Braving the Semantic Gap: Mapping Visual Concepts from Images and Videos

Da Deng

Department of Information Science, University of Otago, New Zealand
ddeng@infoscience.otago.ac.nz

Abstract. A set of feature descriptors have been proposed and rigorously in the MPEG-7 core experiments. We propose to extend the use of these descriptors onto semantics extraction from images and videos, so as to bridge the semantic gap in content-based image retrieval and enable multimedia data mining on semantics level. A computational framework consisting of a clustering process for feature mapping and a classification process for object extraction is introduced. We also present some preliminary results obtained from the experiments we have conducted.

1 Introduction

Seeing is believing, and vision is understanding. For the research on image processing and computer vision, the importance of image analysis, a process spanning from data pre-processing, feature extraction, towards object detection or even image understanding, has never been overlooked. However, despite the rapid theoretical advances observed in relevant research areas such as artificial intelligence, pattern recognition, and more recently machine learning, no significant breakthrough has been achieved in the modeling, manipulation and understanding of image contents.

In the early 1990s, content-based image retrieval (CBIR) [1] was proposed to overcome the limitation of the traditional annotation-based retrieval systems for images and videos. Aimed at effective multimedia asset management and efficient information retrieval, a typical content-based image retrieval system (e.g., [2]) operates basically on low-level visual features such as color, texture, shape or regions. While CBIR revived the research in image analysis and multimedia representation to some extent, it is generally understood that the problem of effective image retrieval is still far from being solved. The similarity of image contents can vary on different levels - locally or globally, on different characteristics, or on account of different psychological effects. Even though techniques such as joint histograms, image classification and relevance feedback have been investigated to more or less improve the retrieval quality, there is still a persisting gap - on one side is the lack of semantic representation in image and video data, but on the other, our capability in deriving meaningful semantics from the varying and multi-dimensional information in images and videos remains rather

P. Perner (Ed.): ICDM 2004, LNAI 3275, pp. 50–59, 2004.

limited. Such a gap implies that significant challenges still exist in areas such as image understanding, image data mining, and image retrieval.

Recently research attempts in bridging the semantic gap are becoming a trend. In [3], color semantics of art works are used for image retrieval, investigating into perceptual concepts on as color qualities and sensation, such as warmth, harmony, and anguish. In [4] support vector machines are used to match image feature clusters onto visual concepts.

On the other hand, from the data mining point of view, image mining has been proposed for knowledge discovery from data clustering and mining association rules [5]. It has been observed that automatic image analysis is in general very difficult, and much effort is required in finding an adequate feature scheme. Therefore most approaches remain domain-limited or require human expert interaction.

Previously we proposed to use CBIR techniques to tackle the problem of image collection profiling and comparison [6]. Feature maps were employed for self-organized profiling of image collections while in the meantime providing an effective graphical user interface for image or video collection navigation. Some distance measures were also proposed to quantitatively assess the similarity of these neural profiles.

In this paper, we extend this approach to investigating the plausibility of building a unified framework that can handle visual query and browsing of image/video contents, detect visual concepts from raw image data, or find interesting patterns or objects. We hope to use this framework to achieve data mining from images and videos on the semantics level, and leverage the semantic retrieval capability of image retrieval systems.

2 Computational Framework

Psychophysical findings have shown that high dimensional semantic structures can be effectively captured in low dimensions on the surface of a neural map [7]. This leads to our motivation of adopting a unified approach to achieve content-based retrieval as well as extraction of high-level semantic structures using labeled feature maps. Visual features are processed in two ways. Firstly clustering analysis of the feature maps can help to extract visual concepts shared by similar visual samples, and then effective classifiers can be built on these feature schemes to allow visual object extraction.

To enable low-level CBIR, global feature codes such as color histogram and texture energy histogram generated by Gabor filters are used. On the other hand, images are segmented into homogeneous regions using the technique in [8] and local feature codes are extracted by sampling the image regions. This not only enables image query using local visual clues, but also allows related visual concepts to be found through clustering analysis of regional visual features.

Upon satisfactory validation of the feature schemes, special objects such as ‘grass’, ‘trees’, and texts can then be detected and recognized by classifiers trained over the selected feature codes.

