D

Detection of High-Level Concepts in Multimedia

Evaggelos Spyrou and Yannis Avrithis
National Technical University of Athens, Greece

Synonyms: *High-Level Concept Detection; Semantic Concept Detection; Image Classification; Object Recognition.*

Definition: *High-level concept detection is the process within which the high-level concepts or objects which are presented in a multimedia document are determined. For example, given an image, a detection scheme would reply that it contains concepts such as "sky", "sand", "sea", the image depicts an "outdoor" and more specifically a "beach" scene. In some cases, the actual position of concepts within the image is also detected.*

Introduction

The continuously growing volume of multimedia content has led many research efforts to high-level concept detection, since the semantics 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 the method that will be used for assigning low-level descriptions to high-level concepts. Finding an automatic transition from the low-level features to semantic entities or equivalently the automatic extraction of high-level characteristics is an extremely hard task, a problem commonly referred to as the "Semantic Gap" [1]. Many descriptors have been proposed that capture the audio, color, texture, shape and motion characteristics of audiovisual documents, or in other words, their low-level features. On the other hand, many techniques such as neural networks, fuzzy systems, and support vector machines have been successfully applied in the aforementioned problem.

For every classification or detection problem, the following steps are almost always followed: First, certain visual features are extracted. These features capture the color, texture and shape properties of the image and in some cases also those of certain image regions. Using a significantly large set of training data, a high-level concept detector is trained. This module acts on every given image, extracts its visual features and decides which high-level concepts exist within it. Moreover, in approaches that exploit the contextual properties of concepts and image regions, a next and final step refines the initial degrees of confidence for the detected concepts and produces the final results.

Visual Descriptors

In every classification, detection or recognition problem, the first step is the extraction of the "low-level features" that exist within the multimedia document. By using the term "low-level", we mean that these features do not carry any semantic information. These features often follow a standardized method of extraction and are called "*descriptors*". Descriptors are extracted in a way that visually similar objects and concepts have similar descriptions and moreover that they should be invariant against conditions of image/video capture such as illumination, rotation and other camera parameters.

The most common visual features that are extracted are color, texture, shape and motion. During the last few years, a large number of various descriptors have been proposed in the literature. To fulfill the needs for standardized and efficient visual descriptors, the MPEG-7 [1] standard has been proposed. Unlike its predecessors, it focuses on description of multimedia content and aims to provide interoperability among applications that use audio-visual content descriptions. MPEG-7 provides various color, texture, shape and motion standardized descriptors that extract visual, low-level, non-semantic information from images and videos and use it to create structural and detailed descriptions of audio-visual information. MPEG-7 color descriptors are mainly color histograms, texture descriptors capture the energies, orientations and distributions of textures and shape descriptors capture certain region or contour properties.

Another popular and effective descriptor is the Scale-Invariant Feature Transform (SIFT) [17]. The SIFT features are locally extracted and based on the appearance of an object at particular interest points. Also they are invariant to scale and rotation and robust to illumination changes, noises, occlusion and minor viewpoint changes. Within the SIFT algorithm, certain points within an image are selected as "points of interest" or "keypoints". The initial set of keypoints is then filtered and a set of them is discarded. At each point, one or more orientations are assigned based on local image gradient directions. The feature descriptor is finally computed as a set of orientation histograms.

A visual descriptor often has the form of a vector, with each constituent corresponding to a certain visual feature. These vectors are often referred to as "feature vectors" and do not carry any semantic information. Depending on the problem in question, it is crucial to specify from which part of the document the selected descriptors are extracted. Thus, apart from extraction of the whole image/video (globally), descriptors are often extracted from grids of the image, from image blocks, from regions resulted from segmentation[1] or from regions near characteristic keypoints of the image.

