University of Wollongong

Research Online

Faculty of Informatics - Papers (Archive)

Faculty of Engineering and Information Sciences

2007

Automatic Image Annotation for Semantic Image Retrieval

Wenbin Shao
University of Wollongong, wenbin@uow.edu.au

G. Naghdy
University of Wollongong, golshah@uow.edu.au

Son Lam Phung
University of Wollongong, phung@uow.edu.au

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

Part of the Physical Sciences and Mathematics Commons

Recommended Citation

Shao, Wenbin; Naghdy, G.; and Phung, Son Lam: Automatic Image Annotation for Semantic Image Retrieval 2007.
https://ro.uow.edu.au/infopapers/739

Research Online is the open access institutional repository for the University of Wollongong. For further information contact the UOW Library: research-pubs@uow.edu.au

Automatic Image Annotation for Semantic Image Retrieval

Abstract

This paper addresses the challenge of automatic annotation of images for semantic image retrieval. In this research, we aim to identify visual features that are suitable for semantic annotation tasks. We propose an image classification system that combines MPEG-7 visual descriptors and support vector machines. The system is applied to annotate cityscape and landscape images. For this task, our analysis shows that the colour structure and edge histogram descriptors perform best, compared to a wide range of MPEG-7 visual descriptors. On a dataset of 7200 landscape and cityscape images representing real-life varied quality and resolution, the MPEG-7 colour structure descriptor and edge histogram descriptor achieve a classification rate of 82.8% and 84.6%, respectively. By combining these two features, we are able to achieve a classification rate of 89.7%. Our results demonstrate that combining salient features can significantly improve classification of images.

Disciplines

Physical Sciences and Mathematics

Publication Details

This article was originally published as Shao, W, Naghdy, G and Phung, SL, Automatic Image Annotation for Semantic Image Retrieval, Lecture Notes in Computer Science, 4781, 2007, 369-378. Copyright Springer-Verlag.

This journal article is available at Research Online: https://ro.uow.edu.au/infopapers/739

Automatic image annotation for semantic image retrieval

Wenbin Shao, Golshah Naghdy, and Son Lam Phung

SECTE, University of Wollongong,
Wollongong NSW, 2522 Australia
{ws909,golshah,phung}@uow.edu.au

Abstract. This paper addresses the challenge of automatic annotation of images for semantic image retrieval. In this research, we aim to identify visual features that are suitable for semantic annotation tasks. We propose an image classification system that combines MPEG-7 visual descriptors and support vector machines. The system is applied to annotate cityscape and landscape images. For this task, our analysis shows that the colour structure and edge histogram descriptors perform best, compared to a wide range of MPEG-7 visual descriptors. On a dataset of 7200 landscape and cityscape images representing real-life varied quality and resolution, the MPEG-7 colour structure descriptor and edge histogram descriptor achieve a classification rate of 82.8% and 84.6%, respectively. By combining these two features, we are able to achieve a classification rate of 89.7%. Our results demonstrate that combining salient features can significantly improve classification of images.

Key words: image annotation, MPEG-7 visual descriptors, support vector machines, pattern classification

1 Introduction

Traditional image retrieval techniques are mainly based on manual text annotation [1]. Given the rapid increase in the number of digital images, manual image annotation is extremely time-consuming. Furthermore, it is annotator dependent. Content-based image retrieval (CBIR) promises to address some of the shortcomings of manual annotation [1–3]. Many existing CBIR systems rely on queries that are based on low-level features. One of the main challenges in CBIR is to bridge the *semantic gap* between low-level features and high-level contents [1, 2, 4]. For example, consider an image of a mountain: in low-level terms, it is a composition of colours, lines of different length, and different shapes; in high-level terms, it is a mountain. If users want to search for mountain images, they need to specify the low-level features such as green texture or they could enter the keyword *mountain.* Automatic annotation at semantic level employs keywords to represent images. It is a powerful approach, because people are better at describing an image with keywords than with low-level features.