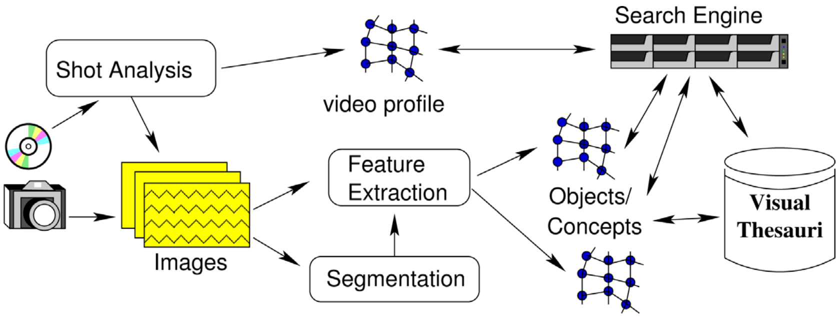

Fig. 1. The system diagram

The system diagram of the computational framework is shown in Fig.1. Three core components in the framework are explained as follows.

2.1 Shot Analysis

Video clips usually consist of a large number of image frames, and this huge volume of image data has to be reduced before further analysis is attempted. The solution is video segmentation and key frame extraction. Video shot boundary detection is largely based on the color histogram method used in image retrieval research. Color histograms of consecutive frames are compared. If the difference overflows a specified threshold, then a shot boundary is marked. Key frames can then be extracted as the representative image within each shot, and further image analysis is done on these key frames. Apart from key frame extraction, motion patterns of the video frames can also be analyzed and this has been found to be helpful for video summarization and retrieval especially for sports videos.

2.2 Feature Extraction

All feature schemes used in this study are based on the XM implementation of Core Experiments on the MPEG Video Group standard MPEG-7 [9], where a group of feature descriptors have been explicitly defined and rigorously tested. Due to the intensive research work in the field of CBIR, various descriptors on color and texture features based on global or local histograms have become mature. We briefly describe several feature descriptors used in our research as follows.

Color Layout Descriptor (CLD). CLD captures the spatial layout of the dominant colors on a grid superimposed on the image or the region of interest. This can be implemented as first dividing the image into 64 (8×8) blocks, and then deriving the average color of each block (or using dominant colors).

Color Structure Descriptor (CSD). CSD expresses local color structure in an image using an 8×8-structuring element. It is implemented as first scanning the image by an 8×8 pixel block, counting the number of blocks containing each color, and then generating a color histogram using the hue-min-max-difference (HMMD) color space. CSD has been found to be quite useful for the retrieval of natural images.

Homogeneous Texture Descriptor (HTD). HTD presents a quantitative characterization of texture for similarity-based image-to-image matching. It can also be used for object-based image retrieval. The computation process is as follows -

1. Partitioning the frequency domain into 30 channels (each modeled by a 2D-Gabor function);
2. Computing the energy and energy deviation for each channel;
3. Computing mean and standard variation of frequency coefficients, resulting a feature code of 62 dimensions:

$$F = \{f_{DC}, f_{SD}, e_1, ..., e_{30}, d_1, ..., d_{30}\}$$

Edge Histogram Descriptor (EHD). A given image is first sub-divided into 4×4 sub-images, and local edge histogram for each of these sub-images is computed. There are five categories of edges: horizontal, vertical, 45°, 135° diagonals, and isotropic. This will result in a descriptor of 80 bins. As commented in [9], unlike HTD, EHD is not appropriate for object based retrieval.

2.3 Features: Mapping and Classification

Among numerous clustering algorithms, SOM features the capability of carrying out vector quantization and multi-dimensional scaling simultaneously. It also has some attractive characteristics, such as readiness for visualization, probability density approximation of the data, and topology preserving. In [6] we proposed to use SOM for image collection navigation and also for collection profiling. For the later purpose some distance measures were introduced on top of the generated collection profiles trained by the SOM algorithm.

In this paper, the use of SOM is also two-fold: visual evaluation of visual feature schemes for conceptual clustering of image contents, as well as providing a user interface for semantic labeling. The visualization of the SOMs is generated from first carrying out two-dimensional principal component analysis (PCA) of the codebook and then using the projection from PCA as initial state for Sammon's mapping [10], which optimizes the low-dimensional visualization space by trying to preserve the distance order between prototypes through a gradient descent process.