Machine Learning Approaches

The main characteristic of learning-based approaches is their ability to adjust their internal structure according to input and respective desired output data pairs in order to approximate the relations implicit in the provided (training) data, thus elegantly simulating a reasoning process. Consequently, machine-learning approaches constitute an appropriate solution when the considered a-priori knowledge cannot be defined explicitly because it is ill-defined, incomplete or too large in terms of amount to be efficiently represented. Among the developed techniques this paragraph describes those that have been widely used in the area of pattern classification in general.

---

[1] *Image Segmentation* is a process that divides images into regions using certain criteria of homogeneity such as color and/or texture

Neural networks [9] encode sampled information in a parallel-distributed framework. The knowledge they gain from training makes them capable of discriminating between objects or patterns.

Fuzzy Systems [8] encode structured, empirical (heuristic) or linguistic knowledge in a similar numerical framework. They are able to describe the operation of the system in natural language with the aid of human-like if-then rules. However, they do not provide the highly desired characteristics of learning and adaptation.

Support Vector Machines [11] are capable of solving classification problems that are non-separable in the input space by performing a non-linear transformation. Thus, the problem driven into the higher-dimension (feature) space may be linearly separable.

Neurofuzzy networks [10] bridge the gap between neural networks and fuzzy systems, as the use of neural networks in order to realize the key concepts of a fuzzy logic system enriches the system with the ability of learning and improves the sub-symbolic to symbolic mapping.

Genetic Algorithms [12] solve problems using an evolutionary process. The algorithm begins with a set of solutions (represented by chromosomes) called a population. Solutions from one population are taken and used to form a new population. This is motivated by a hope, that the new population will be better than the old one. Solutions, which are then selected to form new solutions (offspring), are selected according to their fitness - the more suitable they are the more chances they have to reproduce.

AdaBoost [9] (Adaptive Boosting) is an algorithm that can be used with almost any of the aforementioned techniques. The goal is to linearly combine more than one weak[2] classifiers and to adapt the weight assigned to each one in order to produce a final strong classifier.

Scene Classification

At the early years of image classification, many researchers focused on high-level concepts that globally characterize the semantics of an image or video document. For example, an image may be characterized either as "indoor" or as "outdoor", depending on the location it was taken. The same stands for a video document. In this case, the problem is often called "scene detection". The choice of the visual descriptors to extract appears crucial in such problems and is based on a priori knowledge of the domain properties. The selected descriptors are then extracted globally from an image on a pixel per pixel basis, or from one or more representative frames (keyframes) of a video document. Then, in most cases, a learning approach such as a neural network or a support vector machine is applied to link the extracted low-level features to the semantic concepts. Herein we present some common methods for certain scene classification problems.

As it has already been mentioned, the most common scene detection problem is the "indoor/outdoor" classification [3]. Although this task appears easy to a human observer, it is not very easy to develop a simple yet robust technique, such both indoor and outdoor images are extremely heterogeneous in terms of their visual properties. However by extracting and fusing color and texture descriptors and then using a machine learning approach often the best classification results are achieved. To explain

[2] *Weak* classifiers are those with a nearly random performance, i.e. for a binary problem a weak classifier would have a performance slightly over 50%. With many techniques such as *boosting/AdaBoost*, a combination of many weak classifiers leads to a *strong* classifier.

that, we may consider that a typical outdoor image contains high-level concepts such as "sky", "vegetation", "road", "sea" etc. This gives certain color and texture properties to outdoor images thus they can be distinguishable from the indoor ones.

Other scene detection problems include "city/landscape" classification [4] which also appears as one of the most common problems. In this case, most of the times only texture features are extracted. More specifically texture descriptors that capture the energy of edges and their distribution over images are preferred. That is because in a typical "city" image high-level concepts such as "buildings" appear. As obvious, buildings have strong horizontal and vertical edges. On the other hand, in a typical landscape image, weak and non-oriented edges are common, while horizontal and vertical are rare. Apart from those classical scene detection problems, there are many other variations. For example in [5] a "beach/city" classification problem is dealt with fusion of various MPEG-7 color and texture descriptors, since solely color or texture features appear inadequate. It is also shown that classification performance does not always increase when the number of available descriptors increases. Moreover, different classifier algorithms seem to work better with certain descriptors. Thus, the choice of descriptors should be dependent to the choice of the classification method.