In this paper, we aim to identify visual features that are suitable for automatic semantic annotation tasks. We propose an image classification system that combines MPEG-7 visual descriptors and support vector machines (SVMs), and apply the system to annotate cityscape and landscape images. Note that the system can be also extended to process other image categories. This paper is organised as follows. In Section 2, we review existing techniques for classifying images, especially cityscape versus landscape images. In Section 3, we describe the proposed system that combines MPEG-7 visual descriptors and support vector machines. In Section 4, we present and analyse the experiment results. Finally in Section 5, we give the concluding remarks.

2 Background

Many attempts at bridging the *semantic gap* have been made [1,2,4]. Yiu [5] uses colour histogram and dominant texture orientation to classify indoor and outdoor scenes. When the k-nearest neighbour classifier is used, Yiu finds that colour features outperform texture features. However, when support vector machine classifier is used, the texture features perform better than the colour features. The SVM classifier combining colour and texture features has a classification rate of 92% on a test dataset of 100 images.

Szummer and Rosalind [6] study indoor and outdoor image classification with four features: colour histogram in the Ohta colour space, texture feature based on a multi-resolution, simultaneous autoregressive model, and frequency features based on the two-dimensional DFT and DCT. They report a classification rate of 90.3% on a dataset of 1343 Kodak consumer images.

Vailaya et al. [7] propose a system on classification of cityscape versus landscape images that is based on the k-nearest neighbour classifier. They evaluate five features, namely colour histograms, colour coherence vectors, moments of image DCT coefficients, edge direction histograms, and edge direction coherence vectors. For features based on edge direction histograms and edge direction coherence vectors, Vailaya et al. report a classification rate of more than 93%, on a dataset of 2716 images (mainly Corel stock photos).

Vailaya et al. [8] later use a hierarchical architecture to first separate indoor from outdoor images and then divide outdoor images into subcategories such as *sunset*, *forest* and *mountain*. Their approach is based on Bayesian classifiers. For classifying indoor versus outdoor images, Vailaya et al. report an overall classification rate of 90.5% and find that features based on spatial colour distribution are better than colour and texture features. For classifying sunset versus forest and mountain images, they observe that colour histogram is better than the edge direction features. Vailaya et al. investigate the incremental learning technique and show that, as the training set increases, this technique improves classification accuracy. They also explore the affects of feature subset selection.

Lienhart and Hartmann [9] propose an image classification approach based on the AdaBoost learning algorithm. To classify graphical versus photo-realistic images, they use low-level features based on colour and pixel proportion. To classify real photos versus computer-generated images, they use texture features. In the classification of comics versus presentation slides, they use heuristic features based on text width and text position with respect to the entire image. They use about 5300 images for training and 2250 images for testing, and report classification rates between 93.7% and 98.2% on the test set.

Hu et al. [10] propose a Bayesian method with relevance feedback for indoor and outdoor image classification. The features they use include colour histograms, colour coherence vectors, edge direction histograms, edge direction coherence vectors and texture orientations. Hu et al. show that, on the same dataset, their method achieves a higher classification rate, compared to the method proposed by Szummer and Rosalind [6].

3 Methodology

We propose an image annotation system that combines MPEG-7 visual descriptors and support vector machines. The block diagram of the system is shown in Fig. 1. First, the input image is segmented into regions; this stage is only required for visual descriptors that are based on object shapes. Next, MPEG-7 visual descriptors are extracted from the image. Finally, support vector machines are used to classify the visual descriptors into different image categories such as *landscape*, *cityscape*, *vehicle* or *portrait*.

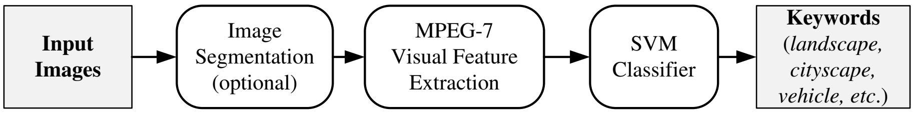

Input Images
Image Segmentation (optional)
MPEG-7 Visual Feature Extraction
SVM Classifier
Keywords (landscape, cityscape, vehicle, etc.)

