Large Scale Concept Detection in Video Using a Region Thesaurus

Evaggelos Spyrou, Giorgos Tolias, and Yannis Avrithis

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,
espyrou@image.ece.ntua.gr,
WWW home page: http://www.image.ece.ntua.gr/∼espyrou/

Abstract. This paper presents an approach on high-level feature detection within video documents, using a Region Thesaurus. A video shot is represented by a single keyframe and MPEG-7 features are extracted locally, from coarse segmented regions. Then a clustering algorithm is applied on those extracted regions and a region thesaurus is constructed to facilitate the description of each keyframe at a higher level than the low-level descriptors but at a lower than the high-level concepts. A model vector representation is formed and several high-level concept detectors are appropriately trained using a global keyframe annotation. The proposed approach is thoroughly evaluated on the TRECVID 2007 development data for the detection of nine high level concepts, demonstrating sufficient performance on large data sets.

1 Introduction

One of the most interesting problems in multimedia content analysis remains the detection of high-level concepts within multimedia documents. Due to the continuously growing volume of audiovisual content, this problem attracts a lot of interest within the multimedia research community. Many research efforts set focus on the extraction of various low-level features, such as audio, color, texture and shape properties of audiovisual content. Moreover, many techniques such as neural networks, fuzzy systems and Support Vector Machines (SVM) have been successfully applied in order to link the aforementioned features to high-level features. However, the well-known "semantic gap" often characterizes the differences between descriptions of a multimedia object by different representations and the linking from the low- to the high-level features.

An important step for narrowing this gap is to provide a description based on higher-level properties than the low-level descriptors. Many research efforts make use of a visual dictionary to describe a decomposed image derived after either clustering or segmentation or keypoint extraction. A mean-shift algorithm is used in [1], to cluster an image and extract local features. In [2] image regions are clustered and a codebook of region types occurs. Moreover, in [3] visual categorization is achieved using a bag-of-keypoints approach. Then, a bag-of-regions approach is used for scene detection.

A hybrid thesaurus approach for object recognition within video news archives is presented in [4], while in [5] a a region-based approach that uses knowledge encoded in the form of an ontology is applied. Finally, a multi-modal machine learning technique is used in [6].

However, the growth of multimedia content has not been accompanied by a similar growth of the available annotated data sets. Very few are the databases that provide an annotation per region, such as LabelMe [7]. On the other hand, annotating an image globally, appears a much easier task. Such an annotation is provided from LSCOM workshop [8]. Therein, a very large number of shots of news bulletins are globally annotated for a large number of concepts. Moreover, during the last few years, TRECVID [9] evaluation continues to attract many researchers interested in comparing their work in various tasks and among them the high-level feature detection within video documents. Within this task, the goal is to globally annotate shots of video for certain concepts.

This work falls within the scope of TRECVID and tackles 9 concepts within the 2007 development data using a common detection approach for all concepts and not specialized algorithms. The concepts that have been selected are *Vegetation*, *Road*, *Explosion_fire*, *Sky*, *Snow*, *Office*, *Desert*, *Outdoor* and *Mountain* and as obvious, they cannot be described as "objects", but rather as "materials" or "scenes". These concepts have been selected since they are the only materials from the TRECVID's set of concepts. Thus, color and texture features are the only applicable MPEG-7 low-level features. For each concept a neural network-based detector is trained based on features extracted from keyframe regions, while keyframes are annotated globally. The presented framework is depicted in figure 1, where the off-line steps, i.e. those that comprise the training part are marked as yellow.

This paper is organized as follows: Section 2 presents the method used for extracting color and texture features of a given keyframe. The construction of the region thesaurus is presented in section 3, followed by the formulation of the model vectors used to describe a keyframe in section 4. Then, training of the neural-network detectors is presented in section 5. Extensive experimental results data are presented in section 6 and finally, conclusions are drawn in section 7.

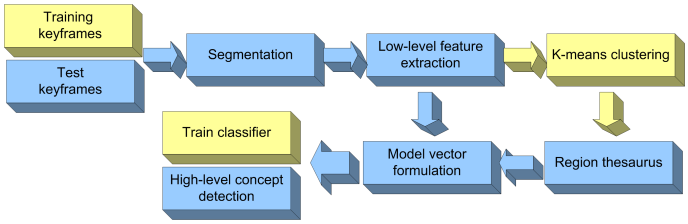


Fig. 1. High-level concept detection framework.

