Recognition of Semantic Content in Image and Video

Punam R. Karmokar
School of Education Technology,
Jadavpur University, Kolkata

Ranjan Parekh
School of Education Technology
Jadavpur University, Kolkata

ABSTRACT
This paper addresses the problem of recognizing semantic content from images and video for content based retrieval purposes. Semantic features are derived from a collection of low-level features based on color, texture and shape combined together to form composite feature vectors. Both Manhattan distance and Neural Networks are used as classifiers for recognition purposes. Discrimination is done using five semantic classes viz. mountains, forests, flowers, highways and buildings. The composite feature is represented by a 26-element vector comprising of 18 color components, 2 texture components and 6 shape components.

General Terms
Semantic content recognition

Keywords
Color, Texture, Shape.

1. INTRODUCTION
Automated schemes for content based image retrieval (CBIR) are usually based on low-level features like color, texture and shape. Recognition of such low-level content has been extensively studied over the last couple of decades. However humans are more interested in image retrieval based on semantic or high-level content i.e. images or video frames need to be retrieved which are semantically similar to some given image, rather than visually. This is a more difficult proposition firstly because automated systems are not capable to directly recognizing semantic concepts, and secondly because semantic concepts are human abstractions and they have no fixed definitions e.g. a house or a car can have all types of colors, textures and shapes. For semantic content recognition therefore systems are first designed to retrieve a set of low-level features and then at a higher level a mapping is done to associate a set of low-level features with one or more high-level concept. This paper addresses the problem of recognizing high level concepts from images and video by utilizing a number of low-level features. The paper is organized as follows: section 2 provides an overview of related work, section 3 outlines the proposed methodology, section 4 provides details of the dataset and experimental results obtained, and section 5 provides the overall conclusions and the scope for future research.

2. PREVIOUS WORKS
A probabilistic framework for semantic video indexing, which can support filtering and retrieval and facilitate efficient content-based access is proposed in [1]. To map low-level features to high-level semantics it proposes probabilistic multimedia objects (*multijects*). Paper [2] proposes a novel framework to make some advances toward the final goal to solve these problems. The goal of [3] is retrieval paradigm (also called image-based search) is to resolve the problems in textual-based search. A system of operations on semi-structured data and how a sample query can be represented as an expression built from the operations have been described in [4]. Low level image histogram features are discussed in [5]. The main advantage of this method is quick generation and comparison of the applied feature vectors. In [6] the main intention is to bridge the video captions and the query term by traditional information retrieval techniques (IR). Paper [7] focuses on Internet video content and social networking to present solutions to the problem of gathering metadata describing user interaction, usage and opinions of that media. In [8] with the intention of retrieving video for a given query, the raw video data is represented by two different representation schemes, video segment representation (VSR) and Optimal key frame representation (OFR). A distributed video retrieval system with support for semantic queries named XUNET is described in [9]. In [10] a novel semantic video retrieval system that integrates web image annotation and concept matching function to bridge images, concepts and videos is designed. In [11] the authors show that integrating existing modalities along with the concept interactions can yield a better performance in detecting semantic concepts.

3. PROPOSED APPROACH
The proposed approach makes use of a number of low-level features based on color, texture and shape and then finally maps them to specific semantic classes. The features can be extracted from images or video frames.

3.1 Color Features
Color information is represented using six different color features derived from color histograms, namely, first order color moments, standard deviation, skew, precision, energy and entropy.

3.1.1 Color Histogram
The histogram H of an image I for a specific color c is the collection of frequency of pixels x for each bin i, the value of i ranging from 1 to n, n being the total number of bins.

$$H_c = \bigcup_{i=1}^{n} x_i \quad (1)$$

3.1.2 Color Moments
Color moments are calculated from the histogram by multiplying each bin number to the pixel frequency at that bin, and summing over all the product values.

$$M_c = \sum_{i=1}^{n} i * H_c(i) \quad (2)$$

3.1.3 Color Standard Deviation
Standard deviation is computed from the histogram by multiplying the pixel frequency at each bin level to its variance.

$$S_c = \sqrt{\sum_{i=1}^{n}(i - M_c)^2 * H_c(i)} \quad (3)$$

3.1.4 Color Skew

Skew is calculated by multiplying the pixel frequencies at each bin level to the third power of the difference with the color moments.

$$K_c = \sum_{i=1}^{n}(i - M_c)^3 * H_c(i) \quad (4)$$

3.1.5 Color Precision

Precision is calculated by taking the reciprocal of the standard deviation and multiplying the difference between the color moment and the maximum frequency value of the histogram.