Apart from the aforementioned techniques that use visual low-level features extracted from the whole image or a keyframe extracted from a video, many research works tend to use features extracted locally, from parts or segments of the image. These methods are presented in the next paragraph, as they are similar to those used for detecting "material"-type high-level concepts.

"Material"-like High Level Concepts

By "materials" we denote those high-level concepts that do not have a specific shape. Thus, the only applicable low-level features in this case are color and texture descriptors. Due to the fact that these high-level concepts are very often in multimedia documents, and easier to detect than other concepts, they have become very popular in the research field. As it will become obvious, materials do not characterize the whole image or video document, but consist only a small part of it. Typical concepts within this category are "sea", "sky", "sand", "vegetation", "road", "snow" etc. The detection methods for this category of objects can be divided into two sub-categories, depending on the available annotation that may be used for training purposes.

When the multimedia documents are annotated globally, usually the goal is to detect whether the concept in question exists within them, and not its actual position. Since there does not exist any annotation available for every region, one cannot train specific detectors for each concept. Thus, researchers choose to apply generic approaches to detect all concepts that can be characterized as materials. Many algorithms use a visual vocabulary that consists of the most common regions that are encountered within the available training data. These regions are often called "region types". For example, in [6], a visual vocabulary is formed after clustering of all regions of the training set. By keeping the centroids of the formed clusters, the set of the region types is defined. Then, for each image or video keyframe a vector is formed. This vector contains the information about the relation to all the words of the visual dictionary in terms of their distances. In 0, a visual dictionary consisting of 8 region types is depicted. For each region, the distance to every visual word is calculated (fig. 1a). The smallest distance is kept for every visual word (fig. 1b) and the model vector is formed. As obvious, each concept is related to certain region types. For example, an image that contains the high-level concept "sky" should also contains light blue, non-textured region types, while one that contains the high-level concept vegetation should contain green and textured region types. A neural network-based detector is trained to learn this mapping.

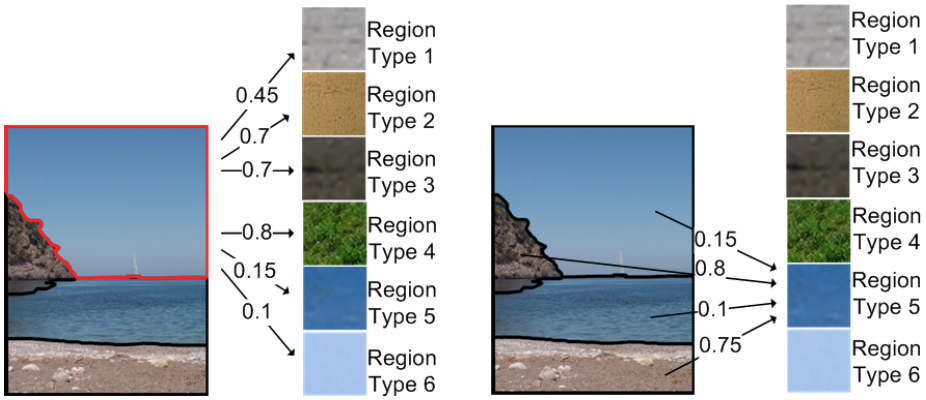

Figure 1: Constructing a model vector to semantically describe a segmented image based on a visual dictionary of region types.