We have found the use of k-Nearest Neighbor classifiers efficient in recognizing visual objects such as sky, grass, and pebbles etc. These classifiers work directly on the feature schemes validated in the previous clustering analysis phase.

3 Experiments and Results

Preliminary results have been obtained from experiments on several resources, including

- A digital image collection of the ISB Project at University of Otago, with site pictures taken over a two-year period. There are 154 images in total, including both indoor and outdoor scenes of various contents: construction sites, campus views, interior designs, and human close-ups;
- Video clips from the Roland Collection of Videos and Films on Art [11];
- Some CNN news video clips.

Algorithms for feature analysis, including SOM, PCA, Sammon's mapping and k-NN classification, are implemented in a Java-coded in-house *Exodus* package we developed for visual data mining purpose. Video processing and feature extraction utilities are coded in C and run on a FreeBSD system. Visual thesauri with sampled templates taken from image segments and all the feature maps generated are maintained using a PostgreSQL database system, based on which a search engine is to handle both content-based and semantic retrieval of images.

3.1 Video Processing

The UC Berkeley MPEG codec [12] is used and modified to decode and then analyze video shots in video clips. A simple global color histogram measured in the RGB space is employed to locate shot boundaries. The key frames are picked as the most stable shot frames (undergoing little change continuously) between the shot boundaries. As an example a CNN news clip 'Butler School' is tested with its key frames extracted.

We adopt a simplified 5-block CLD with average colors as key frame features, which then are used as input to train a SOM, so as to construct a 'snapshot' of the video clip. To preserve the timeline of the video stream, frame number is inserted into the feature code. By generating a one-dimensional map, the key frames can visualized over time while displaying some grouping of visual similarity, as shown in Fig.2. In Fig.3, a two-dimensional map is used instead. Although this profile lacks of indication of the timeline, the grouping on visual similarity is better displayed. There are, for example, clear indications of a cluster of human face close-ups on the lower left, and another cluster of maps on the upper right. Such a feature map will provide a useful interface for visual navigation, as well as for a guided process of concept exploration. On the other hand, by using some distance measure over the generated key frame based profiles, the search engine can compare different video clips quantitatively. This may provide an option for video clip retrieval by example.

3.2 Visual Concept Learning from Images

The major challenge to be met is however the mechanism of concept learning from visual contents. Our basic approach is to use feature maps to validate the effectiveness of feature scheme in clustering analysis, and then construct some classifier to carry out supervised learning and allow for object or concept detection. We have obtained some preliminary results from the experiments.

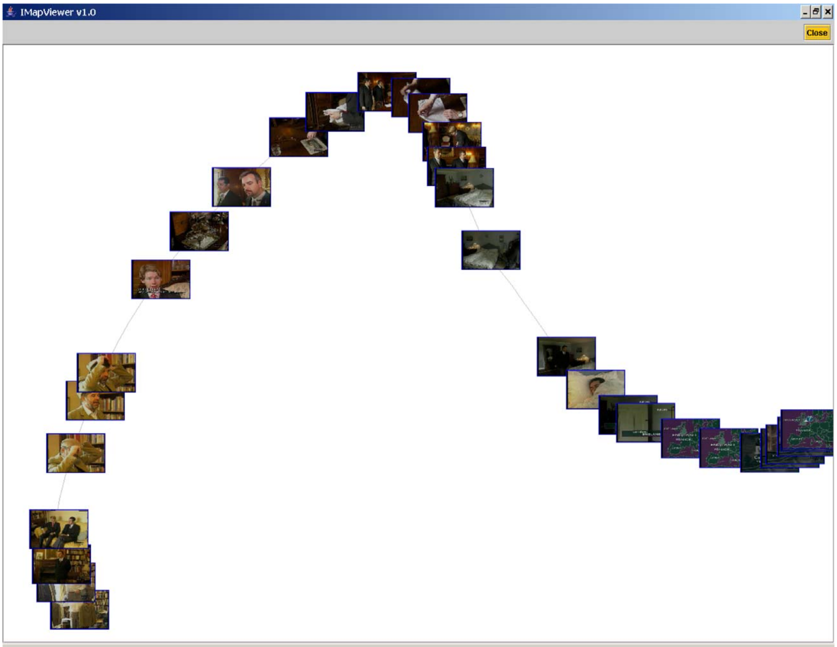

Fig. 2. Video summary of 'Butler School' constructed from a 1-D 'snake' map. Size of the map is 80