2 Low-Level Feature Extraction

At a preprocessing step, a video document, is first segmented into shots and from each shot a representative frame (keyframe) is extracted. Each keyframe k_i is then segmented into regions, using a (color) RSST segmentation algorithm [10], tuned to produce an under-segmented image. Let R denote the set of all regions and $R(k_i) \subset R$ the set of all regions of the keyframe k_i.

Several MPEG-7 descriptors [11] have been selected to capture the low-level features of each region $r_i \in R$. More specifically, *Dominant Color Descriptor* (DC), *Color Structure Descriptor* (CS), *Color Layout Descriptor* (CL) and *Scalable Color Descriptor* (SC) are extracted to capture the color properties and *Homogeneous Texture Descriptor* (HT) and *Edge Histogram Descriptor* (EH) the texture properties.

To obtain a single region description from all the extracted region descriptions, we choose to follow an "early fusion" method, thus merging them after their extraction [12]. The vector formed will be referred to as "feature vector". The feature vector that corresponds to a region $r_i \in R$ is thus depicted in equation (1):

$$f(r_i) = [CL_i, DC_i, CS_i, SC_i, HT_i, EH_i] \tag{1}$$

where $DC(r_i)$ is the Dominant Color Descriptor for region r_i, $CL(r_i)$ is the Color Layout Descriptor for region r_i etc. Each feature vector is denoted by f_i and F is the set of all feature vectors. In other words: $f_i \in F, \quad i = 1 \ldots N_F = N_R$. Herein, we should note that regarding the Dominant Color Descriptor, we choose to keep only the most dominant color and its percentage, since the length of the full descriptor is generally not fixed.

Since many of the MPEG-7 descriptors allow the user to select their level of detail, thus offer a large number of available extraction profiles, we follow a procedure similar to the one presented in [13], in order to select the one that best suits the needs of our approach. The dimensions of the extracted descriptors are depicted in table 1, while the final dimension of the merged feature vector is 286.

Descriptor	DC	SC	CL	CS	EH	HT
Number of Coefficients	4	64	12	64	80	62

Table 1. Dimension of the extracted MPEG-7 color and texture descriptors.

3 Region Thesaurus

Given the entire training set of images and their extracted low-level features, it may be easily observed that regions belonging to similar semantic concepts, have similar low-level descriptions. Also, images containing the same high-level concepts are consisted of similar regions. For example, all regions that belong to the semantic concept *Sky* should be visually similar, i.e. the color of most of them should be some tone of blue.

Moreover, images that contain *Sky*, often contain some similar regions. Finally, and in large problems, such as TRECVID discussed herein, common keyframe extraction algorithms sometimes extract visually similar keyframes, that belong to neighboring shots within the same video.

The aforementioned observations indicate that certain similar regions often co-exist with some high-level concepts. In other words, region co-existences should be able to characterize the concepts that exist within a keyframe. Thus, initially, regions derived from keyframes of the training set are organized in a structure, able to facilitate the description of a given keyframe with respect to a subset of them.

A K-means clustering algorithm is first applied on the feature vectors of the regions of the training set images. The number of clusters N_T is selected experimentally. The definition of the region thesaurus is the one depicted in Eq. (2).

$$T = \Big\{ w_i, \quad i = 1 \dots N_T \Big\}, \quad w_i \subset R \tag{2}$$

where w_i is the i-th cluster, which is a set of regions that belong to R. Then, from each cluster, the region that lies closest to its centroid is selected. These regions will be referred to as "region types" and their corresponding feature vector is depicted in Eq. (3), where $z(w_i)$ is the centroid of the i-th cluster in the feature space.

$$f(w_i) = f\Big(\arg\min_{r \in w_i} \Big\{ d\big(f(r), z(w_i)\big) \Big\} \Big) \tag{3}$$

A region type does not contain conceptual semantic information, although appears a higher description than a low-level descriptor; i.e. one could intuitively describe a region type as "green region with a coarse texture", but would not be necessarily able to link it to a specific concept, which neither is necessary a straightforward process, nor falls within the scope of the presented approach.