On the other hand, in cases where annotation per image region is available, the approaches that are used are completely different. For example in [15], knowledge about the different high-level concepts is stored in the form of manually selection region prototypes in a domain ontology. An expert manually selects the region prototypes from the available dataset. Then, their low-level descriptors are extracted and stored. Using a segmentation algorithm, a given image is segmented into regions based on their color features. For each segment, the confidences to all the concept prototypes of the ontology are calculated. The one with the highest confidence is finally assigned. Finally, a genetic algorithm refines the confidences and produces the final results. However, the basic disadvantage of this method is that it requires a lot of effort both to manually extract and annotate regions used to build the knowledge base and moreover to be applied using a large knowledge base.

Another technique, presented in [14], starts with collecting examples for every "material"-like high-level concept. Using color and texture features, a separate classifier is trained. These classifiers make a decision for every pixel in the given image. This way, every pixel is either assigned to a material or left unclassified. Neighboring pixels that belong to same concept are merged. This way, the regions that contain a single material are formed. As it is obvious, for some concepts represented by non-homogeneous regions, this method would fail. To overcome this, keypoints are defined using the same method as the SIFT algorithm and then color and texture features are extracted from 16x16 blocks around them. By using a visual dictionary, each region is then assigned to a visual concept.

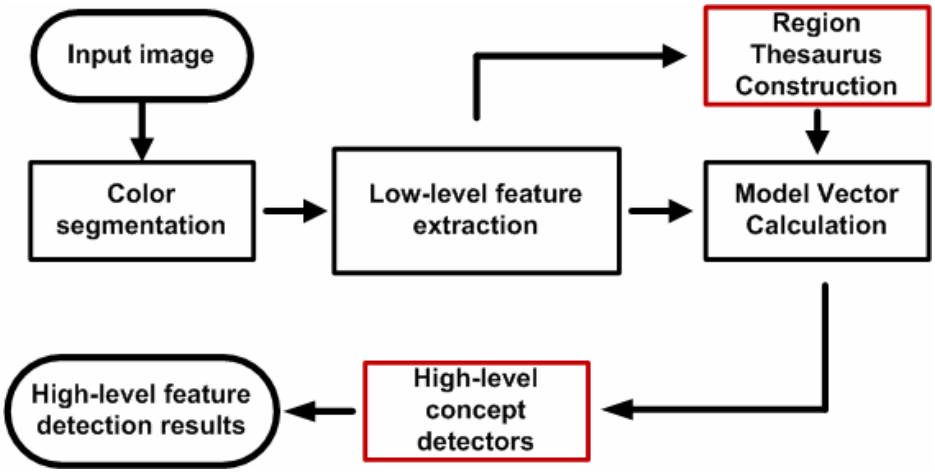

Figure 2: A thesaurus-based approach for the detection of high-level concepts.

Finally, in [18] the classification process starts with the extraction of Local Interest Points (LIPs). These points, often denoted as "salient" tend to have significantly different properties compared to all other pixels in their neighborhood. To extract these points a method called Difference-of-Gaussian (DoG) is applied. From each LIP, a SIFT descriptor is extracted from an elliptic region. Generally the number of LIPs in a keyframe can range from several hundreds to few thousands. A visual dictionary is generated by an offline quantization of LIPs. Then, using this dictionary, each keyframe is described as a vector of visual keywords. This way, direct keyframe comparison is facilitated, since instead of point-to-point LIP matching (large number of comparisons), matching of visual words is performed (significantly decreased number of comparisons). For each high-level concept, a classifier is trained. The basic advantage of such methods is that the extracted LIPs are invariant to many translations such as rotation, scaling etc, while its main disadvantage is the increased computational cost.

Object Detection

This category consists of high-level concepts that can be described explicitly by their shape. Some examples of this category are "person", "boat", "car", "building", "airplane" etc. Concepts that belong in this category are often the most difficult to detect. The most common approaches that are encountered in this research field can be divided into two major categories.

