# A Semantic Multimedia Analysis Approach Utilizing a Region Thesaurus and LSA

Evaggelos Spyrou, Giorgos Tolias, Phivos Mylonas and Yannis Avrithis
Image Video and Multimedia Laboratory
School of Electrical and Electronics Engineering
National Technical University of Athens
{*espyrou,gtolias,fmylonas,iavr*}*@image.ntua.gr*

## Abstract

*This paper presents an approach on high-level feature detection within video documents, using a Region Thesaurus and Latent Semantic Analysis*[1]*. A video shot is represented by a single keyframe. MPEG-7 features are extracted from coarse regions of it. A clustering algorithm is applied on all extracted regions and a region thesaurus is constructed. Its use is to assist to the mapping of low- to high-level features by a model vector representation. Latent Semantic Analysis is then applied on the model vectors to exploit the latent relations among region types aiming to improve detection performance. The proposed approach is thoroughly examined using TRECVID 2007 development data.*

## 1 Introduction

The continuously growing volume of multimedia content has led many research efforts to high-level concept detection, since the semantics that a document contains provide an effective and desirable annotation of its content. However, detecting the actual semantics within image and video documents remains still a challenging, yet unsolved problem. Its two main and most interesting aspects are the selection of the low-level features to be extracted and also the applied method for assigning them to high-level concepts, a problem commonly referred to as the "Semantic Gap". Many descriptors have been proposed that capture the low-level features of multimedia documents and many learning techniques such as neural networks have been successfully applied to map them into semantic concepts.

The idea of using a dictionary to describe a decomposed image derived after either clustering or segmentation or keypoint extraction has been exploited in many research efforts. In [10] a lexicon-driven approach is introduced. A region-based approach in content retrieval using Latent Semantic Analysis is presented in [11], whereas in [4], images are partitioned in regions, which are then clustered to obtain a codebook of region types, and a bag-of-regions approach is applied for scene representation. In [2] visual categorization is achieved using a bag-of-keypoints approach. Finally, in [8] separate shape detectors are trained using a shape alphabet, which is a dictionary of curve fragments.

However, this growth of audiovisual content was not accompanied by a similar growth of annotations. Very few are the databases that provide an annotation per image region, such as LabelMe[2]. On the other hand, annotating an image globally, appears a much easier task. Such an example is the one of LSCOM workshop [7], where a huge number of shots of news bulletins are globally annotated for a large number of concepts. Within TREC conference series, TRECVID evaluation [9] attracts many researchers by comparing their algorithms in various tasks such as high-level feature detection within video documents, where the goal is to globally annotate video shots for certain concepts.

This work falls within the scope of TRECVID and provides an algorithm to tackle 9 concepts within the 2007 development data. These concepts cannot be described as "objects", but rather as "materials" or "scenes". Thus, color and texture features are the only applicable low-level features. For each concept a separate neural network-based detector is trained based on features extracted from keyframe regions, while keyframes are annotated globally.

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

[1] This work was partially supported by the EC under contracts FP6-027026 K-Space and FP6-027685 MESH.

[2] `http://labelme.csail.mit.edu/`

a keyframe in section 4. The Latent Semantic Analysis technique is presented in section 5. Extensive experimental results data are presented in section 6 and conclusions are drawn in section 7.

## 2 Low-Level Feature Extraction

At a preprocessing step, a given video document, is first segmented into shots. From each shot a representative frame (keyframe) is extracted. Let $k_i \in K$ denote the keyframe for shot $s_i \in S$, where $K$ and $S$ denote the sets of all keyframes and shots, respectively. Each keyframe is first segmented into regions, using a (color) RSST segmentation algorithm, tuned to produce coarse regions. Let $r_i \in R$ denote each of the resulting regions where $R$ is the set of all regions and $R(k_i) \subset R$ is the set of all regions of the keyframe $k_i$.

For each region $r \in R$, six MPEG-7 descriptors [6] are extracted: *Dominant Color*, *Color Structure*, *Color Layout*, *Scalable Color*, *Homogeneous Texture* and *Edge Histogram* to capture color and texture features. For a region $r$, these descriptors are merged into a single vector, which will be referred to as "feature vector", denoted by $f(r)$.

## 3 Region Thesaurus

After the extraction of the feature vectors of all keyframe regions, the approach of [12] is followed, in order to provide a higher description, that will be used for the high-level concept detection.

More specifically, a K-means clustering algorithm is used to cluster all regions derived from the keyframes of the training set. The number of clusters $N_T$ is selected experimentally, after a trial and error process. This clustering is applied on the feature vectors, using the Euclidean distance as the clusters' similarity measure. Regions that lie closest to the centroids of the resulting clusters are selected to form the region thesaurus. These regions $w_i,\ i = 1, \ldots, N_T$ will be referred to as "region types". It should be made clear that each region type does not contain any high-level semantic information. However it is a higher description in comparison to a low-level descriptor.