$$L_c = \left(\frac{1}{S_c}\right)\{M_c - \max(H_c)\} \quad (5)$$

3.1.6 Color Energy

Color energy is calculated as the summation of the squares of the frequency values of the histogram at each bin level.

$$E_c = \sum_{i=1}^{n}\{H_c(i)\}^2 \quad (6)$$

3.1.7 Color Entropy

Color entropy is calculated from the color histogram where each value represents the pixel frequency at a particular bin level.

$$N_c = -\sum_{i=1}^{n} x_i . log_2 x_i \quad (7)$$

3.1.8 Color Combined Feature

The six feature values namely moment, standard deviation, skew, precision, energy and entropy is calculated for each of the three primary colors of the RGB color system namely red, green and blue. These values are then combined together into a composite color feature vector consisting of 6 × 3 or 18 elements as shown below, where $c = \{r, g, b\}$.

$$F_C = \{M_c, S_c, K_c, L_c, E_c, N_c\} \quad (8)$$

3.2 Texture Features

Texture information is represented using two features namely (1) the entropy of the gray intensity pixel values of the image and (2) standard deviation calculated from the frequency components of the image obtained as a result of Discrete Fourier Transform (DFT) decomposition.

3.2.1 Texture Entropy

Texture entropy is calculated by converting the color image into gray-scale and using the pixel intensity values.

$$N_t = -\sum_{i=1}^{k} p_i . log_2 p_i \quad (9)$$

3.2.2 Fourier Components

The frequency components of the gray scale image can be obtained by subjecting it to Fourier decomposition.

$$F(u, v) = \sum_{x=0}^{M-1}\sum_{y=0}^{N-1} f(x, y) . e^{-j2\pi\left(\frac{ux}{M} + \frac{vy}{N}\right)} \quad (10)$$

Standard deviation is calculated from the set of Fourier coefficients obtained after decomposition. Here y_i represents the i-th DFT coefficients and μ the mean coefficient value.

$$S_t = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \mu)^2} \quad (11)$$

3.2.3 Texture Combined Feature

The two features namely texture entropy and standard deviation of the Fourier coefficients are combined together into a composite texture feature vector consisting of two components.

$$F_t = \{N_t, S_t\} \quad (12)$$

3.3 Shape Features

Shape information in the image is represented using six shape features namely (1) shape moments (2) shape standard deviation (3) shape skew (4) shape precision (5) shape energy (6) shape entropy.

3.3.1 Edge Histogram

Edge histogram is calculated by subjecting the image I to edge detection using Sobel masks G_X, G_Y and calculating the edge gradient at each point. The set of edge gradient values are combined together into an edge histogram H_s, where each term of the histogram represents the frequency of the edge values.

$$G_X = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix}, G_Y = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} \quad (13)$$

$$S_X = G_X * I, S_Y = G_Y * I, \theta = tan^{-1}\left(\frac{S_X}{S_Y}\right) \quad (14)$$

$$H_s = \bigcup_{i=1}^{n} z_i \quad (15)$$

3.3.2 Shape Moments

Shape moments are calculated from the edge histogram as defined below.

$$M_s = \sum_{i=1}^{n} i * H_s(i) \quad (16)$$

3.3.3 Shape Standard Deviation

Shape standard deviation is calculated from the shape moments and the edge histogram as defined below.

$$S_s = \sqrt{\sum_{i=1}^{n} (i - M_s)^2 * H_s(i)} \quad (17)$$

3.3.4 Shape Skew

Shape skew is calculated from the shape moments and the edge histogram as defined below.

$$K_s = \sum_{i=1}^{n} (i - M_s)^3 * H_s(i) \quad (18)$$

3.3.5 Shape Precision

Shape precision is calculated from the reciprocal of the shape standard deviation, the shape moment and the maximum values of the edge histogram.

$$L_s = \left(\frac{1}{S_s}\right)\{M_s - \max(H_s)\} \quad (19)$$

3.3.6 Shape Energy

Shape energy is calculated by taking the square of the values from the edge histogram as defined below.

$$E_s = \sum_{i=1}^{n} \{H_s(i)\}^2 \quad (20)$$

3.3.7 Shape Entropy

Shape entropy is calculated from the edge histogram as defined below.

$$N_s = -\sum_{i=1}^{n} z_i . log_2 z_i \quad (21)$$

3.3.8 Shape Combined Feature

The composite shape feature vector is formed by combining the 6-elements containing shape information as shown below

$$F_s = \{M_s, S_s, K_s, L_s, E_s, N_s\} \quad (22)$$

3.4 Classification

The final feature is a combination of three features corresponding to color (18 elements), texture (2 elements) and shape (6 elements) and is represented as a 26-element vector.

$$F = \{F_c, F_t, F_s\} \quad (23)$$

Classification is done using two approaches: (1) Manhattan Distance (2) Neural Networks.