The first category consists of methods which, given an image region and by extracting its shape properties, try to match them to a specific object. Usually, certain descriptors similar to Region Shape and Contour Shape descriptors defined by the MPEG-7 standard are used. The basic disadvantage of those methods is that they work well only after nearly perfect image segmentation. However, in real world multimedia problems this case does not exist. Segmentation fails to produce segments that correspond to real objects and also factors such as partial occlusion of objects, lighting conditions etc worsen their results.

The second category is based on object detection by identifying certain parts. For example in [13], a shape vocabulary is first built. This vocabulary consists of curve fragments. Along with them, the vocabulary also contains information about their centroids and the categories of objects to which they apply. The shape of an object is decomposed to a set of "boundary fragments". Then, weak classifiers are trained to detect pairs of fragments. Their combination leads to strong classifiers, that detect specific objects such as "cars", "horses", "bottles" etc. The main differences to the methods presented in the previous sections is that this time the actual locations of objects within images are determined and moreover this method appears robust to occlusion, illumination etc.

In the same category, we should also mention the Viola-Jones approach for object detection [21]. This approach uses a novel image representation which is called "integral image" and measures differences in the intensity between image pixels. Then, an algorithm similar to AdaBoost is used to train a cascade of classifiers. This algorithm is proved to be very fast for real time applications, is robust to changes of illumination, angle, scale etc. The most important is that it is a generic algorithm that can be used to detect many high-level concepts. In the original work it is used to detect faces but can be easily extended to detect other objects.

Finally, in [20], a graphical model relating features, objects and scenes is presented. A large number of different features is extracted from "patches" of the image and after a feature selection process, a subset of them is kept. Then, and by using a boosting process, namely GentleBoost (another variation of AdaBoost) detectors are trained for concepts such as "screen", "car", "pedestrian" etc. To improve speed and accuracy of detectors,

the search space for objects is reduced in terms both of scale and location. Moreover, concepts are detected without determining their actual location and finally models that use contextual knowledge are also presented. For example if a "computer screen" is detected, then a "keyboard" is expected to be seen. This approach falls in between these presented within this section and the contextual approaches presented in the following section.

Using Contextual Relations to Enhance Detection Results

As can be found in the literature, the term *context* may be interpreted and even defined in numerous ways, varying from the philosophical to the practical point of view, none of which is globally applicable or universal. Therefore, it is of great importance to establish a working representation for context, in order to benefit from and contribute to multimedia analysis and media adaptation. The problems to be addressed include how to represent and determine context, both in terms of low- and high-level visual characteristics, and how to use it to optimize the results of multimedia analysis. The latter are highly dependent on the domain an image belongs to and thus in many cases are not sufficient for the understanding of multimedia content. The lack of contextual information in the process significantly hinders optimal analysis performance and together with similarities in low-level characteristics of various objects types, results in a significant number of misinterpretations.

Figure 3: Above: Cloud, Snow and Wave outside their context, Below: The same concepts inside their context

In [16], a visual context ontology is constructed. This ontology models the contextual relations among the various high-level concepts of the given domain of interest, among the various region types that exist within the available training set of images and among high-level concepts and region types. After the application of a typical concept detection algorithm, initial confidence values for each high-level concept are estimated. The visual context algorithm exploits the known relations (see 0) that are encoded within the ontology and based on them and the initial detection confidence values, refines the results. For example, if the concept "sea" has a strong relation with the concept "sky" and the initial confidence value of "sea" is very high, while the one of "sky" is ambiguous (near 0.5), then the latter is increased towards the one of the first.