## 4 Model Vectors

Using the region thesaurus, the distances between each region of the image and all region types are calculated. This way, a "model vector" that semantically describes the visual content of an image, is formed, by keeping the smallest distance of all image regions $r$ to each region type $w_i$. The $j$-th element of a model vector $m_i$ describing keyframe $k_i$ is depicted in eq. 1:

$$m_i(j) = \min_{r \in R(k_i)} \Big\{ d\big(f(w_j), f(r)\big) \Big\} \tag{1}$$

where $i = 1 \ldots N_K$, $j = 1 \ldots N_T$, $d(\bullet)$ denotes the Euclidean distance function, $f(w_j)$ and $f(r)$ denote the feature vectors of a region type $w_j$ and a region $r$, respectively.

## 5 Latent Semantic Analysis

The next step of the presented high-level concept detection approach, is the use of the well known Latent Semantic Analysis technique (LSA) [3], initially introduced in the field of natural language processing. LSA aims to exploit the latent relations among a set of documents and the terms they contain. In this work, a keyframe corresponds to a document and its segmented regions correspond to the terms. The goal is to investigate how these hidden relations among region types may be exploited to improve the semantic analysis.

After the formulation of the model vectors $m_i$, all their values are normalized so that they fall within $[0, 1]$, with 1 depicting the maximum confidence of a region type to a keyframe. The normalized model vectors will be denoted as $m'_i$. This way, the co-occurrence matrix $\mathcal{M}$ of eq. 2 is formed, describing the relations of region types to keyframes.

$$\mathcal{M} = \begin{pmatrix} m'_1(1) & \ldots & m'_{N_K}(1) \\ \vdots & \ddots & \vdots \\ m'_1(N_T) & \ldots & m'_{N_K}(N_T) \end{pmatrix} \tag{2}$$

More specifically, each line of $\mathcal{M}$, $q_i^T = (m'_1(i), \ldots, m'_{N_K}(i))$, describes the relationship of region type $w_i$, with each keyframe $k$ (term vector). Also, each column of $\mathcal{M}$, $m'_j = (m'_j(1) \ldots m'_j(N_T))^T$ corresponds to a specific keyframe, describing its relation with every region type (document vector).

Thus, the co-occurrence matrix $\mathcal{M}$ may be described using the extracted (normalized) model vectors $m'_i$ as:

$$\mathcal{M} = [m'^T_1, \ldots, m'^T_{N_K}] \tag{3}$$

Let $q_i$ and $q_p$ denote two term vectors. Then, their inner product $q_i^T q_p$ denotes their correlation. Thus, it may easily observed that $\mathcal{M}\mathcal{M}^T$ actually consists of all those inner products. Moreover, $\mathcal{M}^T\mathcal{M}$ consists of all inner products between the document vectors $m'^T_i m'_p$, describing their correlation over the terms.

A decomposition of $\mathcal{M}$ is described by eq. 4.

$$\mathcal{M} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \tag{4}$$

When $\mathbf{U}$ and $\mathbf{V}$ are orthonormal matrices and $\mathbf{\Sigma}$ is a diagonal matrix, This is the Singular Value Decomposition (SVD), depicted in eq. (5).

$$\mathcal{M} = \left(\mathbf{u_1} \dots \mathbf{u_{N_T}}\right) \begin{pmatrix} \sigma_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_{N_T} \end{pmatrix} \left(\mathbf{v_1} \dots \mathbf{v_{N_T}}\right)^T \tag{5}$$

In this eq. $\sigma_i$ are the singular values and $\mathbf{u}_i$,$\mathbf{v}_i$ are the singular vectors of $U$ and $V$, respectively.

By keeping the $N_L$ larger singular values $\sigma_i$ along with the corresponding columns of $\mathbf{U}$ and rows of $\mathbf{V}$, an estimate of $\mathcal{M}$, described in eq. 6 occurs.

$$\hat{\mathcal{M}} = \mathcal{M}_{N_L} = \mathbf{U}_{N_L} \mathbf{\Sigma}_{N_L} \mathbf{V}_{N_L}^T \tag{6}$$

This way the vectors of region types and keyframes are transformed to the "concept" space.

A model vector $m'_i$, is transformed to the concept space, using $\mathbf{\Sigma}$ and $\mathbf{U}$, as depicted in eq. 7.

$$\hat{m}_i = \mathbf{\Sigma}_{N_L}^{-1} \mathbf{U}_{N_L}^T m'_i \tag{7}$$

This way, all model vectors extracted from keyframes of the training set are transformed to the concept space and are then used to train several high-level concept detectors, as described in section 6.