In order to create the region thesaurus, a large number of regions is obviously necessary, in order for it to be able to describe effectively every image. When the training set is significantly small, then all available regions are used. On the other side, when the training set is significantly large, i.e. in the TRECVID case, regions derived from images containing the semantic concepts to be detected are selected, accompanied by an equal number of randomly selected regions among the remaining images.

4 Model Vectors

After forming the region thesaurus, described in section 3, a model vector is formed in order to represent the semantics of a keyframe, based on the set of region types. Let r_1, r_2, be two image regions, described by feature vectors f_1, f_2, respectively. The Euclidean distance is applied in order to calculate their distance $d(f_1, f_2)$.

Having calculated the distance between each region of the image and all region types, the "model vector" is then formed by keeping the smallest distance of all image regions to each region type. More specifically, the model vector m_i describing keyframe k_i is the one depicted in eq. (4).

$$m_i = \Big[m_i(1), m_i(2), \dots, m_i(j), \dots, m_i(N_T) \Big] \tag{4}$$

where

$$m_i(j) = \min_{r \in R(k_i)} \left\{ d\Big(f(w_j), f(r) \Big) \right\} \tag{5}$$

and $j = 1 \ldots N_T$.

In fig.2, an under-segmented image and a region thesaurus consisted of 6 regions are depicted. The distances between the upper image region (the one corresponding to *Sky*) and every region type are shown on the left, while those of every image region to a specific region type are shown on the right. As obvious, for the 5th region type the corresponding value of the model vector will be 0.1 as the minimum distance among this region type and all other regions of the given image.

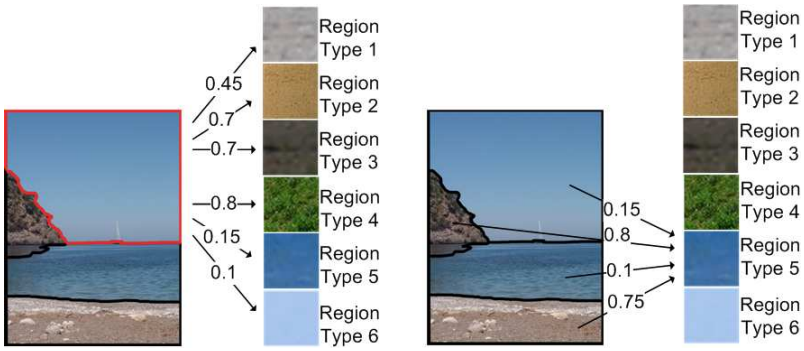


Fig. 2. Distances between regions and region types.

5 High-Level Concept Detectors

After extracting model vectors from all images of the (annotated) training set, a neural network-based detector is trained separately for each high-level concept. The input of the detectors is a model vector m_i describing a keyframe in terms of the region thesaurus. The output of the network is the confidence that the keyframe contains the specific concept. It is important to clarify that the detectors are trained based on annotation per image and not per region. The same stands for their output, thus they provide the confidence that the specific concept exists somewhere within the keyframe in question. Several experiments, presented in 6 indicate that the threshold above which it is decided that a concept exists, also varies depending on the classifier and should be determined experimentally, in a separate process for each concept.

6 Experimental Results

This section presents the results of the aforementioned algorithm, applied on the TRECVID 2007 Development Data, a large dataset consisting of 110 videos, segmented into shots.

A keyframe has been extracted from each shot, thus 18113 keyframes have been made available. The annotation used results from a joint effort among several TRECVID participants [14]. Table 2 summarizes the detected concepts and the number of positive examples within the development data and the constructed training/testing sets for each of them. After the under-segmentation, 345994 regions have been available.

Using the constructed training set, due to the large number of regions derived after segmentation, not all available regions are used. Rather, an adequate number of regions derived from the training sets of all high-level concepts, in other words from keyframes that contain at least one of the high-level concepts, and an equal number of random regions derived from keyframes that do not contain any of the high-level concepts, are used to form the region thesaurus. K-means clustering is applied, with N_T set to 100 region types.

concept	development data	training	testing
Desert	52	36	16
Road	923	646	277
Sky	2146	1502	644
Snow	112	78	34
Vegetation	1939	1357	582
Office	1419	993	426
Outdoor	5185	3000	1556
Explosion_Fire	29	20	9
Mountain	97	68	29

Table 2. Number of positive examples within the development data and the constructed training/testing sets.