3.4.1 Manhattan Distance

The Manhattan distance (MD) between two n-element vectors A and B is defined as follows:

$$D_{MD}(A, B) = \sum_{i=1}^{n} |A_i - B_i| \quad (24)$$

Each class is represented by the set of its training vectors. The i-th class T_i is represented by the mean of the feature values of all its component samples.

$$T_i = \frac{1}{n}\{T_{i1}^F + T_{i2}^F + \cdots + T_{in}^F\} \quad (25)$$

The j-th test sample S_j with feature value S_j^F is classified to class k if the absolute difference $D_{j,i}$ between the j-th test sample and i-th training class is minimum for $i = k$.

$$S_j \rightarrow k, if\ D_{j,i} = \left|S_j^F - T_i\right|\ is\ minimum\ for\ i = k \quad (26)$$

3.4.2 Neural Network

Neural networks (NN) involving Multi-layer Perceptrons (MLP) with feed-forward back-propagation architectures were also employed for classification. A 26-26-5 architecture i.e. 26 input nodes, 26 nodes in the hidden layer and 5 output nodes was found to give best results with log-sigmoid activation functions for both neural layers, a learning rate of 0.01 and a mean square error (MSE) threshold of 0.01 for convergence. The MLP takes approximately 30000 epochs for convergence.

4. EXPERIMENTS AND RESULTS

The dataset consists of 125 images collected from the websites mentioned in [12] and [13], divided into 5 semantic classes viz. mountain, tree, flowers, highway, building. Out of 25 images of each class, 15 are used for training and 10 for testing. Samples of the images are shown below.

Fig 1: Samples of Training images

Fig 2: Samples of Testing images

The figure below shows the plots of the training set based on the 18-element color feature vector, for each of the five classes.

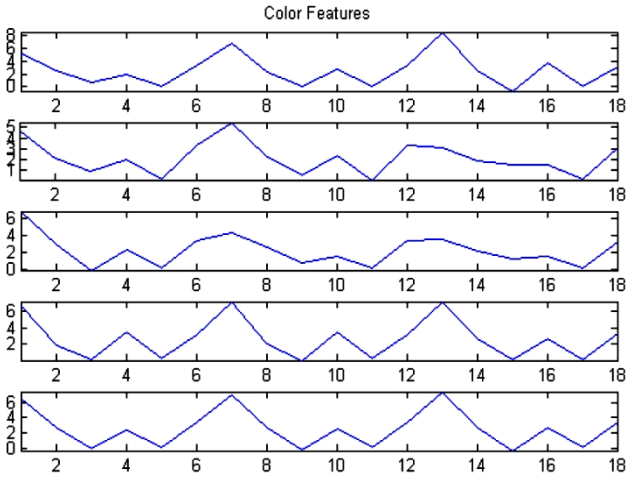

Fig 3: Color feature plots of training set

The figure below shows the plots of the training set based on the 2-element texture feature vector for each of the five classes

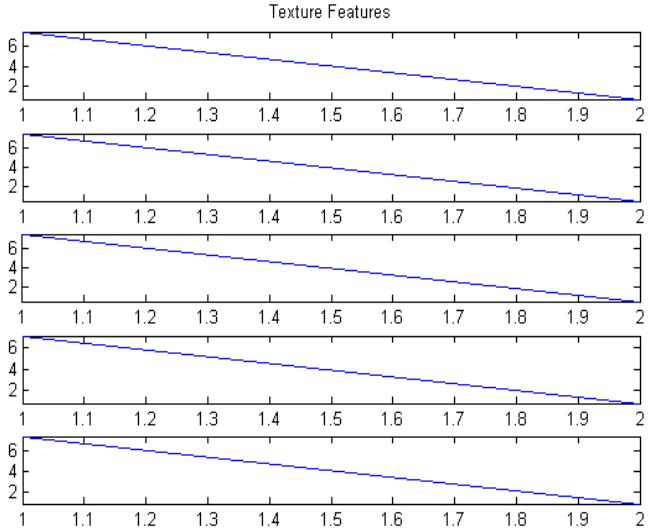

Fig 4: Texture feature plots of training set

The figure below shows the plots of the training set based on the 6-element shape feature vector for each of the five classes

