2007

Automatic annotation of digital images using colour structure and edge direction

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

Automatic annotation of digital images using colour structure and edge direction

Abstract

The focus of this paper is on automatic annotation for semantic image retrieval. This work is aimed at identifying visual descriptors that are most relevant, effective and suitable for semantic annotation tasks. We propose an image annotation system based on support vector machines and a combination of descriptors that includes a gradient direction histogram and several MPEG-7 visual descriptors. The system is tested on a large database of 7200 cityscape and landscape images. The results indicate that when descriptors are used individually, the proposed gradient direction histogram performs best. However, when descriptors are combined, the accuracy is improved. The presented results confirm that combining the gradient direction histogram and colour structure produces the best results.

Disciplines

Physical Sciences and Mathematics

Publication Details

This conference paper was originally published as Shao, W, Naghdy, G and Phung, SL, Automatic annotation of digital images using colour structure and edge direction, 2007 IEEE International Conference on Signal Processing and Communications (ICSPC 2007), 24-27 November 2007, Dubai, United Arab Emirates. Copyright IEEE.

AUTOMATIC ANNOTATION OF DIGITAL IMAGES USING COLOUR STRUCTURE AND EDGE DIRECTION

Wenbin Shao, Golshah Naghdy, Son Lam Phung

School of Electrical, Computer and Telecommunications Engineering
University of Wollongong

ABSTRACT

The focus of this paper is on automatic annotation for semantic image retrieval. This work is aimed at identifying visual descriptors that are most relevant, effective and suitable for semantic annotation tasks. We propose an image annotation system based on support vector machines and a combination of descriptors that includes a gradient direction histogram and several MPEG-7 visual descriptors. The system is tested on a large database of 7200 cityscape and landscape images. The results indicate that when descriptors are used individually, the proposed gradient direction histogram performs best. However, when descriptors are combined, the accuracy is improved. The presented results confirm that combining the gradient direction histogram and colour structure produces the best results.

Index Terms— image classification, SVMs, MPEG-7, gradient direction histogram

1. INTRODUCTION

Text-based image retrieval techniques are mainly based on manual annotation [1, 2]. Given the increasing amount of digital images and videos, manual annotation is greatly time-consuming. Furthermore, it relies heavily on the perception of the person who performs the annotation. Different people might come up with different annotations based on their viewpoints of what are the most prominent content in the image. Content-based image retrieval (CBIR) is a promising approach to overcome the increasing challenge of multi-media management [1, 3, 4].

Many current CBIR systems accept queries that are based on low-level features, shape sketches or sample images. However, this approach is not intuitive and people are better at describing an image with keywords. The main challenge in CBIR is to bridge the semantic gap between the low level features and high level contents. Automatic annotation at semantic level is a powerful approach which generates keywords to describe an image.

In this paper, we propose a combination of a new feature called gradient direction histogram and several MPEG-7 visual descriptors. A number of combination strategies are explored. The proposed system is applied to annotate cityscape and landscape images. This paper is organised as follows. In Section 2, we review existing techniques for classifying images. In Section 3, we describe the proposed system with different combination schemes. In Section 4, we describe the proposed gradient direction histogram and two MPEG-7 visual descriptors. In Section 5, we discuss constructing SVM classifiers. In Section 6, we present and analyse the experimental results. Finally in Section 7, we give the concluding remarks.

2. BACKGROUND

There are many works that deal with bridging the gap between low-level features and high-level contents [5, 6, 7, 8, 9, 10]. Vailaya et al. [7] use a hierarchical architecture, based on Bayesian classifiers, to annotate images. They first separate indoor from outdoor images and then divide outdoor images into subcategories such as sunset, forest and mountain. On the classification of indoor versus outdoor images, Vailaya et al. report an overall classification rate of 90.5%. They find that features based on spatial colour distribution perform better than colour and texture features. On the classification of sunset versus forest and mountain images, they conclude that colour histogram is better than the edge direction features.

Dorado and Izquierdo [8] suggest a semi-automatic approach for image annotation that uses a low-level feature called fuzzy colour signature descriptor. In their approach, there is an image dataset that is manually annotated. Given a new image, fuzzy colour signature is extracted to find similar images from the annotated dataset. The keywords for the new images then are extracted from the similar images. Dorado and Izquierdo argue that the proposed semi-automatic annotation process can enhance CBIR query results and assist the interaction between users and CBIR systems.