Artwork Versus Natural Scene. One experiment is done to distinguish images of artwork from those of natural scenes. By examining feature maps generated from key frames of two video clips from the Roland Collection, we find HTD features are quite promising to achieve the goal. Two Roland Collection videos are analyzed with key frames extracted and HTD features generated. Video clip No.13 is a short story on the artwork of a New Zealand artist, while No.603 is on prehistory sites with natural scenes. By training a SOM on HTD feature codes the frame images of artwork are clearly separated from those with natural scenes, as shown in Fig.4. The HTD features used are of 3 channels and 4 orientations, giving forth a HTD feature code of 14 dimensions (not including the energy gradients). Dimension reduction is then carried out using principal component analysis before generating the map.

Texture Thesaurus Mapping. In Fig.5 a feature map is trained for texture templates extracted from the Otago ISB image collection. A combined feature scheme consisting of both CLD and HTD is used to train the map. One can observe from the map that the *sky* templates are grouped on the lower right, and the *grass* templates grouped in the upper part. The *pebbles* template drifts to the lower left, while the *tree* templates mix either with *grass* or *carpet* templates in the middle. Upon visual assessment one may suggest the granularity of the texture is the major factor in setting forth such a layout. It can be seen from the

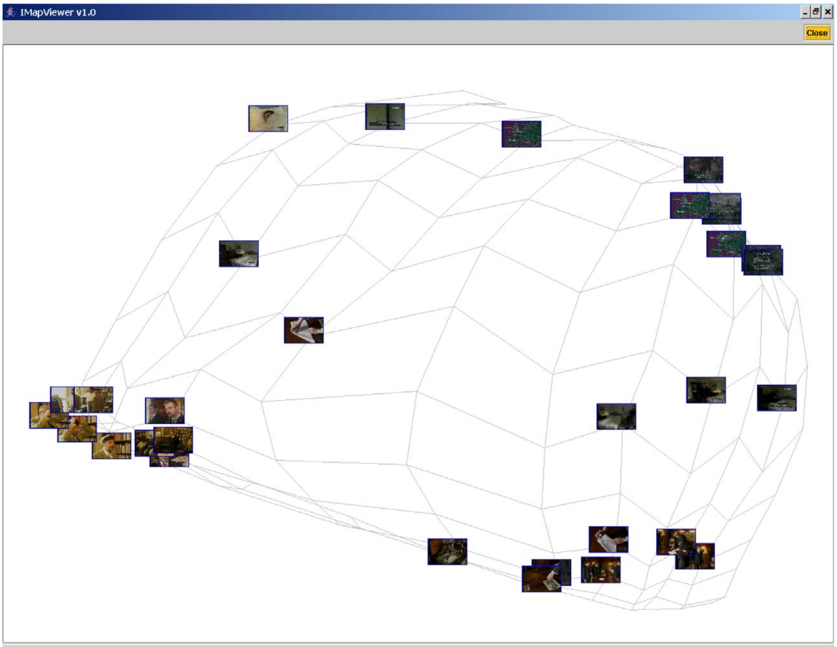

Fig. 3. Video summary visualized with PCA initialization. Size of the map is 12 × 12

figure that the granularity of these texture templates increases monotonously from upper right to the lower left of the map. The feature map can serve as a visual thesaurus for user to navigate through image collections with the help of a search engine operating on feature matching.

3.3 Classification

Once the feature scheme has been validated using a feature map, next thing one can do is to train some classifiers for concepts or objects in interest. The effectiveness of the feature scheme can then be further tested. If the outcome is satisfactory, an automatic process can then proceed to annotate image data with semantic contents.

