Using Local Region Semantics for Concept Detection in Video

Evaggelos Spyrou, George Koumoulos, Yannis Avrithis and Stefanos Kollias

***Abstract*— This paper presents a framework for the detection of semantic features in video sequences. Low-level feature extraction is performed on the keyframes of the shots 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.Then, a "model vector" that contains the distances from each region type is formed and a SVM detector is trained for each semantic concept. Experiments were performed using TRECVID 2005 development data.**

***Index Terms*— semantic analysis, thesaurus, SVM, TRECVID**

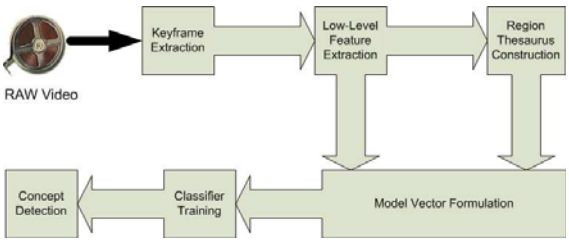

Fig. 1. Presented Framework

I. Introduction

HIGH-level concept detection in video documents remains still an unsolved problem. One aspect of this is the extraction of the low-level features of a video sequence and the other is the method is used for assigning low-level descriptions to high-level concepts, a problem commonly referred to as the "Semantic Gap". Many approaches have been proposed, all sharing the target of bridging the semantic gap, thus extracting high-level concepts from multimedia documents.

In [5], a prototype multimedia analysis and retrieval system is presented, that uses multi-modal machine learning techniques in order to model semantic concepts in video. A region-based approach in content retrieval that uses Latent Semantic Analysis (LSI) is presented in [9]. The extraction of low-level concepts is performed after the image is clustered by a mean shift algorithm thus features are selected locally in [8]. In [11], a region-based approach using MPEG-7 visual features and knowledge in the form of an ontology is presented. Moreover, in the context of TV news bulletins, a hybrid thesaurus approach is presented in [7], a lexicon-driven approach for an interactive video retrieval system is presented in [2] and 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: Low-level features are extracted from keyframes, each representing a shot. A model vector is formed by associating these descriptions with the words of a thesaurus. Then a SVM classifier is used to detect the semantic concepts.The presented framework is depicted in figure 1.

E.Spyrou, G.Koumoulos, Y.Avrithis and S.Kollias are with 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)

II. Low-Level Feature Extraction

Since a set of dominant colors in an image or a region of interest has the ability to efficiently capture its color properties, an approach based on the MPEG-7 Dominant Color Descriptor [6] was selected. The 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 allowing a maximum of eight colors that can be extracted, a fixed number of colors is each time preselected in our approach.the MPEG-7 Homogeneous Texture Descriptor (HTD) [6] was used to capture texture properties of each region. The energy deviations of the descriptors were discarded, in order to simplify the description, preventing biasing towards the texture features.

All the low-level visual descriptions of a keyframe are normalized to avoid scale effects and merged into a unique vector. This vector will be referred to as *feature vector*.

III. Region Thesaurus Construction

Given the entire set of the keyframes extracted from a video, it is obvious that those with similar semantic features should have similar low-level descriptions. 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, *Subtractive Clustering* [3] is the applied method on the low-level description set, since it determines the number of the clusters. Each cluster may or may not represent a high-level feature and each high-level feature may be represented by one or more clusters. For example, the concept *desert* can have many instances differing in i.e. the color of the sand. Moreover, in a cluster that may contain instances from the semantic entity i.e. *sea*, these instances could be mixed up with parts from another concept i.e. *sky*, if present in an image.

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%

TABLE I

CLASSIFICATION RATE USING BOTH VISUAL DESCRIPTORS FOR VARIOUS NUMBERS OF THE REGION TYPES

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 II

CLASSIFICATION RATE USING BOTH VISUAL DESCRIPTORS FOR VARIOUS NUMBERS OF THE DOMINANT COLORS, THESAURUS SIZE = 35

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 members of the cluster are their synonyms. The use of the thesaurus is to facilitate the association of the low-level features of the image with the corresponding high-level concepts. Since the number of the region types can be very large, the dimensionality of the model vector may become very high. To avoid this, principal component analysis (PCA) is applied in order to reduce its dimensionality, thus facilitating the performance of the feature detectors.

IV. MODEL VECTOR KEYFRAME DESCRIPTION

After the construction of the region thesaurus, a "model vector" is formed for each keyframe. Its dimensionality is equal to the number of concepts constituting 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, as in [4].

Having calculated the distance of each region of the image to all the region types 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.

For each semantic concept , a support vector machine [10] is trained. Its input is the model vector and its output determines whether the concept exists or not within the keyframe.

V. EXPERIMENTAL RESULTS

For the evaluation of the presented framework,part of the development data of TRECVID 2005 was used. This set consists of approximately 65000 keyframes, captured from TV news bulletins. The high-level features for which feature detectors were implemented are: *desert, vegetation, mountain, road, sky and snow*. 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 tables I, II and III.

VI. CONCLUSION

The experimental results indicate that the selected concepts can be detected when a keyframe is represented by a model vector with the use of a visual thesaurus. Moreover, future plans include integration of the presented framework to the one of [11] and fusion of their results.

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 III

CLASSIFICATION RATE USING ONLY COLOR, ONLY TEXTURE AND BOTH VISUAL DESCRIPTORS, THESAURUS SIZE = 35

ACKNOWLEDGMENT

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 the Greek Secretariat of Research and Technology (PENED Ontomedia 03 ED 475)

REFERENCES

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

[2] D. C. K. Cees G.M. Snoek, Marcel Worring and A. W. Smeulders, "Learned lexicon-driven interactive video retrieval," 2006.

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

[4] 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.

[5] IBM, "Marvel: Multimedia analysis and retrieval system." [Online]. Available: http://mp7.watson.ibm.com/

[6] B. Manjunath, J. Ohm, V. Vasudevan, and A. Yamada, "Color and texture descriptors," *IEEE trans. on Circuits and Systems for Video Technology*, vol. 11, no. 6, pp. 703–715, 2001.

[7] V. G. 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.

[8] B. Saux and G.Amato, "Image classifiers for scene analysis," in *International Conference on Computer Vision and Graphics*, 2004.

[9] 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.

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

[11] N. Voisine, S. Dasiopoulou, V. Mezaris, E. Spyrou, T. 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.