High-Level Concept Detection in Video Using a Region Thesaurus

Evaggelos SPYROU [a,1] and Yannis AVRITHIS [a]

[a] *Image, Video and Multimedia Systems Laboratory, School of Electrical and Computer Engineering, National Technical University of Athens*

Abstract. This work presents an approach on high-level semantic feature detection in video sequences. Keyframes are selected to represent the visual content of the shots. Then, low-level feature extraction is performed on the keyframes and a feature vector including color and texture features is formed. A region thesaurus that contains all the high-level features is constructed using a subtractive clustering method where each feature results as the centroid of a cluster. Then, a model vector that contains the distances from each region type is formed and a SVM detector is trained for each semantic concept. The presented approach is also extended using Latent Semantic Analysis as a further step to exploit co-occurrences of the region-types. High-level concepts detected are desert, vegetation, mountain, road, sky and snow within TV news bulletins. Experiments were performed with TRECVID 2005 development data.

Keywords. High-level feature detection, MPEG-7, TRECVID, Latent Semantic Analysis, Region Thesaurus, Region Types, Model Vectors

Introduction

The recent advances in telecommunication technologies, along with the World Wide Web proliferation, have boosted the wide-scale creation and dissemination of digital visual content. More specifically, during the last years, a tremendous increase on the number of video documents has been observed and many users tend to share their personal collections via many popular websites. However, this rate of growth has not been matched by a concurrent emergence of technologies to support efficient video retrieval and analysis. Thus, it appears very common for the average internet user to possess a large amount of digital information, without the ability to effectively browse and retrieve it. Moreover, the number of diverse, recently emerging application areas, which rely increasingly on image and video understanding systems, has further revealed the tremendous potential of the effective use of visual content through semantic analysis.

However high-level concept detection in video documents still remains an unsolved problem. As almost all of the typical recognition problems, this one also has two aspects. The first is the extraction of the various features of a video sequence, such as color,

[1]Corresponding Author: Evaggelos Spyrou, Image, Video and Multimedia Systems Laboratory, School of Electrical and Computer Engineering, National Technical University of Athens, 9 Iroon Polytechniou Str., 157 80 Athens, Greece; E-mail: espyrou@image.ece.ntua.gr.

texture, motion and audio a process commonly called low-level feature extraction and then form a description by combining them. The other aspect is the method used for assigning these low-level descriptions to high-level concepts, a problem that is often referred to as the Semantic Gap. Many approaches have been proposed that share the goal of bridging the semantic gap, thus allowing the proper extraction of high-level concepts of multimedia documents using and combining heterogeneous audiovisual features.

In [7], a prototype multimedia analysis and retrieval system is presented, that uses multi-modal machine learning techniques in order to model semantic concepts in video, from automatically extracted multimedia content. A region-based approach in content retrieval that uses Latent Semantic Analysis (LSA) techniques is presented in [15].The choice of global local visual features also appears crucial for good analysis results. In order to exploit the spatial content of a keyframe, the extraction of low-level concepts is performed after the image is modeled using grids, thus color and texture features are selected locally [2]. A similar approach is presented in [13]. Here, the features are extracted by regions of an image that resulted using a mean-shift algorithm. Finally in [18], a region-based approach is presented, that uses knowledge encoded in the form of an ontology. MPEG-7 visual features are extracted and combined and high-level concepts are detected.

In the context of TV news bulletins, a hybrid thesaurus approach is presented in [10]. There, semantic object recognition and identification for video news archives is achieved, with emphasis to face detection and TV channel logos. A lexicon-driven approach for an interactive video retrieval system is presented in [3].The core of this solution is the automatic detection of an unprecedented lexicon of 101 concepts. A lexicon design for semantic indexing in media databases is also presented in [1].

In this work, the problem of concept detection in video is approached in the following way: Each shot is represented by a keyframe, thus, the first step is keyframe extraction from shots. Then a clustering algorithm is applied on the RGB color values of every keyframe and splits the keyframes in homogeneous regions. The centroids of the clusters denote the color description of the image. For each cluster, a texture descriptor is extracted, and this way, the texture description of the image is formed. Fusing both low-level descriptions, a feature vector for each image is formed. Using a significantly large number of keyframes and applying a subtractive clustering method, we construct a region thesaurus, containing all the region types, which may or may not represent the concepts that are chosen to be detected. This thesaurus acts as the knowledge base and facilitates the association of low to high-level features. Each region type of the region thesaurus contains the appropriate merged color and texture description. By measuring the distances of the regions of an image to the region types, a model vector is formed that captures the semantics of an image. A support vector machine is trained to detect each high-level semantic concept, based on the values of the model vectors.

