A relation-based contextual approach for efficient multimedia analysis

E. Spyrou, G. Tolias, Ph. Mylonas
Image, Video and Multimedia Laboratory
National Technical University of Athens
Zographou Campus, PC 157 80, Athens, Greece
{espyrou, gtolias, fmylonas}@image.ntua.gr

Abstract

In this paper we present our research work on the identification of high-level concepts within multimedia documents through the introduction and utilization of contextual relations. A conceptual ontology is introduced, as the means of exploiting the visual context of images, in terms of high-level concepts and region types they consist of. A meaningful combination of these features results in a computationally efficient handling of visual context and extraction of mid-level characteristics towards the ultimate goal of semantic multimedia analysis. Evaluation results are presented on a medium-size dataset, consisting of 1435 images, 25 region types and 6 high-level concepts derived from the beach domain of interest.

1 Introduction

Most current content-based image analysis and retrieval systems are limited by the existing state-of-the-art in image understanding, in the sense that they usually fall short of higher-level interpretation and exploitation of contextual knowledge. Combining the latter with traditional image or scene classification techniques, in order to achieve better semantic results during the content analysis phase, forms a challenging and broad research problem. It was only recently, that multimedia analysis systems have started using semantic knowledge technologies, as they are defined by notions like ontologies [17] or folksonomies [6].

Among the most interesting tasks in multimedia content analysis is the detection of high-level concepts within multimedia documents. Acknowledging the need for providing such an analysis, many research efforts set focus on low-level feature extraction in a way to efficiently describe the various audiovisual characteristics of a multimedia document. However, the widely discussed "semantic gap"[13] characterizes the differences between descriptions of a multimedia object by different representations and the linking from low- to high-level features. An important step for narrowing this gap is to automate the process of semantic feature extraction of multimedia content objects, by enhancing image and video classification with semantic characteristics and knowledge.

Plenty of indicative works exist towards the solution of this problem. In [5] a multimedia analysis and retrieval system is presented, using multi-modal machine learning techniques in order to model semantic concepts in videos. Moreover, in [14], a region-based approach in content retrieval that uses Latent Semantic Indexing (LSI) techniques is proposed. In [11] the features are extracted by segmented regions of an image. Also, in [19] a region-based approach is presented, that uses knowledge encoded in the form of an ontology. In [1] a hybrid thesaurus approach is presented. Finally, a lexicon used in an approach for an interactive video retrieval system is presented in [4].

Furthermore, the aspect of contextual knowledge in multimedia is introduced in [18] and [8], as an extra source of information for both object detection and scene classification. Recent research efforts investigated certain types of context, such as spatial context [20] (i.e. topological relationships between regions in the same scene), temporal context [3] (within video sequences or between images belonging to a particular image collection), or imaging context [2], in conjunction with image content features in the forms of either low-level features (e.g. color, texture, shape) or semantic concepts (e.g. *sky*, *vegetation*, *sand*, and *sea*). In the rest of our paper we shall refer to the term *visual context*, by interpreting it as **all information related to the visual scene content of a still image or video sequence that may be useful during its analysis phase**.

The structure of this paper is as follows: Section 2

978-0-7695-3444-2/08 $25.00 © 2008 IEEE
DOI 10.1109/SMAP.2008.39

deals with the proposed mid-level conceptualization. In Section 3 the overall fuzzy context knowledge formalization is described, together with the proposed contextual adaptation in terms of the visual context algorithm and its optimization steps. Section 4 lists our experimental results derived from the *beach* domain and Section 5 concludes briefly our work.

2 Concepts' Detection using a Region Thesaurus

In this section we propose to tackle the problem of high-level concept detection through an innovative way based on mid-level information and features [15]. This research effort has been initially discussed in [9] and [16] and is further expanded and strengthened herein. In general, visual features that may be extracted from a still image or video document can be divided in two major categories: *low*-level visual features, which may provide a qualitative or quantitative description of the visual properties, and *high*-level features, which describe the visual content of an image in terms of its semantics. One fundamental difference between those categories is that low-level features may be calculated directly from an image or video, while high-level features cannot be directly extracted but are often determined by exploiting the low-level features. In the following we shall describe briefly the extraction of low-level features and the construction of a corresponding *region thesaurus*, whereas Figure 1 presents the overall methodology.