## 6 Experimental Results

This section presents the results of the aforementioned algorithm on TRECVID 2007 Development Data, a set consisting of 110 videos, segmented into shots. A keyframe is extracted from each shot resulting to a set consisting of 18113 keyframes. The annotation used derives from a joint effort [1]. Table 1 summarizes the concepts that are detected and the number of their positive examples. A separate training and testing set has been generated for each concept. 70% of the positive examples was randomly selected for the training set of each concept and the remaining 30% for the testing set. Negative examples were selected randomly from the remaining keyframes.

After segmenting every keyframe of the training set to coarse regions a region thesaurus of 100 region types is constructed, by a K-means clustering. After extracting model vectors from all images of the training set, LSA is applied. The number $k$ of the largest singular values to keep is set to 70. Then a neural network-based detector is trained for each concept. Its input is either a model vector $m_i$ or the output of the LSA algorithm $\hat{m}_i$. Its output denotes the confidence that the specific concept exists.

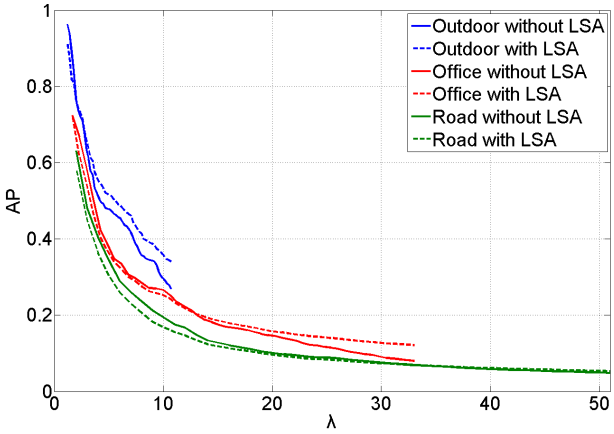


**Figure 1. AP vs $\lambda$ for *Outdoor*, *Office* and *Road*.**

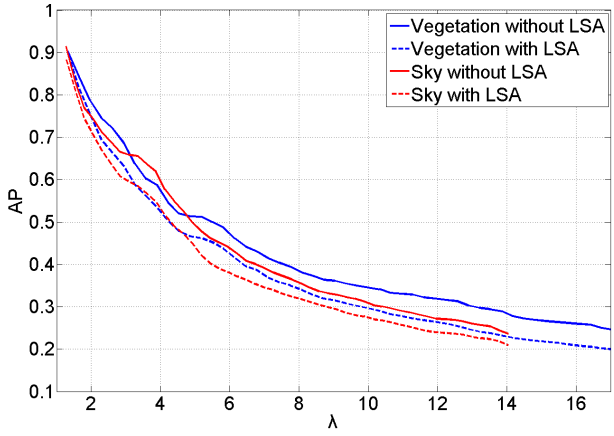


**Figure 2. AP vs $\lambda$ for *Vegetation* and *Sky*.**

Figures 1, 2 and 3 show the Average Precision (AP) [5] vs the ratio $\lambda$ of negative to positive examples. The number of positive examples is kept fixed, while the number of negative increases. It may be observed that when $\lambda$ has a relatively small value, i.e. $\lambda = 4$, as in the case of typical test sets, the performance of the classifiers remains particularly high. When $\lambda$ increases, then the performance falls. Moreover, we may observe that the use of LSA does not always improve the results. For certain concepts such as *Outdoor*, *Office* and *Road*, LSA improves the results, while $\lambda$ increases, as depicted in Fig. 1. This means that positive examples are detected in a lower and more correct rank. The common property of these concepts is that they cannot be described in a straightforward way, such as *Vegetation* and *Sky*, depicted in Fig. 2. That becomes obvious when examining the TRECVID data. Finally, for the concepts depicted in Fig. 3, where the available number of positive examples is particularly small, using LSA improves only the semantic concept *Snow*.

