instance learning; Image retrieve; WWW; EM method

1. Introduction

Due to rapid advances in computer hardware and Internet capability, visual information, incorporating image storage and retrieval, has become a preferred source of information for many consumers. Information in a visual form differs from traditional database contents in many important ways. It requires more space for storage, is highly unstructured, and needs some kind of decoding process to determine its semantic content (Leung, Hibler, & Mwara, 1992; Sethi, 1995). The limitation of the current image analysis techniques necessitates that most image retrieval systems use some form of text description provided by users as the basis to index and retrieve images. These techniques are rather primitive at present and they need further development and refinement (Hwarth & Buxton, 2000; Lin, Wang, & Yang, 1999). Many different methods and techniques have been recently proposed for modeling and retrieving images (Chua, Pung, Lu, & Jong, 1994; Gamanya, Maeyer, & Dapper, 2007; Gudivada & Raghavan, 1997; Jorgensen, 1998; Shakir & Nagao, 1996; Yang & Wu, 1997). They largely fall into two categories, namely, content-based methods and concept-based methods. Content-based image retrieval systems have been proposed to index and search by their low-level features (contents) such as color, texture, shape, spectral, etc. Photobook Pentland, Picard, and Sclaroff (1994), QBIC Ashley et al. (1995), NETRA Ma and Manjunath (1997) and Tsai (2007) are such attempts to enable users to query based on these features. Smith and Chang (1999) present a general framework for querying and retrieving images by spatial and feature attributes. The spatial and feature (SaFe) system integrates content-based techniques with spatial query methods in order to search for image by arrangement of regions. Usually, the contents of an image are very complicated, so an image can be seen as the combination of the small subimages, in which each subimage has its own content. For example, if an image contains an interested subimage such as "*Waterfall*" and a few unin-

[*] Corresponding author. Tel.: +886 35 731578; fax: +886 35 724176.
E-mail address: htpao@cc.nctu.edu.tw (H.T. Pao).

0957-4174/$ - see front matter © 2007 Published by Elsevier Ltd.
doi:10.1016/j.eswa.2007.08.055

terested subimages. One would like to identify this image as "*The Image Containing a Waterfall*". In traditional methods, feature vectors are extracted from the whole image. It is very hard to extract suitable feature vectors from the whole image just to properly represent "*The Image Containing a Waterfall*". Thus, some methods proposed to first segment interested subimages from an image, and then extract feature vectors from the interested subimages. In fact, it is very difficult to segment the interested subimages precisely. Beside, the interested subimages in an image may be different for different users, as a consequence, different feature vectors may be extracted from different interested subimages when the same image are queried by different users. Thus, this approach complicates the system design and confuses users in selecting proper query subimages. In order to represent an image correctly, multiple-instances learning (MIL) is a way to model ambiguity in semi-supervised learning setting, where each training example is a bag of instances and the labels are assigned on the bags instead of on the instances. After being introduced by Dietterich, Lathrop, and Lozano-Perez (1997), MIL has become an active research area and a number of MIL algorithms have been proposed (Andrews, Tsochantaridis, & Hofmann, 2004, Wang & Zucker, 2000), etc. Maron and Ratan (1998) proposed the Multiple-Instance learning method to learn several instances with various diverse densities, and to maximize diverse density (DD) by Quasi-Newton method. In addition, Zhang and Goldman (2001) proposed an EM for multiple-instance learning by using Quasi-Newton search to maximize DD. In this paper, we propose an EM based learning algorithm, which provides a comprehensive procedure to maximizing the measurement of diverse density of the given multiple Instances. Furthermore, the new EM based learning framework converts MI problem into a single-instance treatment by using EM to maximize the instance responsibility for the corresponding label of each bag of instances. As we can see, the proposed algorithm provides comprehensive procedures to maximizing the measurement of diverse density of the given multiple instances.

2. EM based multiple-instance learning algorithm

In the Multiple-Instance learning, conceptual related (positive) and conceptual unrelated (negative) images are used for reinforced and antireinforced learning of a user's desired image class. Each positive training image contains at least one interested subimage related to the desired image class, and each negative training image should not contain any subimage related to the desired image class. The target of the Multiple-instance learning is to search the optimal point of the image class in the feature space, where the optimal point is close to the intersection of the feature vectors extracted from the subimages of the positive training images and is far from the union of the feature vectors extracted from the subimages of the negative training images.