Rui et al. [9] propose an approach that is based on pairwise constrained clustering and semi-naïve Bayesian model for automatic annotation. Their approach is a three-step pipeline: (i) image component decomposition, (ii) image content representation, (iii) and content classification. They argue that clustering-based approaches are significantly influenced by clustering quality, and the assumption that semantic objects in the same image are mutually

1-4244-1236-6/07/$25.00 © 2007 IEEE

independent is not reliable. To address these problems, they investigate a pair-wise constrained clustering algorithm and the semi-naïve Bayesian model. Their results show that the combination of these two techniques outperforms both techniques when used individually. In their experiments, Rui et al. choose the cross-media relevance model and probabilistic SVM as the baselines for comparison. The experiments are based on $4,850$ Corel images, of which 90% images are for training and 10% are for testing. The experimental results show that, compared to the two baselines models, their approach with language model constraints can improve the system by over 38% in terms of F_1 measure.

3. PROPOSED APPROACH

We propose an image annotation system that combines three different visual descriptors, based on colour and texture, feeding a number of support vector machines in different configurations. The block diagram of the proposed system is shown in Fig. 1. MPEG-7 visual descriptors and the gradient direction histogram are extracted from the image first (descriptor A, B, and C). Support vector machines are used in a number of different combinations to classify the images into different categories such as landscape and cityscape.

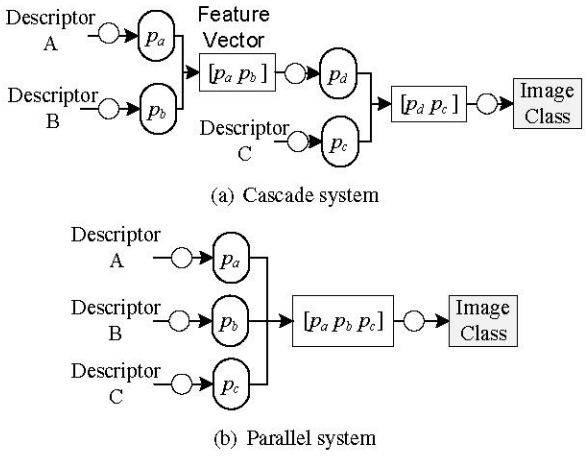


Figure 1. Proposed system models. ◯ represents SVM classifiers; P_a, P_b, P_c and P_d are confidence scores.

4. VISUAL DESCRIPTORS

In this section, we present the descriptors used including gradient direction histogram (GDH) and two MPEG-7 visual descriptors. There are a large number of MPEG-7 descriptors that can be used for classification purposes. The results from earlier work [11] in assessing the best features, confirms that the colour structure and edge histogram are the most effective descriptors. Therefore, two MPEG-7 descriptors based on colour and edge are used in this work.

4.1. Gradient direction histogram

The proposed *gradient direction histogram* is a normalized histogram of gradient directions computed across all edge pixels in the image. This feature is computed as follows. First, we apply an edge operator to calculate the edge magnitude along the horizontal and vertical direction. In this paper, we use the Prewitt operators [12]. For each edge pixel, the gradient angle is calculated as

$$\theta = \arctan \frac{G_y}{G_x} \tag{1}$$

where G_y and G_x are the edge magnitude along the vertical and horizontal direction, respectively.

4.2. MPEG-7 visual descriptors

Multimedia Content Description Interface, generally known as MPEG-7, is a standard for multimedia content description for a broad range of applications involving image, video and audio search [13, 14]. Two MPEG-7 visual descriptors: *colour structure* (CS) and *edge histogram* (EH) are used in this paper.

Colour structure is a colour structure histogram that consists of the information of colour distribution and spatial colour structure. Edge histogram describes the local spatial distribution of edges in an image. It is extracted from 16 sub-images of an image.

5. CONSTRUCTING CLASSIFIERS

In this work, SVMs are used as the basic tool for classification of image features. 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 [15, 16, 17]. Support vector machines are formulated for two-class classification problems. In SVMs, the 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 non-linear problems by projecting the data onto a high-dimensional space and using kernel methods.