Fig. 1. Proposed image annotation system.

3.1 Image segmentation

While most low-level features are extracted from the entire image, some features such as region-based shape and contour-based shape require the image to be segmented into regions or objects. Therefore, our system includes an optional image segmentation stage that relies on multi-resolution watershed segmentation and image morphology [11, 12].

In watershed segmentation, a grey image is considered as a topographic surface. Suppose the surface is flooded by water from its minima. When water from different sources is about to merge, dams are built to prevent the water from merging. The rising water finally partitions the image into two different sets: the catchment basins and the watershed ridge lines. We realise that the watershed method may cause over segmentation if it is applied directly to the gradient image. This problem can be alleviated if the topographic surface is flooded from a set of predefined markers. We propose extra processing steps that are based on two morphological operations: *opening by reconstruction* and *closing by reconstruction.* Image opening operation removes small regions (caused by noise) while preserving the shape of foreground objects. Image closing fills in holes in the objects while keeping their shapes.

3.2 MPEG-7 visual descriptors

Multimedia Content Description Interface (MPEG-7) is a standard for multimedia content description for a wide range of applications involving image, video and audio search. MPEG-7 defines three major categories of visual descriptors for still images: colour, texture and shape descriptors [13–16].

- **Colour descriptors**. All colour descriptors can be extracted from an image or an image region. There are four main descriptors: dominant colour, scalable colour, colour structure and colour layout. *Dominant colour* is a compacted description that consists of the representative colours in an image or an image region. *Scalable colour* is a colour histogram in the HSV colour space. It is encoded by a Haar transform. *Colour structure* is a colour structure histogram that consists of the information of colour content and the corresponding structure. *Colour layout* represents images by spatial colour structure. It is resolution-invariant and extracted in YCbCr colour space.

- **Texture descriptors**. All texture descriptors can be extracted from an image or an image region in MPEG-7 monochrome colour space. Three common texture descriptors are edge histogram, homogeneous texture and texture browsing. *Edge histogram* describes the local spatial distribution of edges in an image. It is extracted from 16 subimages of an image. There are four directional edges and one non-directional edge defined in MPEG-7. *Homogeneous texture* employs the mean energy and the energy deviation to characterise the region texture. *Texture browsing* specifies a texture in terms of regularity, coarseness and directionality.

- **Shape descriptors**. For each two dimensional region, there are two types of shape descriptors: a *region-based shape* descriptor represents the shape of a region whereas a *contour-based shape* descriptor reveals the properties of the object contour.

3.3 Support vector machines

In machine learning and pattern classification, support vector machines are a supervised learning approach that has been demonstrated to perform well in numerous practical applications [17–19]. In two-class classification problem, the SVM's decision boundary is constructed from the training data by finding a separating hyperplane that maximizes the margins between the two classes; this is essentially a quadratic optimization problem. This learning strategy is shown to increase the generalization capability of the classifier. We can apply SVMs to complex nonlinear problems by using kernel methods and projecting the data onto a high-dimensional space. Apart from its good generalisation capability, the SVM approach works well when the number of training samples is small. The main challenge in applying SVMs is to find the appropriate features, the kernel function, and the training parameters. In this work, we will use SVM as the basic tool for classification of image features. Our main aim is to identify salient image features for the task of semantic image annotation.