For example, if one wants to train an image class t with P_t positive images and N_t negative images. Each positive image has V_t^+ subimages, and each negative image has V_t^- subimages. We denote the kth feature vector extracted from the kth subimage of the ith positive image as $\mathbf{X}_{\mathbf{ik}}^+$, and the kth feature vector extracted from the kth subimage of the ith negative example as $\mathbf{X}_{\mathbf{ik}}^-$. The probability that $\mathbf{X}_{\mathbf{ik}}^+$ belongs to class t is $P(t \mid \mathbf{X}_{\mathbf{ik}}^+)$, and the probability that $\mathbf{X}_{\mathbf{ik}}^-$ belongs to class t is $P(t \mid \mathbf{X}_{\mathbf{ik}}^-)$. A measurement called Diverse Density is used to evaluate that how many different positive images have feature vectors near a point t, and how far the negative feature vectors are from a point t. The Diverse Density for a class t is defined as Pearl (1998)

$$DD_t = \prod_{i=1}^{P_t}(1 - \prod_{k=1}^{V_t^+}(1 - P(t \mid \mathbf{X}_{\mathbf{ik}}^+)))\prod_{i=1}^{N_t}(\prod_{k=1}^{V_t^-}(1 - P(t \mid \mathbf{X}_{\mathbf{ik}}^-))). \tag{1}$$

The optimal point of the class t is appeared where the diverse density is maximized. By taking the first partial derivatives of Eq. (1) with respect to parameters of the class t and setting the partial derivatives to zero, the optimal point of the class t can be obtained. Suppose the density function of the class t is a D-dimensional Gaussian mixture with uncorrelated features. The parameters are the mean μ_{tcd}, the variance σ_{tcd}^2, and the cluster prior probability p_{tc} of each cluster in the class t. The estimating parameters of the class t can be derived by $\frac{\partial}{\partial \mu_{tcd}} DD_t = 0$, $\frac{\partial}{\partial \sigma_{tcd}} DD_t = 0$, and $\frac{\partial}{\partial p_{tc}} DD_t = 0$. Thus

$$\begin{aligned}\mu_{tcd} = &\left[\sum_{i=1}^{P_t}\left(\left(\frac{\mathbb{P}_{ti}}{1-\mathbb{P}_{ti}}\right)\sum_{k=1}^{V_t^+} Q_{tc}(\mathbf{X}_{\mathbf{ik}}^+)x_{ikd}^+\right)\right.\\ &\left. - \sum_{i=1}^{N_t}\sum_{k=1}^{V_t^-} Q_{tc}(\mathbf{X}_{\mathbf{ik}}^-)x_{ikd}^-\right]\\ &\bigg/\left[\sum_{i=1}^{P_t}\left(\left(\frac{\mathbb{P}_{ti}}{1-\mathbb{P}_{ti}}\right)\sum_{k=1}^{V_t^+} Q_{tc}(\mathbf{X}_{\mathbf{ik}}^+)\right)\right.\\ &\left. - \sum_{i=1}^{N_t}\sum_{k=1}^{V_t^-} Q_{tc}(\mathbf{X}_{\mathbf{ik}}^-)\right],\end{aligned} \tag{2}$$

$$\begin{aligned}\sigma_{tcd}^2 = &\left[\sum_{i=1}^{P_t}\left(\left(\frac{\mathbb{P}_{ti}}{1-\mathbb{P}_{ti}}\right)\sum_{k=1}^{V_t^+} Q_{tc}(\mathbf{X}_{\mathbf{ik}}^+)\|x_{ikd}^+ - \mu_{tcd}\|^2\right)\right.\\ &\left. - \sum_{i=1}^{N_t}\sum_{k=1}^{V_t^-} Q_{tc}(\mathbf{X}_{\mathbf{ik}}^-)\|x_{ikd}^- - \mu_{tcd}\|^2\right]\\ &\bigg/\left[\sum_{i=1}^{P_t}\left(\left(\frac{\mathbb{P}_{ti}}{1-\mathbb{P}_{ti}}\right)\sum_{k=1}^{V_t^+} Q_{tc}(\mathbf{X}_{\mathbf{ik}}^+)\right)\right.\\ &\left. - \sum_{i=1}^{N_t}\sum_{k=1}^{V_t^-} Q_{tc}(\mathbf{X}_{\mathbf{ik}}^-)\right],\end{aligned} \tag{3}$$