To construct classifiers for the task, one needs to setup a database of templates with class labels. Because of the limited amount of labeled texture templates at the current stage we have only trained a simple k-nearest neighbor classifier for three classes, namely 'sky', 'grass', and 'others'. Table 1 gives the confusion matrix of the testing result on 30 templates extracted from the images.

The relatively poorer performance on grass recognition is mainly due to the lack of training templates. We hope to improve the segmentation quality and then automate the labeling process by flooding annotated templates across their corresponding segmented regions, so as to collect more training and testing image templates so that statistically sound validation results can be obtained and useful object detectors can be constructed using more robust classifiers.

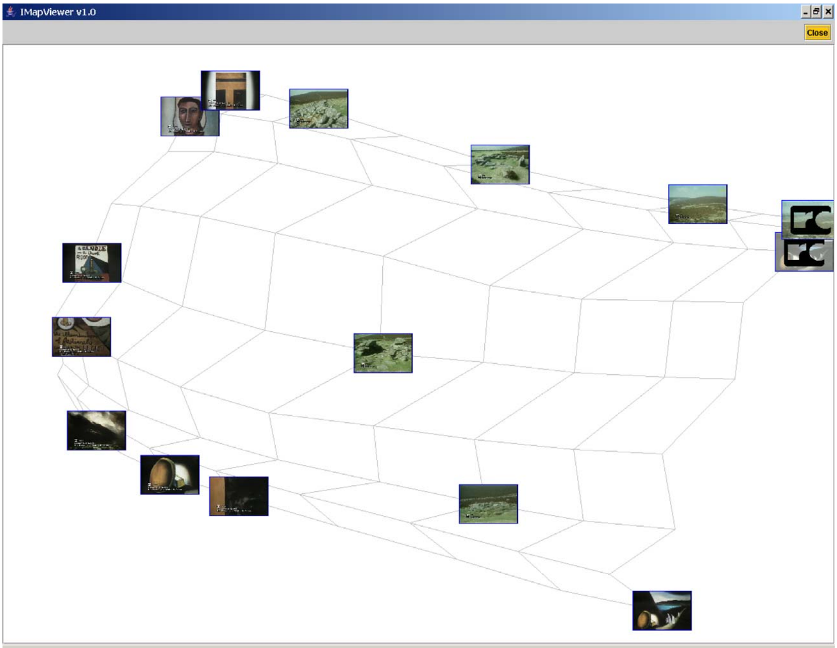

Fig. 4. Feature map trained on the HTD feature of RC video clip No.13 and No.603

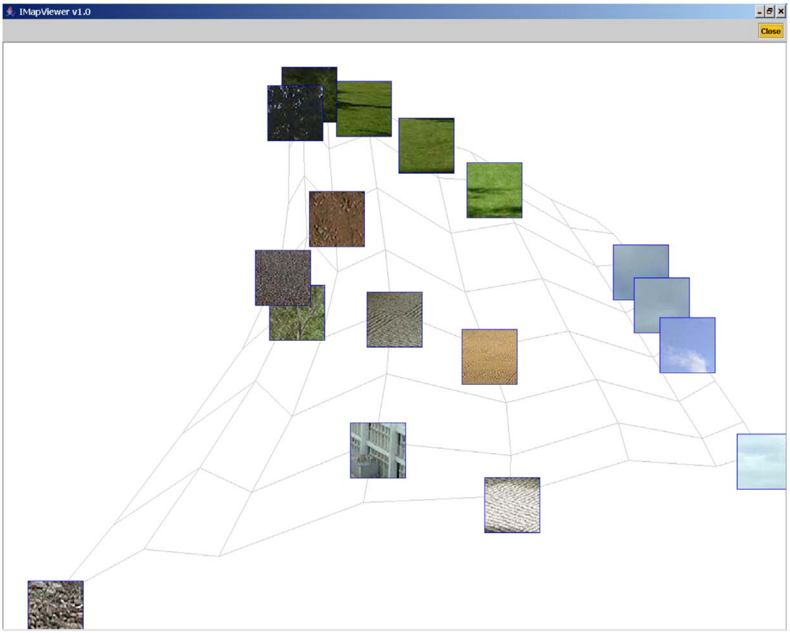

Fig. 5. Feature map trained on the combined CLD+HTD features of template images

Table 1. Confusion matrix of the classifier

Classes	Classified as		
	sky	grass	others
sky	85.5%	14.5%	0
grass	0	66.6%	33.3%
others	5%	0	95%