4 Results and Analysis

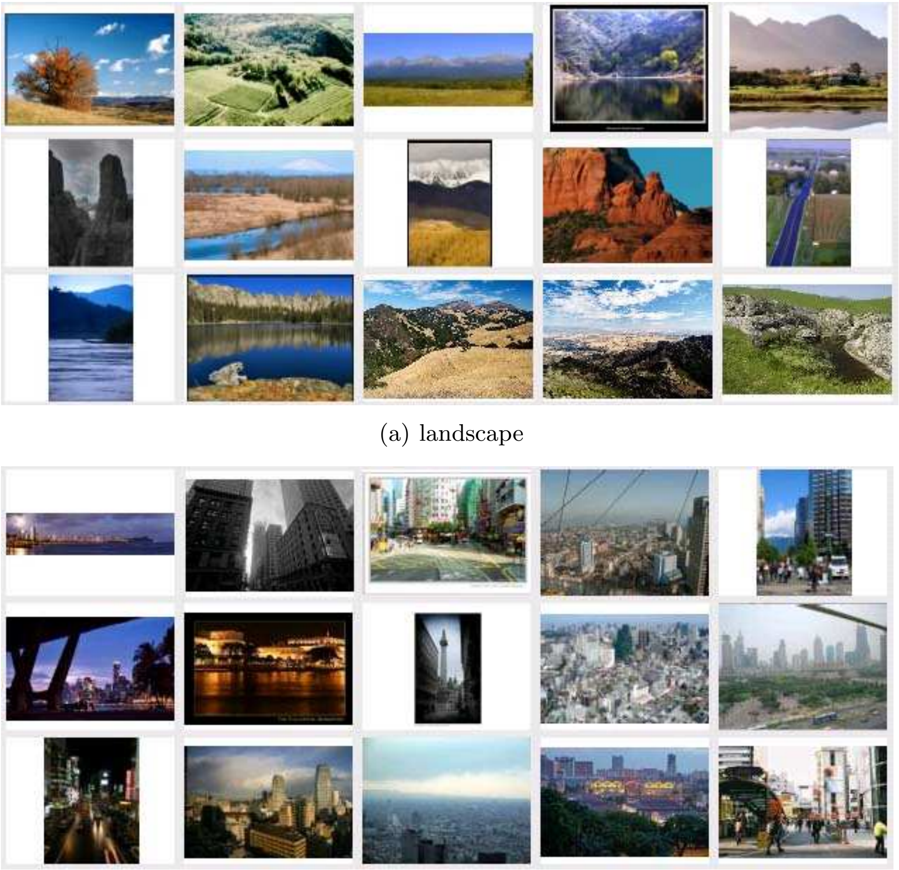
This section describes an application of the proposed image annotation system in classifying landscape and cityscape images. First, we describe the data collection and experimental procedure. Next, we present the classification results of different MPEG-7 visual descriptors and compare our system with an existing image classification system. Finally, we discuss techniques to improve the classification performance of the system.

4.1 Data preparation and experimental steps

Our work is based on a large-scale dataset of images. To provide classification results on real-world images, we have collected about $14,000$ images from online repositories of digital photos and manually annotated these images into four categories: landscape, cityscape, portrait and vehicle. For the task of classifying landscape versus cityscape images, we use a dataset of 3600 landscape images and 3600 cityscape images. These images vary widely in size, quality, and contents; some images even have blurred or monochrome (red=green=blue) appearance. Examples of the images are shown in Fig. 2. We use 4200 images for training and 3000 images for testing; the number of landscape images and cityscape images are equal.

In the experiments, the MPEG-7 reference software called *eXperimentation Model* (XM) [20] was used to extract most MPEG-7 visual descriptors. We used MATLAB to extract the dominant colour descriptor, because there was a bug in the XM software. To train and evaluate SVM classifiers, we chose an SVM library called *LIBSVM* [21], developed by Chang et al. at National Taiwan University. After trying different kernel functions, we selected the radial basis function kernel. We experimented with different SVM parameters: training cost c and kernel radius γ.

Because of the large number of MPEG-7 visual descriptors, we conducted a preliminary experiment on a small set of about 900 images to short-list the descriptors. We excluded from further analysis any descriptor that is computationally intensive or does not perform well. For example, the texture browsing descriptor for each image of size 600×800 pixels requires over 60 seconds to compute (using XM software on a P4 3GHz computer). This descriptor is an extension of the homogeneous texture descriptor, which can be computed more efficiently.