$$p_{tc} = \left[\sum_{i=1}^{P_t}\left(\left(\frac{\mathbb{P}_{ti}}{1-\mathbb{P}_{ti}}\right)\sum_{k=1}^{V_t^+} Q_{tc}(\mathbf{X}_{\mathbf{ik}}^+)\right) - \sum_{i=1}^{N_t}\sum_{k=1}^{V_t^-} Q_{tc}(\mathbf{X}_{\mathbf{ik}}^-)\right] \Bigg/ \left[\sum_{i=1}^{P_t}\left(\left(\frac{\mathbb{P}_{ti}}{1-\mathbb{P}_{ti}}\right)\sum_{k=1}^{V_t^+}\left(\frac{P(t|\mathbf{X}_{\mathbf{ik}}^+)}{1-P(t|\mathbf{X}_{\mathbf{ik}}^+)}\right)\right) - \sum_{i=1}^{N_t}\sum_{k=1}^{V_t^-}\left(\frac{P(t|\mathbf{X}_{\mathbf{ik}}^-)}{1-P(t|\mathbf{X}_{\mathbf{ik}}^-)}\right)\right], \quad (4)$$

where

$$P(c|\mathbf{X}_{\mathbf{ik}}^{\star}, t) = \frac{p_{tc}\cdot P(\mathbf{X}_{\mathbf{ik}}^{\star}|c,t)}{P(t|\mathbf{X}_{\mathbf{ik}}^{\star})}, \quad (5)$$

$$P(t|\mathbf{X}_{\mathbf{ik}}^{\star})^{(l)} = \frac{P(\mathbf{X}_{\mathbf{ik}}^{\star(l)}|t)P_t}{P(\mathbf{X}_{\mathbf{ik}}^{\star})}, \quad (6)$$

$$\mathbb{P}_{ti} = \prod_{k=1}^{P_i}\left(1 - P\left(t|\mathbf{X}_{\mathbf{ik}}^{\star}\right)\right),$$

$$Q_{tc}(\mathbf{X}_{\mathbf{ik}}^{\star}) = \frac{P(t|\mathbf{X}_{\mathbf{ik}}^{\star})P(c|\mathbf{X}_{\mathbf{ik}}^{\star}, t)}{1 - P(t|\mathbf{X}_{\mathbf{ik}}^{\star})},$$

$$P(\mathbf{X}_{\mathbf{ik}}^{\star}|c,t) = \frac{1}{\prod_{d=1}^{D}(2\pi\sigma_{tcd}^2)^{\frac{1}{2}}} \cdot \exp\left(-\frac{1}{2}\sum_{d=1}^{D}\left(\frac{x_{ikd}^{\star}-\mu_{tcd}}{\sigma_{tcd}}\right)^2\right),$$

and the notation $\star$ represents + or −.

According to Eqs. (2)–(4), we proposed an EM based Multiple-Instance learning algorithm to learn these parameters. The EM based Multiple-Instance learning algorithm contains two steps: the expectation step (E-step) and the maximization step (M-step). The algorithm is described as follows.

1. Choose an initial point in the feature space, and let its parameters are $\mu_{tcd}^{(0)}$, $\sigma_{tcd}^{2(0)}$, and $p_{tc}^{(0)}$.
2. *E-Step*: Using the calculated model parameters $\mu_{tcd}^{(l)}$, $\sigma_{tcd}^{2(l)}$, $p_{tc}^{(l)}$, Eqs. (5) and (6), estimate $P^{(l)}(c|\mathbf{X}_{\mathbf{ik}}^{\star}, t)$ and $P^{(l)}(t|\mathbf{X}_{\mathbf{ik}}^{\star})$.*M-Step*: Using the estimated $P^{(l)}(c|\mathbf{X}_{\mathbf{ik}}^{\star}, t)$ and $P^{(l)}(t|\mathbf{X}_{\mathbf{ik}}^{\star})$, compute the *new* model parameters $\mu_{tcd}^{(l+1)}$, $\sigma_{tcd}^{2(l+1)}$, and $p_{tc}^{(l+1)}$ according to Eqs. (2)–(4).
3. Calculate the diverse density $DD^{(l+1)}$. If $(DD^{(l+1)} - DD^{l})$ is smaller than a predefined threshold ϵ, then stop the process. Otherwise, loop step 2.

As we can see, the proposed algorithm provides comprehensive procedures to maximizing the measurement of diverse density of the given multiple instances. Furthermore, the new EM based learning framework converts MI problem into a single-instance treatment by using EM to estimate and to maximize the instance responsibility for the corresponding label of each bag of instances.