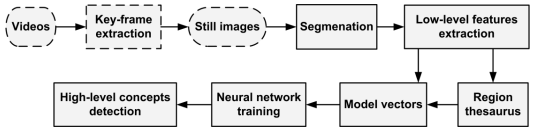

Figure 1. High-level concept detection algorithm

After a trivial step of basic MPEG-7 color and texture features extraction, the next important step aims to bridge the low-level features to the high-level concepts. To achieve this, we construct a visual dictionary and with its aid we form a mid-level image description. This description will contain all the necessary information to connect an image with every visual word of the dictionary. This way, we achieve to keep a fixed-size image description and face the problem that the number of segmented regions is not fixed *a priori*. Moreover this mid-level description will prove useful when contextual relations will be exploited.

Initially we select the appropriate region types. We start from an arbitrary large number of segmented regions and we apply an efficient *hierarchical clustering* algorithm [10] on them. After this process, each cluster may or may not contain a high-level feature and each high-level feature may be contained in one or more clusters; i.e. the concept *sand* may be represented by many instances differing e.g. in the color or the texture. Moreover, in a cluster that may contain instances from a semantic entity (e.g. *sea*), these instances could be mixed up with parts from another visually similar concept (e.g. *sky*). Finally, we do select the region type that represents each cluster as the closest region to its centroid.

We move forward by formally describing the constructed visual dictionary (thesaurus) T as a set of visual words w_i by equation (1).

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

$$\bigcup_i w = R, \quad i = 1 \dots N_T \tag{2}$$

$$\bigcap_{i,j} w = \emptyset, \quad i \neq j \tag{3}$$

where N_T is the number of region types of the thesaurus (and, obviously, the number of clusters) and w_i is the i-th cluster, which is a set of regions that belong to R, as it is presented in equation (1). The region types are the centroids of the clusters (4) and the rest feature vectors of a cluster are their synonyms. By using a significantly large training set of images/keyframes, the entire thesaurus is constructed. Its purpose is to formalize a conceptualization between the low- and the high-level features and facilitate their association.

$$z(w_i) = \frac{1}{|w_i|} \sum_{r \in w_i} f(r) \tag{4}$$

$$f(w_i) = f\Big(\arg\min_{r \in w_i} \Big\{ d\big(f(r), z(w_i)\big) \Big\} \Big) \tag{5}$$

Each region type is represented by its feature vector that contains all the extracted low-level information for it (5). As it is rather obvious, a low-level descriptor does not carry any semantic information. It only constitutes a formal representation of the extracted visual features of the region. On the other hand, a high-level concept carries only semantic information. Thus, a region type lies in-between those features. It contains the necessary information to formally describe the color and texture features, but can also be described with a *lower* description than the high-level concepts. I.e., one can describe a region type as *a green region with a coarse texture*.

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 mid-level concept (region type). In particular, the model vector describing image p_i, will be:

$$m_i = \Big[m_i(1), \ldots, m_i(j), \ldots, m_i(N_T)\Big],\ i = 1 \ldots N_K \tag{6}$$

where:

$$\begin{aligned} m_i(j) &= \min_{r \in R(k_i)} \Big\{ d\Big(f(w_j), f(r)\Big)\Big\} \qquad (7) \\ i &= 1 \ldots N_K, \quad j = 1 \ldots N_T \end{aligned}$$

Each model vector is denoted by $m_i \in M,\ i = 1 \ldots N_K$, where M is the set of all model vectors and m_i is the model vector of image/keyframe k_i. More specifically, the j-th element of a model vector contains the minimum distance amongst all distances between the j-th region type and all the image' s regions (8).

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 an image/keyframe in terms of the region thesaurus. The output of the network is the confidence that the image/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 image/keyframe in question.