During the last years, there has been an effort to effectively evaluate and benchmark various approaches in the field of information retrieval, by the TREC conference series. Within this series the TRECVID [16] evaluation attracts many organizations and research interested in comparing their research in tasks such as automatic segmentation, indexing, and content-based retrieval of digital video. For its first years, the interest of TRECVID has been the TV news domain. Within the context of TRECVID, the presented work is applied on TV news bulletins from TRECVID 2005 development data, to

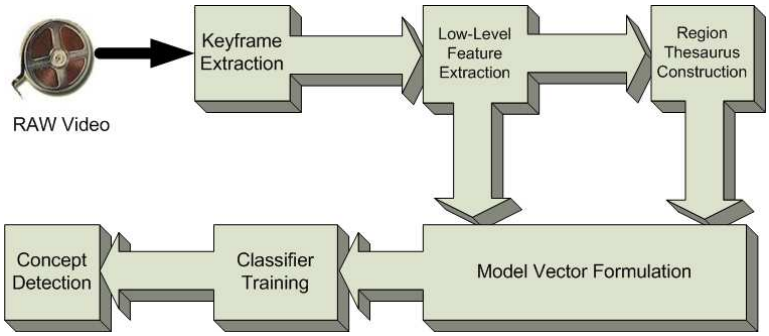

Figure 1. Presented Framework

detect specific high-level features: *desert*, *vegetation*, *mountain*, *road*, *sky* and *snow*. The presented framework is depicted in figure 1.

This paper is organized as follows: Section 1 presents the method used for the extraction of the color and texture features of a given keyframe. The method for the construction of the thesaurus containing all the region types derived from the training set is presented in section 2, followed by the construction of the model vectors that include the semantic image features in section 3. Then, section 4 presents the application of the Latent Semantic Analysis technique to the presented problem. Section 5 presents SVM-based high-level feature (concept) detectors, followed by experimental results in section 6. Finally, conclusions are drawn in section 7 accompanied by plans for future work.

1. Low-Level Feature Extraction

For the representation of the low-level features of a given keyframe, this work considers only color and texture features. For the color properties, a simple description similar to the approach of the MPEG-7 Dominant Color Descriptor and for the texture properties the actual MPEG-7 Homogeneous Texture Descriptor [8] have been applied, respectively.

1.1. Color Features

It is shown in many image retrieval applications that a set of dominant colors in an image or a region of interest is usually capable of efficiently capturing its color properties. The standardized MPEG-7 Dominant Color Descriptor [8] is formed after the clustering of the present colors within an image or a region of interest. This way, the representative colors of each keyframe are calculated. The selected low-level visual features of the image consist of the representative (dominant) colors, their percentages in the region, and optionally their spatial coherencies and their variances. What discriminates the use of the dominant color description of an image, instead of i.e. a color histogram is that the representative colors are computed each time, based on the features of the given image rather than being fixed in the color space.

Figure 2. An input image used for the extraction of four dominant colors

In our approach, the well-known K-means clustering method is applied on the RGB values of a given keyframe. As opposed to the MPEG-7 Dominant Color descriptor, where the number of the extracted representative colors varies from image to image allowing a maximum of eight colors that are allowed to be extracted from each image, a fixed number of colors is each time preselected in our approach. Using this predefined number, an image is then represented by the features of a fixed number of regions. Since the colors are clustered and the average color of each image region (cluster of the color space) is considered, our approach describes the color properties in a similar way to the Dominant Color Descriptor and there is no need to further extract more colors from its region, since the regions already occur from color clustering and share similar color properties.

The color description is then formed as follows:

$$DCD_N = [\{C_1, P_1\}, \{C_2, P_2\}, \ldots, \{C_N, P_N\}]$$

An example of an input image is depicted in figure 2. For the case of four dominant colors the four images each one containing one of the four regions are depicted in figure 3.