First of all, several experiments are performed by varying the ratio λ_{tr} of negative to positive examples within the training set. For a complex problem, such as TRECVID, it is very difficult to "model" positive examples of each concept. Of course, modeling the negative examples appears more difficult. Thus, a larger number of negative examples is needed, but should be selected appropriately, in order to avoid biasing the detectors towards the negative examples. To test the performance of the differently trained classifiers, a testing set with a ratio $\lambda_{ts} = 1$, i.e. consisting of an equal number of positive and negative values is used. Results are summarized in table 3. It may be easily observed that for almost every concept, a value of λ_{tr} between 4-5 is appropriate to achieve the highest possible average precision (AP) [15] by the classifiers. We should note that the number of the available examples is inadequate in order to investigate more values of λ for most of the concepts. Thus we choose to stop at the value of $\lambda_{tr} = 5$, in order to have comparable results for most of the examined high-level concepts.

Having selected the training set, experiments on the threshold confidence value for each classifier are performed. As testing set for each concept, the set of all remaining keyframes is used. Precision and recall measures are calculated for each high-level concept and for a range of threshold values, starting from 0 and increasing with a step of 0.1 until they reach 0.9. Then, the threshold value where precision is almost equal to

concept	$\lambda_{tr}=1$	$\lambda_{tr}=2$	$\lambda_{tr}=3$	$\lambda_{tr}=4$	$\lambda_{tr}=5$
Desert	0.6593	**0.6994**	0.3653	0.4775	0.6634
Road	0.5944	0.6091	0.5954	0.6062	**0.6957**
Sky	0.6791	0.723	0.6883	0.7197	**0.7369**
Snow	0.9144	0.9054	0.9293	0.9174	**0.9504**
Vegetation	0.7175	0.7731	0.7649	0.7522	**0.7802**
Office	0.6337	0.7073	**0.7382**	0.7077	0.7235
Outdoor	0.6832	0.6842	**0.6978**	-	-
Expl._Fire	0.3879	0.3679	0.3485	**0.647**	0.3827
Mountain	0.6878	0.6119	0.5458	0.625	**0.7662**

Table 3. Average Precision on a test set with $\lambda_{ts}=1$, for several values of the ratio λ_{tr} within the training set.

recall is selected, as depicted in fig. 3. This way, both measures are kept in equally good values, as it is generally desirable. Table 4 summarizes the selected threshold values for all 9 concepts. As it may be observed, for those concepts that their positive examples do not vary a lot, in respect to their model vectors, such as *Desert* and *Mountain*, a high threshold value is selected.