3 Mid-level Visual Context

In order to fully exploit the notion of visual context and combine it with the aforementioned mid-level region types, we further refine the initial high-level concept detection results by exploiting solely contextual (in comparison to visual) relations between high-level concepts. This approach differentiates itself from most of the related research works, because it deals with a global interpretation of the image and the concepts that are present in it. In other words, high-level concepts either exist or do not exist within the entire image under consideration and not within a specific region of interest (e.g., the image might contain concept *water*, but there is no information regarding its spatial allocation). The same approach is adopted by the well-known TRECVID experiments series [12].

In order to further adapt the results of low-level, descriptor-based multimedia analysis, utilizing the notion of mid-level region-types, we introduce a concept-based method, founded on an enhanced high-level contextual ontology; the latter is described as a set of *concepts* and *semantic relations* between concepts within a given universe. In general, we may decompose such an ontology $\mathcal{O}_C$ into two parts, i.e.

1. the set C of all semantic concepts $c_i \in C, i = 1 \ldots n$ and
2. the set R_{c_i,c_j} of all semantic relations amongst any two given concepts $c_i, c_j, j = 1 \ldots n$

More formally:

$$\mathcal{O}_C = \{C, R_{c_i,c_j}\}, \quad R_{c_i,c_j} : C \times C \rightarrow \{0,1\} \tag{8}$$

The utilized relations need to be meaningfully combined, so as to provide a view of the knowledge that suffices for context definition and estimation. Since modelling of real-life information is usually governed by uncertainty and ambiguity, it is our belief that these relations must incorporate fuzziness in their definition. The constructed ontology may be described by the "fuzzified" version of the concept ontology (eq. 9), where C represents again the set of all possible concepts, $F(R_{c_i,c_j}) = r_{c_i,c_j} : C \times C \rightarrow [0,1]$ denotes a fuzzy ontological relation amongst two concepts c_i, c_j and R_{c_i,c_j} denotes the non-fuzzy semantic relation amongst the two concepts. The final combination of the MPEG-7 originating relations forms an RDF graph and constitutes the abstract contextual knowledge model to be used (Fig. 2).

$$\mathcal{O}_C^f = \left\{C, r_{c_i,c_j}\right\}, \quad i,j = 1 \ldots n, \quad r_{c_i,c_j} : T \times T \rightarrow [0,1] \tag{9}$$

Herein, $r \in \mathcal{R}$ denotes a fuzzy ontological relation and

$$\mathcal{R} = \{Sp, P, Ex, Ins, Loc, Pat, Pr\} \tag{10}$$

denotes the set of all available relations. A meaningful combination of relations is described by:

$$\mathcal{C}_{ij} = (\underset{r \in \mathcal{R}}{\cup} r_{c_i,c_j}{}^{p_{r,ij}}), \quad p_{r,ij} \in \{-1,0,1\}, \quad i = 1 \ldots n \tag{11}$$

The value of $p_{r,ij}$ is determined by the semantics of each relation R_{c_i,c_j} used in the construction of $\mathcal{C}_{ij}$. We remind that:

- $p_{r,ij} = 1$, if the semantics of r_{t_i,t_j} imply it should be considered as is
- $p_{r,ij} = -1$, if the semantics of r_{t_i,t_j} imply its inverse should be considered
- $p_{r,ij} = 0$, if the semantics of r_{c_i,c_j} do not allow its participation in the construction of the combined relation $\mathcal{C}_T$.

Table 1. Fuzzy semantic relations between concepts.