1.2. Texture Features

To efficiently capture the texture features of an image, the MPEG-7 Homogeneous Texture Descriptor (HTD) [8] is applied, since it provides a quintative characterization of texture and comprises a robust and easy to compute descriptor. The image is first filtered with orientation and scale sensitive filters. The mean and standard deviation of the fil-

Figure 3. The four extracted regions of image depicted in figure 2 by the K-means clustering described in section 1.1

tered outputs are computed in the frequency domain. The frequency space is divided in 30 channels, as described in [9], and the energy and energy deviation of each channel are computed and logarithmically scaled.

The texture description for a region of an image is then formed as follows:

$$HTD = [f_{DC}, f_{SD}, e_1, e_2, ..., e_{30}, d_1, d_2, ..., d_{30}]$$

Where f_{DC} and f_{SD} denote the mean and the standard deviation of the image texture respectively. Since each image is divided into four regions based on its color features, its texture properties are described by four homogeneous texture descriptors. The energy deviation of each channel is discarded, in order to simplify the description and moreover to prevent biasing towards the texture features, since they already have a significantly higher dimensionality than the color ones.

1.3. Fusion of color and texture features

All the low-level visual descriptions of a keyframe are merged into a unique vector. If DCD, HTD_1, HTD_2, HTD_3, $\dots HTD_N$ are respectively the N dominant (representative) colors and the homogeneous texture descriptors of each region of the keyframe referenced before, then the merged keyframe description D_{KF} is defined as:

$$D_{KF} = [DCD|HTD_1|HTD_2|HTD_3|\dots|HTD_4] \quad (1)$$

It is necessary that all color and texture features should have more or less the same numerical values to avoid scale effects. To achieve that, the dominant color and the ho-

mogeneous texture descriptions are normalized before their fusion into this unique vector which will be referred to as *feature vector*.

2. Region Thesaurus Construction

By observing the set of the keyframes extracted from the entire video collection and their low-level visual features extracted as described in section 1, becomes obvious that keyframes with similar semantic features should have similar low-level descriptions. Apart from that, there are many keyframes that share almost identical color and texture features since in our case they are taken in a studio with a still camera and have only a small time difference. Moreover, many shots within the same TV news bulletin contain only the anchorman and the same artificial background. To exploit this, clustering is performed on all the descriptions of the training set. Since we cannot have a priori knowledge for the exact number of the required classes, a K-means or a Fuzzy C-means clustering approach does not appear useful enough. To overrun this problem *Subtractive clustering* [4] is the method we choose to apply on the low-level description set. This method assumes each data point is a potential cluster center and calculates a measure of the likelihood that each data point would define the cluster center, based on the density of surrounding data points. In other words, this algorithm defines the number of the clusters and their corresponding centroids.

After the application of the clustering technique on the training data set, the following observations become obvious: First, each cluster may or may not represent a high-level feature. There are some clusters that contain region types belonging to the same high-level concept. Apart from those, most of the clusters contain region types that do not belong to the same high-level feature and are mixed up because they share similar color and texture features. Second, some concepts can be found in more than one clusters, since they cannot be described uniquely by their visual characteristics. For example, the concept *desert* can have more than one instances differing in i.e. the color of the sand, each represented by the centroid of a cluster. Moreover, in a cluster that may contain instances from the semantic entity i.e. *sea*, these instances could be mixed up with parts from i.e. *sky*, if present in the data set.

Generally, a *thesaurus* combines a list of every term in a given domain of knowledge and a set of related terms for each term in the list. In our approach, the constructed Region Thesaurus contains all the Region Types that are encountered in the training set. These region types are the centroids of the clusters and all the other feature vectors of a cluster are their synonyms. It is important to mention that when two region types are considered to be synonyms, they belong to same cluster, thus share similar visual features, but do not necessarily share the same semantics. By using a significantly large training set of keyframes, our thesaurus is constructed and enriched. As it will be presented in section 3, the use of the thesaurus is to provide a means of association of the low-level features of the image with the high-level concepts.

Since the number of the region types can be very large depending on the selected thresholds for the potential above or below from which a data point will be selected or rejected respectively as a cluster center, the dimensionality of the model vector may become very high. It is then possible that the extracted region types may carry redundant information. However, since two region types may sometimes be strongly correlated

although they may appear visually different. To avoid this, principal component analysis (PCA) is applied in order to reduce the dimensionality and facilitate the performance of the high-level feature detectors which are presented in section 5.