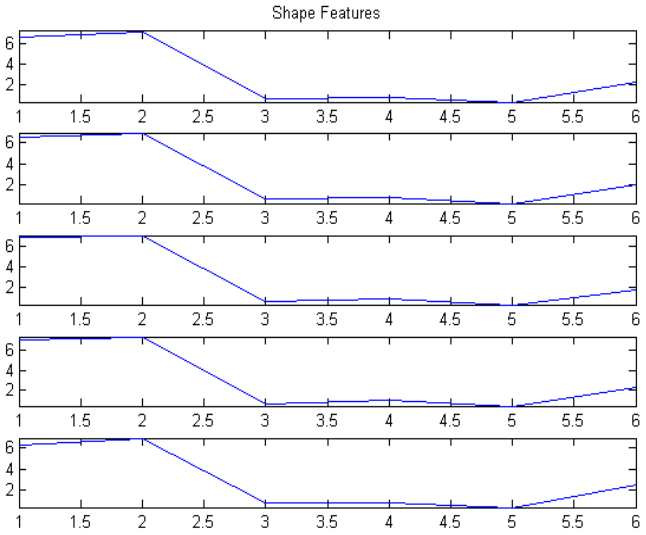

Fig 5: Shape feature plots of training set

The figure below indicates the classification plots based on the three types of features using Manhattan distance. The 50 test files (10 per class) are arranged along the horizontal axis and the values along the vertical axis depict their differences with the training files of each class as indicated in the legends.

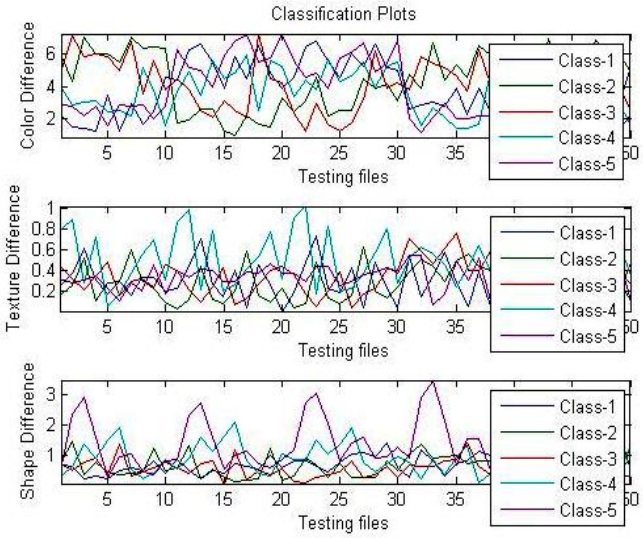

Fig 6: Classification plots of testing set using MD

The figure below shows the classification plots based on the neural network. The 50 test files (10 per class) are arranged along the horizontal axis and the values along the vertical axis depict their probabilities of belonging to the training classes as indicated in the legends.

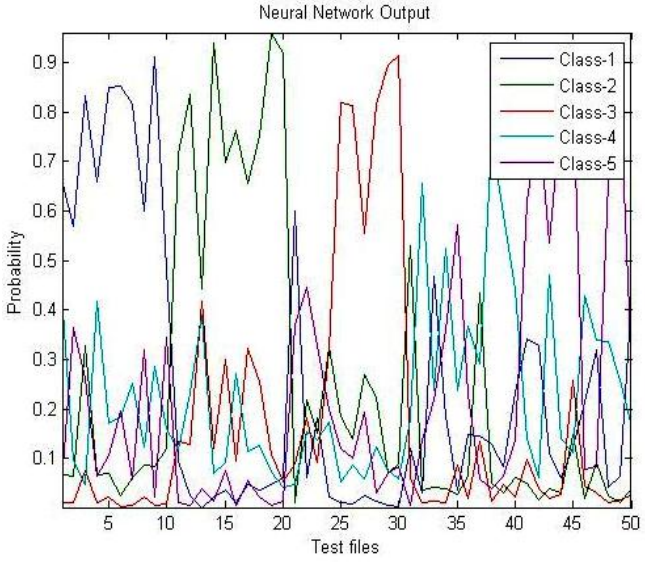

Fig 7: Classification plots of testing set using NN

The corresponding accuracy values are tabulated below. It indicates that the composite feature vector achieves an accuracy of 94% using MD and 84% using NN, while individual features produce 72% and 86% (color), 40% and 58% (texture), 60% and 66% (shape).