3. Image feature extraction and indexing

Before training the system for indexing images, multiple feature vectors are extracted from the multiple instances of several exemplar training images. Then, the system are trained according to the proposed EM based Multiple-Instance learning algorithm. Finally, the testing images are evaluated using Bayesian decision rule for indexing and classification.

3.1. Image feature extraction

The image features extraction we used are similar to the method proposed in (Pearl, 1998). First, a number of instances are randomly selected from an image. Then, the feature vectors are extracted from the instances as shown in Fig. 1. The feature vector in the position (*i*,*j*) is defined as $\mathbf{X} = \{x_1,\cdots,x_{15}\}$, where

- $\{x_1, x_2, x_3\}$ is the average YUV values of C.
- $\{x_4, x_5, x_6\}$ is the average YUV values of A minus average YUV values of *C*.
- $\{x_7, x_8, x_9\}$ is the average YUV values of B minus average YUV values of *C*.
- $\{x_{10}, x_{11}, x_{12}\}$ is the average YUV values of D minus average YUV values of *C*.
- $\{x_{13}, x_{14}, x_{15}\}$ is the average YUV values of E minus average YUV values of *C*.

It is clear to see that the proposed feature extraction provides not only the color information but also some of the spatial information.

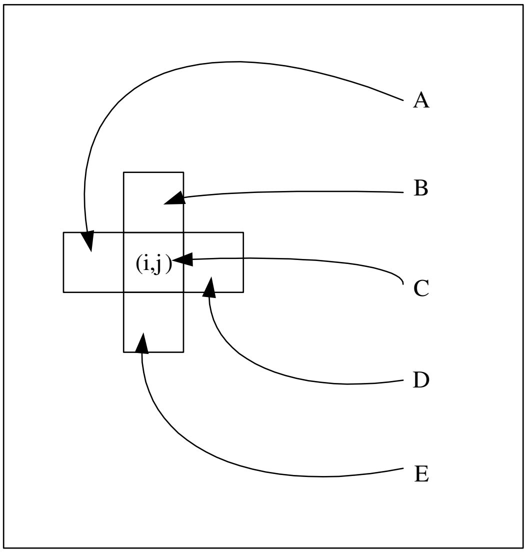


Fig. 1. Feature vectors are extracted from "+" shaped subimage (instance). The instance consists of 5 subregions: *A*, *B*, *C*, *D* and *E*. Each subregion is composed of 2 × 2 pixels. The feature vector, $\mathbf{X} = \{x_1, \ldots, x_{15}\}$, is the YUV value of *C*, and the difference values of *C* and its 4 neighbors.