(a) landscape

(b) cityscape

Fig. 2. Example images in the dataset of 7200 images.

Furthermore, both region-based and contour-based shape descriptors performed poorly: after training, the system based on these descriptors misclassified most test images. There are two possible explanations for this result. First, robust image segmentation is still a challenging task; in the shape-based approach, failure to segment objects will impede the shape descriptors from representing the objects accurately. Second, landscape and cityscape images may contain many common shapes.

4.2 Comparison of MPEG-7 visual descriptors

In this experiment, we aim to identify the visual descriptors that perform well in the task of landscape and cityscape image classification. We constructed SVM classifiers that use each of the following MPEG-7 visual descriptors:

- dominant colour,
- colour layout
- scalable colour
- colour structure,
- homogeneous texture and
- edge histogram.

On the training set of 4200 images, the classification rates for the above descriptors are 84.5%, 80.9%, 89.3%, 99.0%, 85.5% and 93.1%, respectively. Note that the classification rates on the training set can change depending on two parameters: cost c and kernel radius γ. The training performance is reported for the parameter combination that gives the best performance on the test set.

The classification rates on the training set and test set for the scalable colour, colour structure and edge histogram descriptors are shown in Tables 1, 2, and 3, respectively. We have experimented with c values from 2^{-5} to 2^{15} and γ values from 2^{-15} to 8.

Table 1. Classification rates of the scalable colour descriptor on (*training set*, *test set*) for different SVM parameters.

γ c	2^{-7}	2^{-5}	2^{-3}	0.25	0.5	4
0.5	76.2/73.4	81.0/78.5	89.4/80.4	94.3/78.4	99.1/74.2	99.7/49.1
4	80.8/78.2	89.3/80.5	99.3/78.4	99.5/78.8	99.7/78.6	99.7/49.1
16	84.9/80.1	96.5/78.0	99.6/77.9	99.7/79.2	99.7/78.7	99.8/49.2
64	89.8/80.0	99.4/74.7	99.7/78.7	99.7/79.3	99.8/79.0	99.8/49.4
256	96.3/76.9	99.6/74.8	99.7/78.5	99.8/79.7	99.8/79.3	99.9/49.4
1024	99.4/73.4	99.7/75.9	99.8/77.8	99.7/79.3	99.8/79.0	99.9/49.4

Table 2. Classification rates of the colour structure descriptor (*training set*, *test set*) for different SVM parameters.

γ c	2^{-7}	2^{-5}	2^{-3}	0.25	0.5	4
0.5	76.3/72.8	79.7/77.7	84.6/80.0	86.9/80.8	90.7/82.6	100.0/55.6
4	79.3/76.3	84.0/79.4	90.9/81.7	95.6/82.3	99.0/82.8	100.0/72.6
16	82.0/78.1	86.9/80.7	95.9/81.7	99.1/81.7	99.9/82.3	100.0/72.6
64	84.3/79.7	90.9/80.9	99.0/80.7	99.9/81.4	100.0/82.5	100.0/72.6
256	87.1/80.4	95.8/81.0	99.9/80.2	100.0/81.4	100.0/82.2	100.0/72.6
1024	90.9/80.2	98.9/80.2	100.0/80.3	100.0/81.2	100.0/82.3	100.0/72.6

Classification performance, on the test set of 3000 images, of the six MPEG-7 visual descriptors is shown in Fig. 3. The dominant colour descriptor is the

Table 3. Classification rates of the edge histogram descriptor on (*training set*, *test set*) for different SVM parameters.