Name	Inverse	Symbol	Meaning
Specialization	Generalization	$Sp(a,b)$	b is a specialization in the meaning of a
Part	PartOf	$P(a,b)$	b is a part of a
Example	ExampleOf	$Ex(a,b)$	b is an example of a
Instrument	InstrumentOf	$Ins(a,b)$	b is an instrument of or is employed by a
Location	LocationOf	$Loc(a,b)$	b is the location of a
Patient	PatientOf	$Pat(a,b)$	b is affected by or undergoes the action of a
Property	PropertyOf	$Pr(a,b)$	b is a property of a

The final ontology that occurs after the combination of the aforementioned relations is denoted by:

$$\mathcal{O}_C^{\mathcal{C}} = \{C, \mathcal{C}_{ij}\}, \quad i,j = 1,\ldots,n \quad i \neq j \tag{12}$$

The graph of the proposed model contains nodes (i.e. domain concepts) and edges (i.e. an appropriate combination[1] of contextual fuzzy relations between concepts). The degree of confidence of each edge represents fuzziness in the model. Non-existing edges imply non-existing relations (i.e. relations with zero confidence values are omitted). An existing edge between a given pair of concepts is produced based on the set of contextual fuzzy relations that are meaningful for the particular pair. Each concept has a different probability to appear in the scene, thus a flat context model would not have been sufficient in this case; on the contrary, concepts are related to each other, implying that the graph relations used are in fact transitive. The degree of confidence is implemented using the RDF reification technique [21].

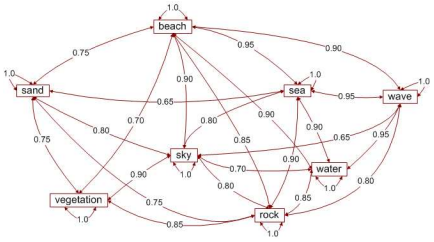

Figure 2. A fragment of the beach domain ontology. Concept $beach$ is the "root" element.

[1] The combination of different contextual fuzzy relations towards the generation of a practically exploitable knowledge view is conducted by utilizing fuzzy algebra's operations, in general and the default t-norm, in particular.

3.1 Contextualization effect

Once the contextual knowledge structure is finalized and the corresponding representation is implemented, a variation of the context-based confidence value readjustment algorithm [7] is applied on the output of the neural network-based classifier. The proposed contextualization approach empowers a post-processing step on top of the initial set of mid-level region types extracted. It provides an optimized re-estimation of the initial concepts' degrees of confidence for each region type and updates each model vector. In the process, it utilizes the high-level contextual knowledge from the constructed contextual ontology.

An estimation of each concept's degree of membership is derived from direct and indirect relationships of the concept with other concepts, using a meaningful compatibility indicator or distance metric. Again, depending on the nature of the domains provided in the domain ontology, the best indicator could be selected using the max or the min operator, respectively. The general structure of the degree of membership re-evaluation algorithm is now as follows:

1. The considered domain imposes the use of a domain (dis-)similarity measure: $\mu \in [0,1]$.

2. For each region type t consider a fuzzy set L_t with a degree of membership $\mu_t(c)$, containing the possible concepts' degrees of confidence.

3. For each concept c_i in the fuzzy set L_t with a degree of membership $\mu_t(c_i)$, obtain the particular contextual information in the form of its relations to the set of any other concepts:
$\{r_{c_i,c_j} : c_i, c_j \in C, \quad i \neq j\}$.

4. Calculate the new degree of membership $\mu_t(c)$, taking into account each domain's similarity measure. In the case of multiple concept relations in the ontology, when relating concept c to more than

the *root* concept (Fig. 2), an intermediate aggregation step should be applied for the estimation of $\mu_t(c)$ by considering the *context relevance* notion, cr_c: $cr_c = \max\{r_{c,c_1}, \ldots, r_{c,c_k}\}$, $c_1 \ldots c_k \in C$. We express the calculation of $\mu_t(c)$ with the recursive formula:

$$\mu_t^n(c) = \mu_t^{n-1}(c) - \mu(\mu_t^{n-1}(c) - cr_c) \quad (13)$$

where n denotes the iteration used. Equivalently, for an arbitrary iteration n:

$$\mu_t^n(c) = (1-\mu)^n \cdot \mu_t^0(c) + (1-(1-\mu)^n) \cdot cr_c \quad (14)$$

where $\mu_t^0(c)$ represents the initial degree of membership for concept c.