3.2. Image indexing

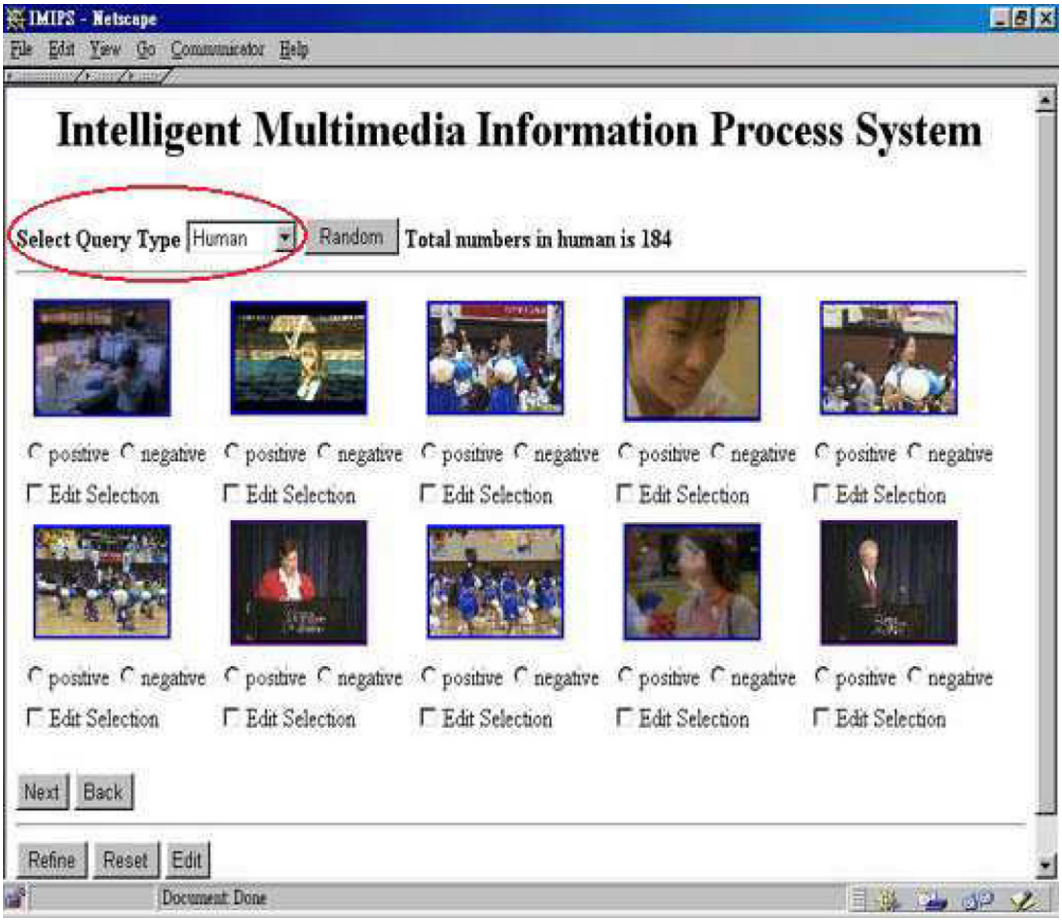
After the feature vectors are extracted from the sub-images of the training images, the system can be trained to perform indexing and classification for the images with respect to the class t. First, a user needs to select some related and unrelated images for a class as training images, then feature vectors of these images are extracted. Once the training feature vectors are ready, the EM based Multiple-Instance learning algorithm is used to compute the mean μ_{tcd}, the variance σ^2_{tcd}, and the cluster prior probability p_{tc}. By using these parameters, the posterior probabilities $P(t|\mathbf{X}_i)$ of an unindex image can be computed for each class. The unindex image is indexed to the class f if the $P(f|\mathbf{X}_i)$ is the highest among all the $P(t|\mathbf{X}_i)$.

4. Experimental results

We have built a prototype system to evaluate the proposed image classification and indexing method. This system is called the "*Intelligent Multimedia Information Processing Systems*" (IMIPS) (IMIPS, xxxx). This system has been used as a video search engine over the WWW. Once a new video file is found by a video spider in IMIPS, the system will download the file and save several key frames in a database.

When a desired image class is to be trained, the positive and the negative exemplar images are selected from the stored key frames. Then, the system learns the desired image class using the proposed EM based Multiple-Instance learning algorithm. When the optimal model for desired class is trained, each key frame in the database is indexed by its posterior probability associated with the desired class.

A web-based user interface is depicted in Fig. 2. Each of the title of the trained classes are display in a pulldown menu: "Select Query Type". When a user is trying to search an image of a certain class, one can use "Select Query Type" to select a class. Then, the associated images will be shown. Suppose, the class "Human" is selected, the system responds with all the images belonging to "Human" from the database in a descending order according to their computed value of posterior probability. As we can see, most of the shown images are human-related. If the user wants to create a new class, one can click the "Random" button, then a set of randomly selected images will be shown. One can select the "Positive" button to include conceptual related images, and select the "Negative" button to exclude conceptual unrelated images. When the "Refine"


Intelligent Multimedia Information Process System

Select Query Type: Human

Total numbers in human is 184

Document: Done

Fig. 2. When a user enters the IMIPS, one can select a class on interested of images, which are key frames of a video over the WWW. Then, the user can click on an image to view its associated video program.

Table 1
Results of natural scenes classification

	Total of retrieve				
Templates	10	20	30	40	50
Animal	10	20	30	39	48
Human	10	15	18	22	23
Sky	7	13	19	26	29
Star	10	20	30	35	40
Fire	9	13	N/A	N/A	N/A

The number in each row indicates the correctly retrieved images of each classes with respect to total retrieved images.

button is pressed, the system will train a new class according to the selections.

In this prototype system, we have trained five classes of nature scenes: “Human”, “Star”, “Sky”, “Animal”, and “Fire”. The experimental results are shown in Table 1. The number in the intersection of row “Animal” and column “10” indicates that the correctly retrieved images is 10 out of the total 10 retrieved images. The correctness of retrieve is judged according to human perception.