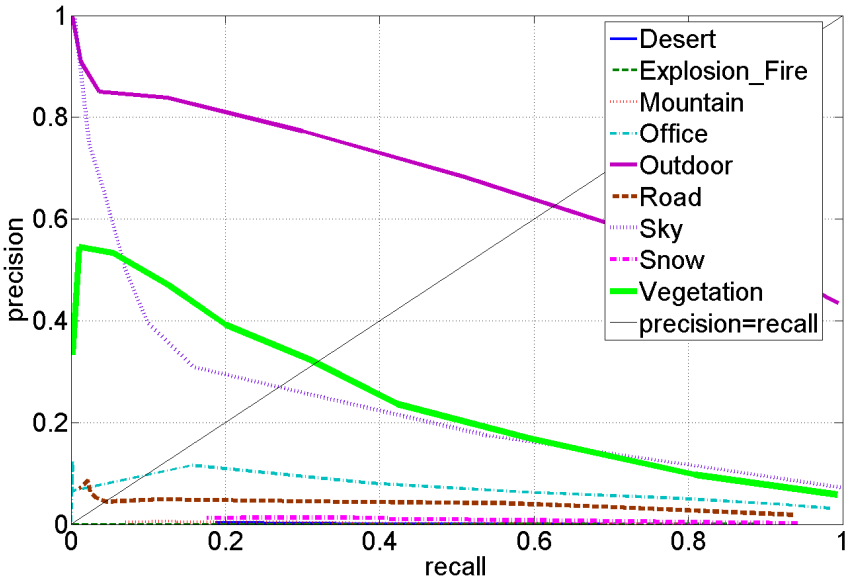


Fig. 3. Precision-Recall for increasing threshold values

In the last part of the experiments, the proposed approach is evaluated on the testing sets derived from the TRECVID 2007 development data. The testing set of each concept contains 30% of all positive examples and is complemented using part from negative

concept	threshold
Desert	0.8
Road	0.5
Sky	0.3
Snow	0.6
Vegetation	0.4
Office	0.5
Outdoor	0.3
Explosion_Fire	0.2
Mountain	0.8

Table 4. Thresholds for the high-level concept detectors.

examples, i.e. from all keyframes that do not contain the specific concept. The number of negative keyframes increases gradually, until it reaches certain values of λ_{ts}. For each concept, the value of λ_{ts} is increased until it reaches its maximum possible value. Each time the AP is calculated, with a window equal to all the testing set.

concept	$\lambda_{ts} = 4$			$\lambda_{ts} = max$		
	P	**R**	**AP**	**P**	**R**	**AP**
Vegetation	0.643	0.312	0.560	0.322	0.313	0.232
Road	0.295	0.046	0.407	0.045	0.047	0.043
Explosion_Fire	0.291	0.777	0.252	0.000	0.000	0.001
Sky	0.571	0.304	0.603	0.258	0.304	0.214
Snow	0.777	0.411	0.610	0.013	0.412	0.008
Office	0.446	0.157	0.418	0.117	0.157	0.072
Desert	0.333	0.312	0.457	0.003	0.313	0.064
Outdoor	0.425	0.514	0.361	0.425	0.514	0.361
Mountain	0.444	0.137	0.401	0.003	0.379	0.037

Table 5. Precision (P), Recall (R) and Average Precision (AP) for all concepts.

Figures 4 and 5 show how AP changes with respect to λ_{ts}. The number of positive examples is kept fixed, while the number of negative increases. It may be observed that when the value of λ_{ts} is relatively small, i.e. $\lambda_{ts} = 4$, as in the case of typical test sets that are used for evaluation, the performances remain particularly high. When λ_{ts} increases, then the performances fall as expected.

Finally some comments regarding the detection results are presented, focusing on examples of true and false detections of high-level concepts. Examples are depicted for concepts *Sky* and *Vegetation* in figures 6, 7 and 8 for true positive, false negative and false positive examples, respectively. Typical regions of *Sky* and *Vegetation* detected successfully are depicted in figure 6. Figure 7 presents false negative keyframes, containing images in which the existing small regions of *Vegetation* and *Sky* are also merged with other regions as a result of the under-segmentation of the image. Thus visual features of the regions are degraded. Our approach was unable to detect an artificial region of sky, such as the one depicted in figure 7(a), because of the abnormal yellowish tone. Some false positive examples are depicted in figure 8. Images falsely detected as positive for the concept *Sky* (figure 8(a)) contain light blue regions which are similar to a