γ c	2^{-7}	2^{-5}	2^{-3}	0.25	0.5	4
0.5	84.3/81.6	86.7/83.2	90.1/84.4	92.8/84.5	95.3/83.5	100.0/50.4
4	86.4/83.2	89.5/84.3	97.7/84.5	99.9/84.4	100.0/84.9	100.0/50.4
16	87.8/83.7	93.1/84.6	100.0/83.8	100.0/84.2	100.0/84.9	100.0/50.4
64	89.7/84.4	97.4/83.8	100.0/83.5	100.0/84.2	100.0/84.9	100.0/50.4
256	93.3/84.2	99.9/81.9	100.0/83.5	100.0/84.2	100.0/84.9	100.0/50.4
1024	96.9/83.2	100.0/81.8	100.0/83.5	100.0/84.2	100.0/84.9	100.0/50.4

only one with a classification rate below 70%. There are three MPEG-7 visual descriptors that have a classification rate above 80%: scalable colour, colour structure and edge histogram. The edge histogram descriptor has the highest classification rate of 84.6%; this result shows that edge histogram is a salient visual feature in differentiating landscape and cityscape images.

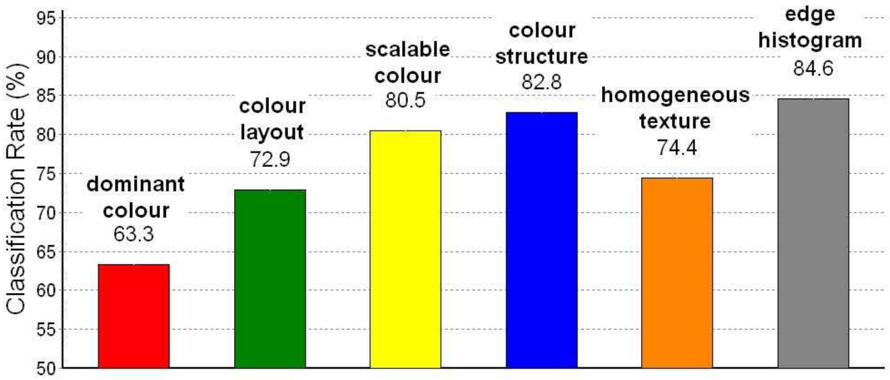


Fig. 3. Comparison of MPEG-7 visual descriptors in landscape versus cityscape image classification task, on a test set of 3000 images.

4.3 Comparison with other techniques

For comparison purposes, we study the performance, on the same dataset described in Section 4.1, of an image classification approach proposed in [7]. Vailaya et al. [7] use a weighted k-nearest neighbour classifier to differentiate cityscape versus landscape images and study several feature vectors. They find that the edge direction histogram (EDH) feature vector performs better compared to the others. An EDH feature vector has 73 elements. The first 72 elements are the normalized histogram of edge directions (72-bin). The last element is the normalized count of non-edge pixels.

We experimented with the k-NN classifier where k varies from 1 to 15; note that in [7], k goes from 1 to 9. Table 4 shows the classification results of Vailaya et al.'s method on our image dataset: the best classification rate achieved is 82.8% when k is equal to 7. This classification rate is similar to that of our colour structure descriptor (82.8%), and is lower compared to the CR of our edge histogram descriptor (84.6%).

Table 4. Classification rates of the k-NN classifier and the EDH feature.

Number of nearest neighbours k	1	3	5	7	9	11	13	15
Classification Rate (CR) (%)	80.4	81.9	82.7	82.8	82.6	82.7	82.7	82.6

4.4 Improving the system

So far, we have identified a number of MPEG-7 visual descriptors that are suitable for the task of classifying landscape and cityscape images. The system performance can be improved by combining these salient features and there are different approaches in doing so.

- **Using a single SVM**: We assemble all salient descriptors into a single feature vector and use only one SVM to classify the feature. When we combine the edge histogram and the colour structure in this way, the classification rate on the test set is 88.5%.
- **Using multiple SVMs**: We build individual SVMs that use separate visual descriptors and a final SVM to process the ensemble of confidence scores produced by the individual SVMs. The system implementing this approach has a CR of 89.7% on the test set. However, when we combine the best three salient descriptors, the classification rate increases to only 88.6%.

Our results show that combining salient features has a clear advantage to classification accuracy and is a promising research direction.