| concept $c_i$ | number of positives | $\lambda$=4 | | | | | | $\lambda$=max | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Before LSA | | | After LSA | | | Before LSA | | | After LSA | | |
| | | P | R | AP | P | R | AP | P | R | AP | P | R | AP |
| **Vegetation** | 1939 | 0.643 | 0.312 | 0.460 | 0.626 | 0.221 | 0.395 | 0.322 | 0.313 | 0.232 | 0.268 | 0.222 | 0.179 |
| **Road** | 923 | 0.295 | 0.046 | 0.280 | 0.400 | 0.050 | 0.210 | 0.045 | 0.047 | 0.043 | 0.036 | 0.051 | 0.044 |
| **Explosion_Fire** | 29 | 0.291 | 0.777 | 0.182 | 0.200 | 0.111 | 0.148 | 0.000 | 0.000 | 0.001 | 0.001 | 0.111 | 0.000 |
| **Sky** | 2146 | 0.571 | 0.304 | 0.436 | 0.559 | 0.271 | 0.372 | 0.258 | 0.304 | 0.214 | 0.288 | 0.207 | 0.184 |
| **Snow** | 112 | 0.777 | 0.411 | 0.460 | 0.818 | 0.264 | 0.529 | 0.013 | 0.412 | 0.008 | 0.023 | 0.265 | 0.012 |
| **Office** | 1419 | 0.446 | 0.157 | 0.318 | 0.406 | 0.147 | 0.285 | 0.117 | 0.157 | 0.072 | 0.095 | 0.148 | 0.110 |
| **Desert** | 52 | 0.333 | 0.312 | 0.287 | 0.215 | 0.687 | 0.246 | 0.003 | 0.313 | 0.064 | 0.001 | 0.438 | 0.063 |
| **Outdoor** | 5185 | 0.425 | 0.514 | 0.361 | 0.331 | 0.634 | 0.382 | 0.683 | 0.510 | 0.515 | 0.601 | 0.646 | 0.522 |
| **Mountain** | 97 | 0.444 | 0.137 | 0.241 | 0.110 | 0.035 | 0.072 | 0.003 | 0.379 | 0.037 | 0.003 | 0.172 | 0.001 |

**Table 1. Experimental results for all high-level concepts. P=precision, R=recall, AP=average precision.**

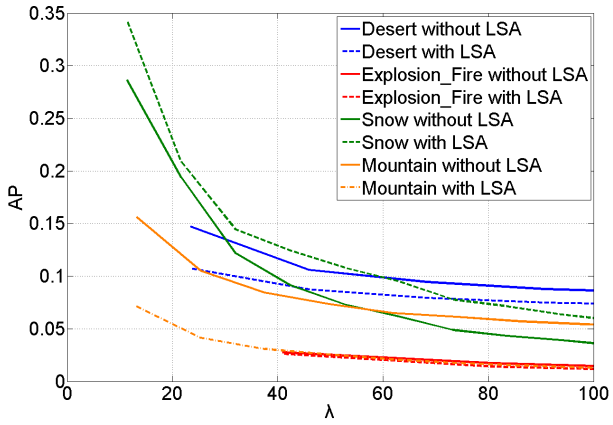


**Figure 3. AP vs $\lambda$ for *Desert*, *Explosion_Fire*, *Snow* and *Mountain*.**

## 7 Conclusions and Future Work

In this paper we presented our proposal towards efficient semantic multimedia analysis based on a region thesaurus and LSA. We used a clustered set of region types to produce a model vector and thus map low-level features to high-level concepts. LSA was then applied on the latter, using the TRECVID 2007 Development Data. Within our future goals are the utilization of contextual information in the process and the extension of the proposed methodology to support detection of a large number of concepts.

## References


[1] S. Ayache and G. Quenot. TRECVID 2007 collaborative annotation using active learning. In *TRECVID 2007 Workshop, Gaithersburg*.

[2] C. Dance, J. Willamowski, L. Fan, C. Bray and G. Csurka. Visual categorization with bags of keypoints. In *ECCV - International Workshop on Statistical Learning in Computer Vision*, 2004.

[3] S. Deerwester, S. 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.

[4] D. Gokalp and S. Aksoy. Scene classification using bag-of-regions representations. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[5] K. Kishida. Property of average precision and its generalization: an examination of evaluation indicator for information retrieval. NII Technical Reports, NII-2005-014E, 2005.

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

[7] 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.

[8] A. Opelt, A. Pinz and A. Zisserman. Incremental learning of object detectors using a visual shape alphabet. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2006.

[9] A. F. Smeaton and P. Over and W. Kraaij. Evaluation campaigns and TRECVid. In *MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, 2006.

[10] C. Snoek, M. Worring, D. Koelma and A. Smeulders. A learned lexicon-driven paradigm for interactive video retrieval. *IEEE Trans. on Multimedia*, 9(2):280–292, 2007.

[11] F. Souvannavong, B. Merialdo and B. Huet. Region-based video content indexing and retrieval. In *4th International Workshop on Content-Based Multimedia Indexing*, 2005.

[12] E. Spyrou and Y. Avrithis. A region thesaurus approach for high-level concept detection in the natural disaster domain. In *2nd International Conference on Semantics And digital Media Technologies (SAMT)*, 2007.