3. Model Vectors for keyframe representation

After the construction of the region thesaurus, a model vector is formed for each keyframe. Its dimensionality is equal to the number of concepts that constitute the thesaurus. The distance of a region to a region type is calculated as a linear combination of the dominant color and homogeneous texture distances respectively. The MPEG-7 standardized distances are used for each case and a linear combination is used to fuse the distances as in [6].

3.1. Similarity Measures

We begin with the similarity measure used for the case of the color descriptor. Since the color representation is rather simple, the well known Euclidean distance is used since it works effectively in many applications and retrieval systems.

$$D(F_1, F_2) = \sqrt{(R_1 - R_2)^2 + (G_1 - G_2)^2 + (B_1 - B_2)^2} \tag{2}$$

where $F_i = (R_i, G_i, B_i), i = 1, 2$ are the two RGB values. The distance between 2 *Homogeneous Texture Descriptors* is computed as:

$$D(HTD_1, HTD_2) = \sum_k \left| \frac{HTD_1(k) - HTD_2(k)}{a(k)} \right| \tag{3}$$

where $a(k)$ is the standard deviation of the Homogeneous Texture Descriptors for a given database. In this approach the standard deviation of the database is ignored, because it does not affect the result since the values are normalized afterwards.

However we should notice that the MPEG-7 standard does not strictly define the distance functions to be used, thus leaving the developers the flexibility to develop their own dissimilarity/distance functions and to exploit other well-known similarity functions such as i.e. the Minkowski distance.

3.2. Model Vector Formulation

Having calculated the distance of each region (cluster) of the image to all the words of the constructed thesaurus, the model vector that semantically describes the visual content of the image is formed by keeping the smaller distance for each high-level concept. More specifically, let: $d_i^1, d_i^2, ..., d_i^j, i = 1, 2, 3, 4$ and $j = N_C$, where N_C denotes the number of words of the lexicon and d_i^j is the distance of the i-th region of the clustered image to the j-th region type. Then, the model vector D_m is the one depicted in equation 4.

$$D_m = [min\{d_i^1\}, min\{d_i^2\}, ..., min\{d_i^{N_C}\}], i = 1, 2, 3, 4 \tag{4}$$

4. Latent Semantic Analysis

Apart from the obvious next step of simply training classifiers using the aforementioned model vectors as the means of representing the extracted features of the given keyframe, we also perform some experiments using a Latent Semantic Analysis [5](LSA) approach. LSA is a technique in natural language processing, which exploits the relationships between a set of documents and the terms they contain more often by producing a set of concepts related to the documents and terms. In our approach, since a keyframe is described as a set of region types, it appears obvious that LSA can easily be applied with the keyframe and the region types corresponding to a document and the terms it contains, respectively.

We first construct the co-occurrence matrix of region types in given keyframes of the training set in contexts (region types in the thesaurus). The distance function we use to compare a given region type with one of the thesaurus, in order to assign each region of the image to the correct prototype region is a linear combination of a Euclidean distance for the dominant color and the MPEG-7 standardized distance for the HTD. After of the construction of the co-occurrence matrix, we solve the SVD problem and transform all the model vectors to the semantic space. For each semantic concept, a separate SVM is then trained having as input the model vector in the semantic space.

5. SVM Feature Detector Training

Support Vector Machines [17] are feed-forward networks that can be used for pattern classification and nonlinear regression. Their main idea is to construct a hyperplane that acts as a decision space in such a way that the margin of separation between positive and negative examples is maximized. This hyperplane is not constructed in the input space, where the problem may not be linearly solvable, but in the feature space where the problem is driven. This is generally referred as the Optimal Hyperplane, a property that is achieved as the support vector machines are an approximate implementation of the method of structural risk minimization. Despite the fact that a support vector machine does not incorporate domain-specific knowledge, it provides a good generalization performance, a unique property among the various different types of neural networks. Support vector machines have been used for image classification based on their histogram as in [12] and for the detection of semantic concepts such as goal, yellow card and substitution in the soccer domain [14].