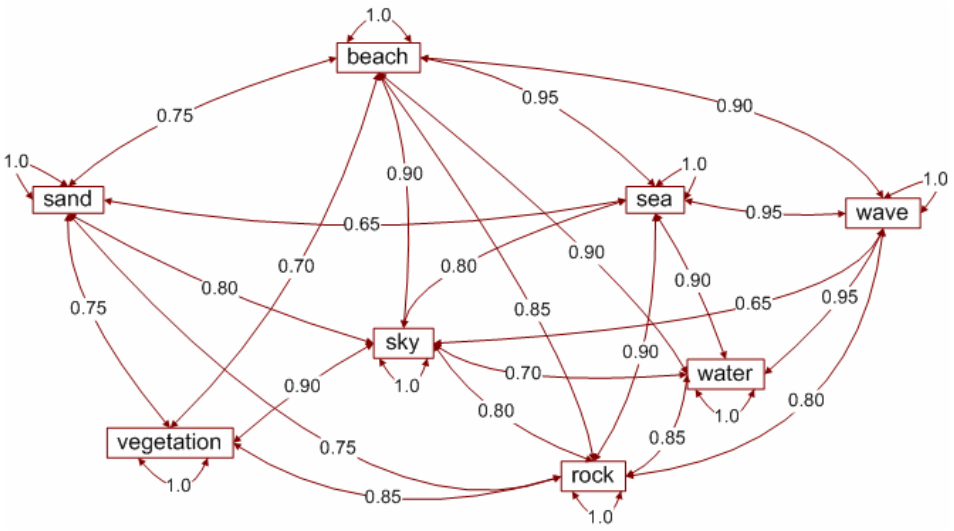

Figure 4: A visual context ontology for the "beach" domain. The relations between high-level concepts (nodes of the graph) are depicted in the edges of the graph.

Moreover, in [19] another category of visual context is presented, that deals with the spatial relations among the different materials. High-level concept detection based on the co-occurrence of high-level concepts, without considering their structure within an image, often fails. That is because certain materials sometimes present similar visual properties. For example, "sky" and "sea" regions may sometimes be indistinguishable when observed outside their context. Since "sky" almost always is *above* "sea", when two indistinguishable regions occur and their possible labels are among "sky" and "sea", then the upper is labeled as "sky" and the lower as "sea". This work makes use of a probabilistic framework to model a set of 7 spatial relations. Initial detection confidence for a material is based on low-level color and texture features and a neural network approach. Then, using a Bayesian network, an initial belief vector is generated for each segmented region in an image, based on the individual material detection. This belief is updated by the spatial context module and the final degrees of confidence occur.

Finally, in [7] a similar approach to the aforementioned contextual ones is applied, based on a technique derived from natural language processing, called Latent Semantic Analysis. Its goal is to uncover the hidden (latent) relations among the existing region types that are encountered within the documents of the available training set. Latent relations among a set of documents and the terms they contain. In this work, a video keyframe corresponds to a document and its segmented regions correspond to the terms. Experimental results denote that for certain high-level concepts that are better "defined", detection performance improves. What is different than the previous methods presented within this section is that this time the relations are not defined by a domain expert or statistically but are "discovered" by the algorithm.

Datasets and Benchmarks

During the last years and following the growing research interest in the area of high-level concept detection, an increasing number of available annotations has started to appear. One example of such an annotation is the one of the LSCOM workshop [22], where a huge number of shots of news bulletins are globally annotated for a large number of concepts. On the other hand, annotated data sets per region are very rare. Special note should be given to an effort to effectively evaluate and benchmark various approaches in the field of information retrieval, by the TREC conference series, during the last few years. Within this series the TRECVID [23] 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 the high-level feature detection task of TRECVID, global annotations have been offered by different organizations and a huge database of video keyframes has been available to active participants.

Conclusions

This article presented some state-of-the-art methods for detecting high-level concepts within multimedia documents. As it became obvious, this problem is huge, having many different aspects. There does not exist a single algorithm able to tackle any semantic concept in any set of multimedia documents. Thus, a researcher should each time observe every aspect of the problem in question, select appropriate visual features to extract and combine them efficiently to train concept detectors.