Mathematically, given a training set $\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_L, y_L)\}$ where $y_i \in \{1, -1\}$ and $\mathbf{x}_i$ is a vector of n elements. Usually, the input samples are mapped to another higher-dimensional space using a function $\phi(\mathbf{x}_i)$. In the new space, a linear separating hyperplane that produces maximum separation between the two classes can be found by minimizing

$$Q(w, b, \epsilon) = \frac{1}{2}\mathbf{w}^T\mathbf{w} + c \cdot \sum_{i=1}^{L} \epsilon_i \tag{2}$$

subject to constraints

$$\{\mathbf{w}^T \phi(\mathbf{x}_i) + b\} \cdot y_i \geq 1 - \epsilon_i \text{ and } \epsilon_i \geq 0$$

where $\mathbf{w}$ is a vector perpendicular to the optimal separating hyperplane, b is a bias term, ϵ_i is a non-negative slack variable and c is a learning cost. The learning cost represents a compromise between margin maximization and classification error minimization. Note that $H(\mathbf{x}_i, \mathbf{x}_j) = \langle \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j) \rangle$ is the kernel function. The radial basis function kernel used in our work is defined as

$$H(\mathbf{x}_i, \mathbf{x}_j) = e^{-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2} \quad (3)$$

where γ is the kernel radius, $\gamma > 0$.

5.1. Determining SVM training parameters

Finding proper training parameters is a challenging task in classifier design and evaluation. The k-fold cross validation method is used in the proposed system [18]. In the cross validation, the training set is divided into k partitions. In each training turn, $(k-1)$ partitions are used to train, and the remaining partition is used to validate the classifier. This step is repeated k times until all partitions have been evaluated. Finally, the average classification rate across k folds is calculated. The parameters with the highest classification rate are selected for constructing the SVM classifier. The classifier is evaluated on the test set.

6. RESULTS AND ANALYSIS

In this section, we describe an application of the proposed image annotation system in classifying landscape versus cityscape images.

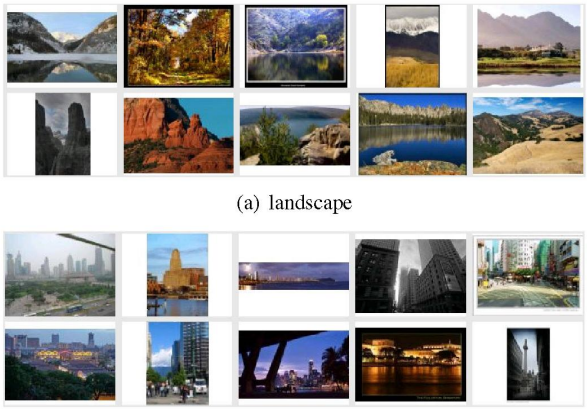
6.1. Data preparation

In this paper, we use a dataset of 3600 landscape images and 3600 cityscape images for the task of differentiating landscape versus cityscape images. These images vary widely in size, quality and contents. There are some images that are blurred or have perceptual monochrome appearance included in the database. Figure 2 shows a number of images from the dataset. We use 4200 images for training and 3000 images for testing with equal number of landscape and cityscape images.

6.2. Experimental steps

In the experimental process, the MPEG-7 reference software called *eXperimentation Model* (XM) [19] was used to extract the two MPEG-7 visual descriptors. To train and evaluate SVM classifiers, we chose a SVM library called *LIBSVM* [18], developed by Chang et al. at National Taiwan University.

After trying different kernel functions, we selected the radial basis function kernel. A number of experiments were conducted where the performance of each descriptor is evaluated on its own and then in combination with other descriptors. All classifier parameters are found through five-fold cross validation.



(a) landscape

(b) cityscape

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

6.3. Performance of each descriptor

Initially each descriptor is tested separately. Results are shown in Table 1. Among the three features the GDH performs best, and MPEG-7 EH outperforms CS.

Table 1. Classification rates of each descriptor.

	Classification Rate (CR)
Proposed GDH	87.2%
MEPG-7 EH	84.4%
MPEG-7 CS	82.3%

6.4. Cascaded SVMs

In this approach the SVMs are cascaded as shown in Fig. 1a. Results are shown in Table 2. The combination of GDH, CS and EH has the same CR with that of EH, CS and EDH. Note that the CR of GDH, EH and CS combination has the same performance as individual GDH.

Table 2. Classification rates of cascade SVMs.