An inner-product kernel between an input vector $\mathbf{x}$ and a *support vector* $\mathbf{x}_i$ is the main characteristic on the support vector machines. The support vectors consist of a small subset of the training set vectors and are extracted by the optimization algorithm. The kernel can be implemented in various ways, thus leading to different types of nonlinear learning machines. The most important are *Polynomial* learning machines, *Radial-Basis Function* networks and Single-hidden layer *Perceptrons*, where the kernel function is polynomial, exponential or a hyperbolic tangent function, respectively.

The nonlinear mapping may be denoted by a set of nonlinear transformations as $\{\phi_j(\mathbf{x})\}_{j=1}^{m_1}$. Then, a hyperplane in the feature space is defined as:

Table 1. Classification rate using both visual descriptors for various numbers of the region types

Concept	35 Region Types	62 Region Types	125 Region Types
Desert	82.5%	77.5%	70.1%
Vegetation	80.5%	71.3%	67.2%
Mountain	83.6%	77.7%	67.0%
Road	72.0%	67.0%	65.9%
Sky	80.1%	77.4%	70.0%
Snow	70.5 %	62.1%	55.2%

$$\sum_{j=1}^{m_1} w_j \phi_j(\mathbf{x}) + b = 0 \tag{5}$$

If we denote the inner-product kernel of the support vector machine as $K(\mathbf{x}, \mathbf{x}_i)$, it is defined by:

$$K(\mathbf{x}, \mathbf{x}_i) = \boldsymbol{\phi}^T(\mathbf{x})\boldsymbol{\phi}(\mathbf{x}_i) \tag{6}$$

For each semantic concept, a separate support vector machine is trained, thus solving a binary problem, of the existence or not of the concept in question. The input of the SVM is the model vector D_m described in section 3. The well known polynomial support vector machine, described in equation 7 is selected in our framework.

$$K(\mathbf{x}, \mathbf{x}_i) = (\mathbf{x}^T\mathbf{x}_i + 1)^p \tag{7}$$

6. Experimental Results

For the evaluation of the presented framework, part of the development data of TRECVID 2005 were used. This set consists of approximately 65000 keyframes, captured from TV news bulletins. For the following experiments, a set of 5000 keyframes was selected in order to include examples of all the selected features and also some keyframes not containing any of them. The high-level features for which feature detectors were implemented are: *desert, vegetation, mountain, road, sky and snow*. The annotation was provided by the LSCOM Lexicon Definitions and Annotations [11]. The color visual features were extracted using a standard K-means clustering algorithm and the texture visual features using the MPEG-7 eXperimentation Model (XM).

Experiments were performed on the size of the region thesaurus, the number of dominant colors and the presence or not of both visual descriptors. Results are shown in table 1 for different sizes of the region thesaurus, in table 2 for fixed size of the region thesaurus and the use of different numbers of dominant colors and finally in table 3 for each descriptor and their combination. Also, experiments were performed for the case of fixed size of the region thesaurus and for each one or both of the descriptors, using LSA as an intermediate step between model vector formulation and SVM training and classification. The performance of these experiments is presented in table 4

Table 2. Classification rate using both visual descriptors for various numbers of the dominant colors, thesaurus size = 35

Concept	2 DC + HT	3 DC + HT	4 DC + HT	5 DC + HT
Desert	77.5%	80.5%	82.5%	79.0%
Vegetation	70.5%	77.5%	80.5%	81.2%
Mountain	70.3%	82.0%	83.6%	78.6%
Road	68.0%	70.0%	72.0%	70.0%
Sky	77.5%	80.1%	80.1%	79.0%
Snow	57.2%	62.0%	70.5%	72.2%

Table 3. Classification rate using only color, only texture and both visual descriptors, thesaurus size = 35

Concept	DC	HT	DC+HT
Desert	80.2%	77.2%	82.5%
Vegetation	72.5%	75.0%	80.5%
Mountain	72.1%	77.5%	83.6%
Road	71.5%	70.2%	72.0%
Sky	85.0%	70.1%	80.1%
Snow	75.0%	60.1%	70.5%

Table 4. Classification rate using only color, only texture and both visual descriptors and LSA in all cases, thesaurus size = 35

Concept	DC	HT	DC+HT
Desert	83.2%	75.2%	87.2%
Vegetation	74.5%	75.2%	82.5%
Mountain	77.5%	77.5%	80.6%
Road	78.2%	73.7%	76.7%
Sky	88.2%	72.5%	82.2%
Snow	79.0%	65.0%	72.5%

7. Conclusions - Future Work

The experimental results indicate that the extraction of the aforementioned low-level features is appropriate for semantic indexing. The selected concepts can be successfully detected when a keyframe is represented by a model vector that contains the distances to all the semantic entities of a constructed lexicon containing unlabeled semantic features. Latent Semantic Analysis was also successfully applied in the given problem and led to an improvement of the results. Plans for future work include the extraction of more visual features, exploitation of the spatial context of a keyframe and extension of this method for applications such as shot/image classification. Finally, integration of the presented framework to the one of [18] and fusion of their results is also intended.