5 Conclusion

In this paper, we have presented an image classification system that combines MPEG-7 visual descriptors and support vector machines. We analysed a wide range of MPEG-7 visual descriptors including colour, edge, texture and shape. On a large dataset of 7200 landscape and cityscape photos, our system achieves a classification rate of 89.7%. We find that for landscape versus cityscape classification, the edge histogram and colour structure descriptors outperform other MPEG-7 visual descriptors and classification rate is improved by combining these two features.

References

1. F. Long, H. Zhang, and D. Feng. Fundamentals of content-based image retrieval. In D. Feng, W.C. Siu, and H.J.Zhang., editors, *Multimedia Information Retrieval and Management - Technological Fundamentals and Applications*. Springer, Berlin / Heidelberg, 2002.
2. A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1349–1380, 2000.
3. R. Datta, J. Li, and J. Z. Wang. Content-based image retrieval: approaches and trends of the new age. In *The 7th ACM SIGMM International Workshop on Multimedia Information Retrieval*, pages 253–262, 2005.
4. J. P. Eakins. Retrieval of still images by content. In *Lecture Notes in Computer Science: Lectures on Information Retrieval*, volume 1980/2001, pages 111–138. Springer, Berlin / Heidelberg, 2001.
5. E. C. Yiu. *Image classification using color cues and texture orientation.* PhD thesis, Massachusetts Institute of Technology, 1996.
6. M. Szummer and R. W. Picard. Indoor-outdoor image classification. In *IEEE International Workshop on Content-Based Access of Image and Video Database*, pages 42–51, 1998.
7. A. Vailaya, A. Jain, and H. J. Zhang. On image classification: city vs. landscape. In *IEEE Workshop on Content-Based Access of Image and Video Libraries*, pages 3–8, 1998.
8. A. Vailaya, M. Figueiredo, A. Jain, and H. J. Zhang. Content-based hierarchical classification of vacation images. In *IEEE International Conference on Multimedia Computing and Systems*, volume 1, pages 518–523, 1999.
9. R. Lienhart and A. Hartmann. Classifying images on the web automatically. *Journal of Electronic Imaging*, 11(4):445–454, 2002.
10. G. H. Hu, J. J. Bu, and C. Chen. A novel bayesian framework for indoor-outdoor image classification. In *International Conference on Machine Learning and Cybernetics*, volume 5, pages 3028–3032, 2003.
11. R. C. Gonzalez and R. E. Woods. *Digital image processing.* Prentice Hall, New York, 2002.
12. R. C. Gonzalez, R. E. Woods, and S. L. Eddins. *Digital image processing using MATLAB.* Prentice Hall, New York, 2004.
13. B. S. Manjunath, P. Salembier, and T. Sikora, editors. *Introduction to MPEG-7: Multimedia content description interface.* Wiley, Milton, 2002.
14. MPEG-7 Video Group. Text of ISO/IEC 15938-3/FDIS information technology - Multimedia Content Description Interface - Part 3 Visual. In *ISO/IEC JTC1/SC29/WG11/N4358*, Sydney, 2001.
15. F. Nack and A. T. Lindsay. Everything you wanted to know about MPEG-7, part 1. *IEEE Multimedia*, 6(3):65–77, 1999.
16. F. Nack and A. T. Lindsay. Everything you wanted to know about MPEG-7, part 2. *IEEE Multimedia*, 6(4):64–73, 1999.
17. C. J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, 1998.
18. N. Cristianini and J. Shawe-Taylor. *An introduction to support vector machines and other kernel-based learning methods.* Cambridge University Press, Cambridge, 2001.

19. S. Abe. *Support vector machines for pattern classification.* Springer, New York, 2005.
20. Institute for Integrated Systems. *MPEG-7 eXperimentation Model (XM)*, 2005. Software available at `http://www.lis.e-technik.tu-muenchen.de/research/bv/topics/mmdb/mpeg7.html`.
21. C. C. Chang and C. J. Lin. *LIBSVM : a library for support vector machines*, 2007. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.