Descriptor A	GDH	GDH	EH
Descriptor B	CS	EH	CS
Descriptor C	EH	CS	GDH
CR	89.9%	87.3%	89.9%

6.5. Parallel SVMs

When the system is configured as shown in Fig. 1b, a classification rate of 87.8% is achieved. However, if we use only two descriptors, the system performance is significantly improved. The results for different combinations of two descriptors are shown in Table 3.

The results presented above show that parallel SVMs achieve a higher classification accuracy than cascaded SVMs. With parallel SVMs, three descriptors do not achieve the best classification rate. In our work, two-descriptor combination is superior to three-descriptor combination. By

Table 3. Classification rates of two-descriptor combinations in parallel SVMs.

Descriptor A	GDH	GDH	EH
Descriptor B	EH	CS	CS
CR	89.9%	91.6%	89.7%

combing gradient direction histogram and MPEG-7 colour structure in parallel SVMs, the system can achieve a classification rate of 91.6%.

7. CONCLUSION

In this paper, an image annotation system that combines salient visual descriptors and support vector machines is presented. A wide range of descriptor combinations are explored. On a dataset of 7200 landscape and cityscape photos, our system achieves a classification rate of 91.6% by combining gradient direction histogram and a colour based descriptor. Our results show that combining salient features can improve classification accuracy significantly and this is a promising research direction.

8. REFERENCES

[1] Fuhui Long, Hongjiang Zhang, and David D. Feng, "Fundamentals of content-based image retrieval," in *Multimedia Information Retrieval and Management - Technological Fundamentals and Applications*, D. Feng, W.C. Siu, and H.J.Zhang., Eds. Springer, Berlin / Heidelberg, 2002.

[2] Y. Chen, J. Li, and J. Z. Wang, *Machine Learning and Statistical Modeling Approaches to Image Retrieval*, Kluwer Academic Publishers, New York, 2004.

[3] 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*, vol. 22, no. 12, pp. 1349–1380, 2000.

[4] Ritendra Datta, Jia Li, and James Z. Wang, "Content-based image retrieval: approaches and trends of the new age," in *Proceedings of the 7th ACM SIGMM International Workshop on Multimedia Information Retrieval*, Hilton, Singapore, 2005, ACM Press.

[5] M. Szummer and R. W. Picard, "Indoor-outdoor image classification," in *IEEE International Workshop on Content-Based Access of Image and Video Database*, 1998, pp. 42–51.

[6] A. Vailaya, A. Jain, and Hong Jiang Zhang, "On image classification: city vs. landscape," in *Proceedings of IEEE Workshop on Content-Based Access of Image and Video Libraries*, 1998, pp. 3–8.

[7] A. Vailaya, M. Figueiredo, A. Jain, and Hong Jiang Zhang, "Content-based hierarchical classification of vacation images," in *IEEE International Conference on Multimedia Computing and Systems*, 1999, vol. 1, pp. 518–523 vol.1.

[8] A. Dorado and E. Izquierdo, "Semi-automatic image annotation using frequent keyword mining," in *7th International Conference on Information Visualization*, 2003, pp. 532–535.

[9] 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 *11th International Multimedia Modelling Conference*, 2005, pp. 322–327.

[10] Rainer Lienhart and Alexander Hartmann, "Classifying images on the web automatically," *Journal of Electronic Imaging*, vol. 11, no. 4, pp. 445–454, 2002.

[11] W. Shao, G. Naghdy, and S. L. Phung, "Automatic image annotation for semantic image retrieval," in *Lecture Notes in Computer Science: VISUAL2007*, vol. 4781, pp. 372–381. Springer-Verlag, Berlin Heidelberg, 2007.

[12] Rafael C. Gonzalez and Richard E. Woods, *Digital image processing*, Prentice Hall, 2002.

[13] B.S. Manjunath, Phillipe Salembier, and Thomas Sikora, Eds., *Introduction to MPEG-7: multimedia content description interface*, Wiley, Chichester, 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] Christopher J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 121–167, 1998.

[16] Nello Cristianini and John Shawe-Taylor, *An Introduction to support vector machines and other kernel-based learning methods*, Cambridge University Press, Cambridge, 2001.

[17] Shigeo Abe, *Support vector machines for pattern classification*, Springer, New York, 2005.

[18] Chih-Chung Chang and Chih-Jen Lin, *LIBSVM : a library for support vector machines*, 2007, Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[19] 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.