References

1. A.W.M. Smeulders, M. Worring, S. Santini, A. Gupta and R. Jain,"Content-based image retrieval at the end of the early years", IEEE Trans. PAMI 22 (2000)
2. S.F. Chang, T. Sikora and A. Puri, "Overview of the MPEG – 7 standard", IEEE Trans. on Circuits and Systems for Video Technology Vol.11 – no.6, June 2001
3. M. Szummer and R. Picard, "Indoor-outdoor image classification". In: IEEE international workshop on content-based access of images and video databases, (1998)
4. A. Vailaya, A. Jain and H.J. Zhang, "On image classification: City images vs. landscapes". Pattern Recognition 31 (1998) 1921–1936
5. E. Spyrou, H. Le Borgne, T. Mailis, E. Cooke, Y. Avrithis, and N. O´Connor, "Fusing MPEG-7 Visual Descriptors for Image Classification", W. Duch et al. (Eds.): ICANN 2005, LNCS 3697, pp. 847-852, 2005.
6. E. Spyrou and Y. Avrithis, "A Region Thesaurus Approach for High-Level Concept Detection in the Natural Disaster Domain", 2nd international conference on Semantics And digital Media Technologies (SAMT), 2007, Genova
7. E. Spyrou, G. Tolias, Ph. Mylonas and Y. Avrithis, "A Semantic Multimedia Analysis Approach Utilizing a Region Thesaurus and LSA", International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS), 2008, Klagenfurt, Austria
8. G.J. Klir and B. Yuan, "Fuzzy Sets and Fuzzy Logic, Theory and Applications", Prentice Hall, 1995
9. S. Haykin, "Neural Networks: A comprehensive foundation", Second Edition, Prentice Hall, 1999
10. C.-T. Lin and C.S. Lee, "Neural fuzzy Systems: A neuro-fuzzy synergism to intelligent systems", Prentice-Hall, Englewood Cliffs, NJ, 1995
11. V.N. Vapnik, "The nature of statistical learning theory", Springer, New York, 1995.
12. M. Mitchell, "An introduction to Genetic Algorithms", MIT Press, 1996
13. A. Opelt, A. Pinz and A. Zisserman, "Incremental learning of object detectors using a visual shape alphabet", In Proc. of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '06), Washington, DC, USA
14. D. Gokalp and S. Aksoy, "Scene Classiffcation Using Bag-of-Regions Representations", In Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR '07)
15. 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 Proc. of 6th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS 2005)

16. Ph. Mylonas, E. Spyrou and Y. Avrithis, "Enriching a context ontology with mid-level features for semantic multimedia analysis",1st Workshop on Multimedia Annotation and Retrieval enabled by Shared Ontologies, co-located with SAMT 2007
17. D. Lowe, "Distinctive image features from scale-invariant keypoints", International Journal of Computer Vision. 60(2):91:110, 2004.
18. W.L. Zhao, Y.G. Jiang, and C.W. Ngo, "Keyframe retrieval by keypoints: Can point-to-point matching help?" Proc. of International Conference on Image and Video Retrieval, Tempe, AZ, USA, 72-81, Springer, 2006.
19. A. Singhal, J. Luo and W. Zhu, "Probabilistic Spatial Context Models for Scene Content Understanding", in Proc of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'03)
20. P. Murphy, A. Torralba and W.T. Freeman "Using the forest to see the trees: a graphical model relating features, objects and scenes", Adv. in Neural Information Processing Systems 16 (NIPS), Vancouver, BC, MIT Press, 2003.
21. P. Viola and M.J. Jones, "Robust Real-Time Face Detection", International Journal of Computer Vision 57(2), 137–154, 2004
22. "LSCOM Lexicon Definitions and Annotations Version 1.0", DTO Challenge Workshop on Large Scale Concept Ontology for Multimedia, Columbia University ADVENT Technical Report #217-2006-3, March 2006.
23. A.F. Smeaton, P. Over and W. Kraaij, "Evaluation campaigns and TRECVid", In Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval (Santa Barbara, California, USA, October 26 - 27, 2006). MIR '06.