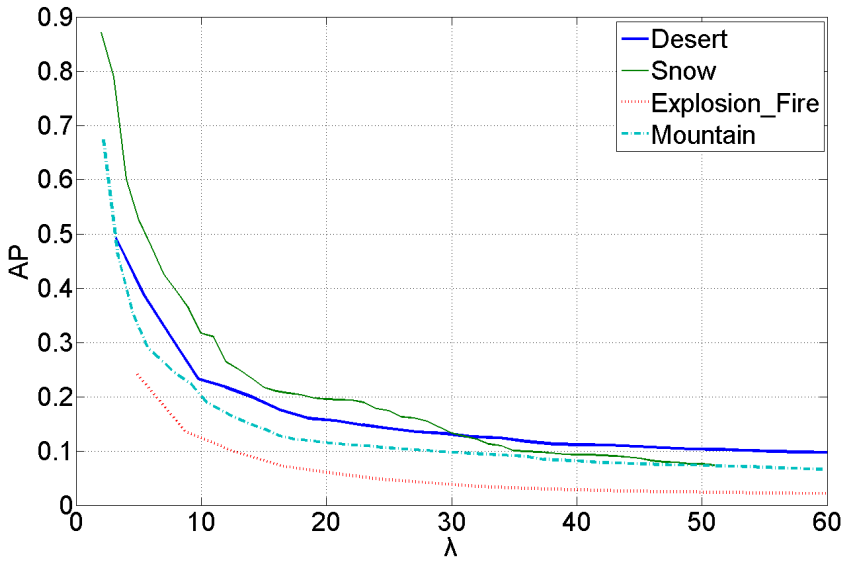


Fig. 4. AP vs. λ_{ts} for *Desert*, *Snow*, *Explosion_Fire* and *Mountain*.

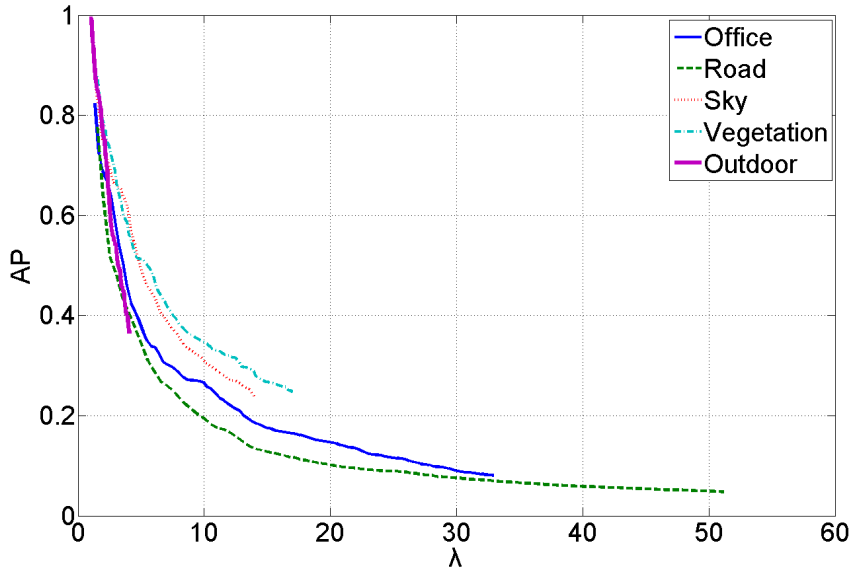
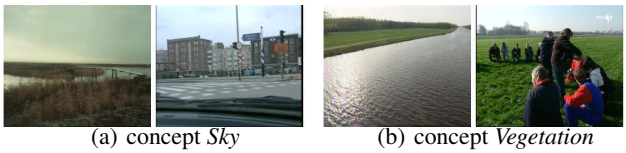


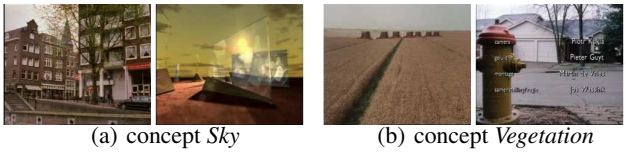
Fig. 5. AP vs. λ_{ts} for *Office*, *Outdoor*, *Road*, *Sky* and *Vegetation*.

typical region of *Sky* in both color and texture, while the right image in figure 8(b) has a dark green tone and texture too similar with the ones that a typical *Vegetation* region would have.



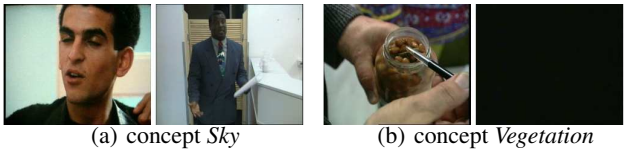
(a) concept *Sky* (b) concept *Vegetation*

Fig. 6. True positive examples.



(a) concept *Sky* (b) concept *Vegetation*

Fig. 7. False negative examples.



(a) concept *Sky* (b) concept *Vegetation*

Fig. 8. False positive examples.