Table 1: Percentage Recognition Accuracies

Features	MD	NN
Color	72	86
Texture	40	58
Shape	60	66
Composite	94	84

5. CONCLUSIONS & FUTURE SCOPES

This paper proposes a system for recognizing semantic content from images and video frames using multiple features based on color, texture and shape. Both MD and NN are used as classifiers for testing the efficiency of the system. The Manhattan distance approach shows higher accuracy and has lower complexity thereby resulting in faster retrieval. Individually the color feature produces the highest accuracy. The Neural Network approach only gives good accuracy result with color feature but shape and texture is not satisfactory, which can be improved in future. And for probabilistic approach and complex procedure it's a bit time consuming.

This work can be extended further by incorporating a dynamic learning system wherein success fully retrieved videos can be reused to teach the system further thereby fine tuning the process. Here only image based semantic concept is taken in account and only five classes have been included. In future, more image classes and data samples can be included and audio features also can be considered. With other classifiers also the work can be experimented.

6. REFERENCES

[1] H. R. Naphide and T. S. Huang. "A probabilistic framework for semantic video indexing, filtering, and retrieval". Proc. of Multimedia, IEEE Transactions on, pp. 141 - 151, mar 2001.

[2] Jianping Fan, HangzaiLuo ; Elmagarmid, A.K., "Concept oriented indexing of video databases: toward semantic sensitive retrieval and browsing" , Proc. of Image Processing, IEEE Transactions on (Volume:13 , Issue: 7), on July 2004, pp. 974 – 992.

[3] Y. Peng and C. W. Ngo, "Clip-Based Similarity Measure for Query-Dependent Clip Retrieval and Video Summarization," Proc. of IEEE Transactions on Circuits and Systems for Video Technology, vol.16, no.5, pp.612-627, 2006.

[4] Al-Safadi, Getta, J.R, "Application of Semi-structured Data Model to the Implementation of Semantic Content Based Video Retrieval System", Mobile Ubiquitous Computing, Systems, Services and Technologies, 2007. Proc. Of UBICOMM '07. International Conference on 4-9 Nov. 2007, pp. 217 – 222.

[5] Szabolcs Sergyan Budapest Tech John, "Color Histogram Features Based Image Classification in Content-Based Image Retrieval Systems", von Neumann Faculty of Informatics Institute of Software Technology, B´ecsi ´ut96/B, Budapest, H-1034, Hungary IEEE. 2008.

[6] A. Haubold, A. (Paul) Natsev, "Web-based Information Content and its Application to Concept-Based Video Retrieval," Proc. of international conference on Content based image and video retrieval, pp.437-446, 2008.

[7] S. J. Davis, C. H. Ritz, "Using social networking and collections to enable video semantics acquisition", IEEE Multimedia, Oct-Dec 2009, pp. 52-60.

[8] Padmakala, S., AnandhaMala, G.S., Shalini, M. ,"An Effective Content Based Video Retrieval Utilizing Texture, Color and Optimal Key Frame Features", Image Information Processing (ICIIP), Proc. of 2011 International Conference on 3-5 Nov. 2011, pp. 1-6.

[9] Quan Zheng, Zhiwei Zhou, "An MPEG-7 compatible video retrieval system with support for semantic queries" Proc. of Consumer Electronics, Communications and Networks (CECNet), 2011 International Conference on 16-18 April 2011, pp. 1035 – 1041.

[10] Bo-Wen Wang, Ja-Hwung Su ; Chien-Li Chou ; Tseng, V.S., "Semantic Video Retrieval by Integrating Concept and Content-Aware Mining" Proc. of , Technologies and Applications of Artificial Intelligence (TAAI), 2011 International Conference on 11-13 Nov. 2011, pp. 32-37.

[11] Gulen, E., Yilmaz, T., & Yazici, A. "Multimodal Information Fusion for Semantic Video Analysis". International Journal of Multimedia Data Engineering and Management (IJMDEM), 3(4), 2012, pp. 52-74.

[12] Free Best Wallpapers [www.freebestwallpapers.info].

[13] Free Big Pictures [www.freebigpictures.com].