8. Acknowledgements

The work presented in this paper was partially supported by the European Commission under contracts FP6-027026 K-Space and FP6-027685 MESH. Evaggelos Spyrou is funded by PENED 2003 Project Ontomedia 03ED475.

References

[1] M. Naphade A. Natsev and J. Smith. Lexicon design for semantic indexing in media databases. In *International Conference on Communication Technologies and Programming*, 2003.

[2] S. Aksoy, A. Avci, E. Balcuk, O. Cavus, P. Duygulu, Z. Karaman, P. Kavak, C. Kaynak, E. Kucukayvaz, C. Ocalan, and P. Yildiz. Bilkent university at trecvid 2005. 2005.

[3] Dennis C. Koelma Cees G.M. Snoek, Marcel Worring and Arnold W.M. Smeulders. Learned lexicon-driven interactive video retrieval. 2006.

[4] S.L. Chiu. *Extracting Fuzzy Rules from Data for Function Approximation and Pattern Classification*. John Wiley and Sons, 1997.

[5] S. Deerwester, Susan Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the Society for Information Science*, 41(6):391–407, 1990.

[6] E.Spyrou, H.LeBorgne, T.Mailis, E.Cooke, Y.Avrithis, and N.O'Connor. Fusing mpeg-7 visual descriptors for image classification. In *International Conference on Artificial Neural Networks (ICANN)*, 2005.

[7] IBM. Marvel: Multimedia analysis and retrieval system.

[8] B.S. Manjunath, J.R. Ohm, V.V. Vasudevan, and A. Yamada. Color and texture descriptors. *IEEE trans. on Circuits and Systems for Video Technology*, 11(6):703–715, 2001.

[9] MPEG-7. Visual experimentation model (xm) version 10.0. ISO/IEC/ JTC1/SC29/WG11, Doc. N4062, 2001.

[10] V. Gouet N. Boujemaa, F. Fleuret and H. Sahbi. Visual content extraction for automatic semantic annotation of video news. In *IS&T/SPIE Conference on Storage and Retrieval Methods and Applications for Multimedia, part of Electronic Imaging symposium*, January 2004.

[11] M. R. Naphade, L. Kennedy, J. R. Kender, S.-F. Chang, J. R. Smith, P. Over, and A. Hauptmann. A light scale concept ontology for multimedia understanding for trecvid 2005. IBM Research Technical Report, 2005.

[12] O.Chapelle, P.Haffner, and V.N.Vapnik. Support vector machines for histogram-based image classification. *IEEE Transactions on Neural Networks*, 10(5):1055–1064, 1999.

[13] B.Le Saux and G.Amato. Image classifiers for scene analysis. In *International Conference on Computer Vision and Graphics*, 2004.

[14] C. G. M. Snoek and M. Worring. Time interval based modelling and classification of events in soccer video. In *Proceedings of the 9th Annual Conference of the advanced School for Computing and Imaging (ASCI)*, 2003.

[15] F. Souvannavong, B. Mérialdo, and B. Huet. Region-based video content indexing and retrieval. In *CBMI 2005, Fourth International Workshop on Content-Based Multimedia Indexing, June 21-23, 2005, Riga, Latvia*, Jun 2005.

[16] TREC. - video retrieval evaluation. http://www-nlpir.nist.gov/projects/t01v/.

[17] V. Vapnik. *Statistical Learning Theory*. John Wiley and Sons, 1998.

[18] N. Voisine, S. Dasiopoulou, V. Mezaris, E. Spyrou, Th. Athanasiadis, I. Kompatsiaris, Y. Avrithis, and M. G. Strintzis. Knowledge-assisted video analysis using a genetic algorithm. In *6th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS 2005)*, April 13-15, 2005.