7 Conclusions and Discussion

In this paper we presented our current work towards efficient semantic multimedia analysis based on a region thesaurus and a methodology on working with large data sets, such as the TRECVID collections. Extensive experiments have been presented and the effect of the ratio λ_{tr} and λ_{ts} of negative to positive examples on training and testing data has been examined. The applied generic approach tackled successfully most of the selected high-level concepts. The proposed approach is difficult to be compared to other approaches on the TRECVID data set, since we use a test set derived from the 2007 development data collection and not the actual TRECVID 2007 test data. Moreover, there does not exist any annotation for this test set and in the evaluation of all submissions the ratio λ_{ts} is kept to the maximum value. Future work aims to compare the presented algorithm with other approaches, within the same data sets derived from the TRECVID collections and to exploit the model vector representation for other applications such as keyframe extraction from videos and content based image retrieval.

8 Acknowledgements

This research was partially supported by the European Commission under contracts FP6-027685 - MESH, FP6-027026 - K-Space and FP7-215453 - WeKnowIt. Evaggelos Spyrou is partially funded by PENED 2003 Project Ontomedia 03ED475.

References

1. Saux, B., Amato, G.: Image classifiers for scene analysis. In: International Conference on Computer Vision and Graphics. (2004)
2. Gokalp, D., Aksoy, S.: Scene classification using bag-of-regions representations. In: IEEE Conference on Computer Vision and Pattern Recognition. (2007)
3. Dance, C., Willamowski, J., Fan, L., Bray, C., Csurka, G.: Visual categorization with bags of keypoints. In: ECCV - International Workshop on Statistical Learning in Computer Vision. (2004)
4. Boujemaa, N., Fleuret, F., Gouet, V., Sahbi, H.: Visual content extraction for automatic semantic annotation of video news. In: IS&T/SPIE Conf. on Storage and Retrieval Methods and Applications for Multimedia. (2004)
5. Voisine, N., Dasiopoulou, S., Mezaris, V., Spyrou, E., Athanasiadis, T., Kompatsiaris, I., Avrithis, Y., Strintzis, M.G.: Knowledge-assisted video analysis using a genetic algorithm. In: 6th International Workshop on Image Analysis for Multimedia Interactive Services. (WIAMIS 2005)
6. IBM: MARVEL Multimedia Analysis and Retrieval System. IBM Research White paper (2005)
7. Russell, B.C., Torralba, A., Murphy, K.P., Freeman, W.T.: Labelme: a database and web-based tool for image annotation. International Journal of Computer Vision (2008)
8. Naphade, M.R., Kennedy, L., Kender, J.R., Chang, S.F., Smith, J.R., Over, P., Hauptmann, A.: A Light Scale Concept Ontology for Multimedia understanding for trecvid 2005. (IBM Research Technical Report, 2005)
9. Smeaton, A.F., Over, P., Kraaij, W.: Evaluation campaigns and trecvid. In: MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval, ACM Press (2006) 321–330
10. Avrithis, Y., Doulamis, A., Doulamis, N., Kollias, S.: A stochastic framework for optimal key frame extraction from mpeg video databases. Computer Vision and Image Understanding **75 (1/2)** (1999) 3–24
11. Manjunath, B., Ohm, J., Vasudevan, V., Yamada, A.: Color and texture descriptors. IEEE trans. on Circuits and Systems for Video Technology **11**(6) (2001) 703–715
12. Spyrou, E., LeBorgne, H., Mailis, T., Cooke, E., Avrithis, Y., O'Connor, N.: Fusing MPEG-7 visual descriptors for image classification. In: International Conference on Artificial Neural Networks (ICANN). (2005)
13. Molina, J., Spyrou, E., Sofou, N., Martinez, J.M.: On the selection of mpeg-7 visual descriptors and their level of detail for nature disaster video sequences classification. In: 2nd International Conference on Semantics and Digital Media Technologies (SAMT). (2007)
14. Ayache, S., Quenot, G.: TRECVID 2007 collaborative annotation using active learning. (In: TRECVID 2007 Workshop, Gaithersburg)
15. Kishida, K.: Property of average precision and its generalization: an examination of evaluation indicator for information retrieval. NII Technical Reports, NII-2005-014E (2005)