5. Conclusion

In this paper, we propose an EM based Multiple-Instance learning algorithm and implement a user friendly video search engine (IMIPS) over WWW. The multiple-instances learning (MIL) is a way to model ambiguity in semi-supervised learning setting, where each training example is a bag of instances and the labels are assigned on the bags instead of on the instances. In this paper, we propose an EM based learning algorithm, which provides a comprehensive procedure to maximizing the measurement of diverse density of the given multiple Instances. Furthermore, the new EM based learning framework converts MI problem into a single-instance treatment by using EM to maximize the instance responsibility for the corresponding label of the bag. The experimental results show that the retrieved images is quite match with human perception. How to properly determine the correct number of the clusters in the mixture Gaussian model of each class is a problem we want to solve in the future. In order to build a more powerful model, some features, such as shapes, textures,locations and domain knowledge, etc., will also be included in the future systems.

References

Andrews, S., Tsochantaridis, I., & Hofmann, T. (2004). Support vector machines for multiple-instance learning. In *NIPS*.

Ashley, J., Flickner, M., Hafner, J. L., Lee, D., Niblack, W., & Petkovic, D. (1995). The query by image content (QBIC) system. In *SIGMOD Conference*, (p. 475).

Chua, T. S., Pung, H. K., Lu, G. J., & Jong, H. S. (1994). A concept-based image retrieval system. *IEEE Computer*, 590–598.

Dietterich, T. G., Lathrop, R. H., & Lozano-Perez, T. (1997). Solving the multiple-instance problem with axis-parallel rectangles. *Artificial Intelligence Journal, 89*.

Gamanya, Ruvimbo, Maeyer, Philippe De, & Dapper, Morgan De (2007). An automated satellite image classification design using object-oriented segmentation algorithms: A move towards standardization. *Expert System with Applications, 32*(2), 616–624.

Gudivada, Venkat N., & Raghavan, Vijay V. (1997). Modeling and retrieving images content system. *Information Processing and Management, 33*(4), 427–452.

Hwarth, R. J., & Buxton, H. (2000). Conceptual-description from monitoring and watching image sequences. *Image and Vision Computing, 18*, 105–135.

IMIPS: The intelligent multimedia information processing system, http://imips.csie.nctu.edu.tw/imips/imips.html.

Jorgensen, Corinne (1998). Corinne jorgensen: Attributes of images in describing tasks. *ACM Transactions on Information Systems, 34*(2), 61–174.

Leung, C. H. C., Hibler, D., & Mwara, N. (1992). Image retrieval by content description. *Journal of Information Science, 18*, 111–119.

Lin, H. C., Wang, L. L., & Yang, S. N. (1999). Regular-texture image retrieval based on texture-primitive extraction. *Image and Vision Computing, 17*, 51–63.

Ma, W.Y., & Manjunath, B.S. (1997). NETRA: A toolbox for navigating large image databases. In *Proceedings of the IEEE International Conference Image Processing 1*, (Vol. 1, pp. 568, V571).

Maron, O., & Lakshmi Ratan, A. (1998). Multiple-instance learning for natural scene classification. In *Machine learning: Proceedings of the 15th international conference*.

Pearl, Judea (1998). *Probablistic reasoning in intelligent systems: Networks of plausible inference*. Morgan Kaufmann.

Pentland, A., Picard, R. W., & Sclaroff, S. (1994). Photobook: Content-based manipulation of image databases. In *Proceedings of the SPIE Storage Retrieval Image Video Databases II*, (pp. 34–47).

Sethi, Ishwar K. (1995). Image computing for visual information systems. *IEEE Computer*, 06–10.

Shakir, Hussain Sabir, & Nagao, Makoto (1996). Context-sensitive processing of semantic queries in an image database system. *Information Processing and Management, 32*(5), 573–600.

Smith, J. R., & Chang, S. F. (1999). Integrated spatial and feature image query, http://disney.ctr.columbia.edu/safe/. Multimedia System, 7(2):129–140.

Tsai, Chih-Fong (2007). Image mining by spectral features: A case study of scenery image classification 1. *Expert system with applications, 32*, 135–142.

Wang, J. & Zucker, J. D. (2000). Solving the multiple-instance problem: A lazy learning approach. In *Proceedings of the 17th international conference on machine learning*, (pp. 1119–1125).

Yang, Li, & Wu, Jiankang (1997). Towards a semantic image database system. *Data and Knowledge Engineering, 22*, 207–227.

Zhang, Q. & Goldman, S. A. (2001) Em-dd: An improved multiple-instance learning technique. In *Neural Information Processing Systems*.