4 Discussion

In this paper, a computational approach is proposed to deal with the semantic gap acknowledged in image retrieval and data mining research. Self-organizing maps are used to produce profiles for video clips and to assist the exploration of various visual feature spaces proposed in the MPEG-7 core experiments. Semantic clusters related to some visual concepts or objects can emerge from the feature maps, whereupon classifiers can be constructed for accurate classification of visual concepts and objects.

Some preliminary results are reported in the paper, showing the capabilities of using the computational framework for image/video data navigation, as well as for visual discovery of concepts and objects. There are a number of unaddressed questions, such as

- How scalable is the current approach? This not only applies to the number of images or patterns being analyzed, but also the number of different objects or concepts that the current XM-based feature schemes can accommodate. For example, a simple task for human beings such as distinguishing indoor scenes from outdoor scenes [13] has found to be not easy to cope with. Further work needs to be done in improving the classification accuracy by examining their difference in lighting conditions combined with results from object detection.
- How to model the interrelationship between objects and concepts? When hierarchical feature maps are to be used for scalable feature clustering analysis, how does the map hierarchy guide concepts and objects related to the clusters formed on hierarchical maps to organize themselves into some semantic structure? Notably, there are some graphical representations of objects and concepts proposed, for instance, in [14].

We hope, in an incremental process our approach can lead to the construction of a mathematical framework for visual concepts and object extraction, as well as for the modeling of the interrelationship between these concepts, so that high-level multimedia data mining and semantic retrieval can eventually be achieved.

Acknowledgement. This work was supported by Otago Research Grant ORG 200200621, University of Otago, 2003.

References

1. Smeulders, A., Worring, M., Santini, S., Gupta, A., R., J.: Content-based image retrieval at the end of the early years. IEEE Transaction on Pattern Analysis and Machine Intelligence **22** (2000) 1349–1380
2. Carson, C., Thomas, M., Belongie, S., et al.: Blobworld: A system for region-based image indexing and retrieval. In: Proc. Int. Conf. Visual Inf. Sys. (1999) 509–516
3. Corridoni, J., Del Bimbo, A., Pala, P.: Image retrieval by color semantics. Multimedia Systems **7** (1999) 175–183
4. Wang, L., Manjunath, B.: A semantic representation for image retrieval. In: Proc. ICIP 2003, IEEE (2003) 523–526
5. Perner, P.: Image mining: issues, framework, a generic tool and its application to medical-image diagnosis. Eng. Appl. of Art. Intell. **15** (2002) 193–203
6. Deng, D.: Content-based comparison of image collections via distance measuring of self-organised maps. In: Proc. IEEE MMM 2004. (2004) 233–240
7. Edelman, S., Intrator, N.: Learning as formation of low dimensional representation spaces. In: Proc. 19th The Cognitive Science Meeting. (1997) 199–204
8. Deng, Y., Manjunath, B.: Unsupervised segmentation of color-texture regions in images and video. IEEE Trans. PAMI **23** (2001) 800–810
9. Manjunath, B., Ohm, J., Vinod, V., Yamada, A.: Color and texture descriptors. IEEE Trans. Circuits and Systems for Video Technology **2** (2001) 703–715
10. Sammon, W.: A nonlinear mapping for data analysis. IEEE Trans. on Computers **5** (1969) 401–409
11. RC: Rolland collection of videos and films for art (2004) `http://www.roland-collection.com/`
12. UCB: Berkeley mpeg tools (2004) `http:bmrc.berkeley.edu/frame/research/mpeg/`
13. Szummer, M., Picard, R.W.: Indoor-outdoor image classification. In: IEEE Intl. Workshop on Content-Based Access of Image and Video Databases, CAIVD, Bombay, India (1998) 42–51
14. Naphade, M., Kozintsev, I., Huang, T.: A factor graph framework for semantic video indexing. IEEE Trans. on Circuits and Systems for Video Tech. **12** (2002) 40–52