4 Experimental Results

In this section we provide some experimental results facilitating the conceptual approach and we shall try to demonstrate the usefulness of the visual context algorithm when applied to real-life multimedia content problems and data. We carried out a set of experiments utilizing a set of 1435 images, 25 region types and 6 high-level concepts {*sea*, *wave*, *sky*, *sand*, *rock*, *vegetation*} derived from the *beach* domain. All images were acquired from our personal image collections and the Internet. We further utilized a clustering training set of 300 images (i.e. merely 21% of the dataset) and selected $\mu = 0.125$ as the best normalization parameter for the considered domain of interest. Typically, a number of $n = 3$ iterations were used.

Evaluation results for 6 high-level *beach* concepts are presented in Table 2. Each concept's row displays the precision and recall values *before* and *after* the use of context. Observing the results it is rather obvious that the proposed contextualization algorithm exploits semantic relations in order to favor or disfavor the degrees of confidence for the detection of a concept that exists within an image. Thus, it strengthens the concepts' differences, but at the same time it treats smoothly the confidence values of almost certain concepts (e.g. *sea*). Finally, exploiting the constructed ontological knowledge, the algorithm is able to disambiguate cases of similar concepts or even concepts being difficult to be detected solely based on traditional low-level analysis steps.

5 Conclusions

Our research effort indicates clearly that high-level concepts can be efficiently detected when an image is represented by a mid-level model vector with the aid of a visual thesaurus and additional contextual knowledge. Amongst the core contribution of this work has been the implementation of a novel mid-level visual context interpretation utilizing a fuzzy, ontology-based representation of knowledge. Experimental research results were presented, indicating a significant high-level concept detection optimization (i.e. precision improvement per concept varies from 6.98% to 25.86%) over the entire dataset utilized.

6 Acknowledgements

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

References

[1] N. Boujemaa, F. Fleuret, V.G., 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, part of Electronic Imaging symposium, 2004.

[2] M. Boutell and J. Luo, *Bayesian fusion of camera metadata cues in semantic scene classification*, in Proc. IEEE Conf. Computer Vision Pattern Recognition (CVPR), Washington, DC, 2004, vol. 2, pp. 623–630.

[3] M. Boutell, J. Luo, C.M. Brown, *A generalized temporal context model for classifying image collections*, ACM Multimedia Syst. J., 11(1), pp. 82–92, Nov. 2005.

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

[5] IBM: (Marvel: Multimedia analysis and retrieval system)

[6] A. Mathes, *Folksonomies - Cooperative Classification and Communication Through Shared Metadata*, Computer Mediated Communication - LIS590CMC, Graduate School of Library and Information Science, University of Illinois Urbana-Champaign, 2004.

[7] Ph. Mylonas, Th. Athanasiadis and Y. Avrithis, *Improving image analysis using a contextual approach*, In Proc. of 7th International Workshop on

Table 2. Overall precision/recall scores before and after the use of visual context per high-level *beach* concept.

Concepts	Precision			Recall		
	before	after	%	before	after	%
sea	0.86	0.92	6.98%	0.66	0.58	-12.12%
wave	0.89	0.96	7.87%	0.68	0.63	-7.35%
sky	0.79	0.91	15.19%	0.55	0.48	-12.73%
sand	0.72	0.83	15.28%	0.45	0.39	-13.33%
rock	0.58	0.73	25.86%	0.32	0.27	-15.63%
vegetation	0.38	0.42	10.53%	0.33	0.28	-15.15%
Total	**0.70**	**0.80**	**13.03%**	**0.50**	**0.44**	**-12.04%**

Image Analysis for Multimedia Interactive Services (WIAMIS), Seoul, Korea, 2006.

[8] Ph. Mylonas and Y. Avrithis, *Context modelling for multimedia analysis*, In Proc. of 5th International and Interdisciplinary Conference on Modeling and Using Context (CONTEXT ŏ05), Paris, France, 2005.

[9] 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.

[10] Ph. Mylonas, M. Wallace, S. Kollias, *Using k-nearest neighbor and feature selection as an improvement to hierarchical clustering* Proceedings of 3rd Hellenic Conference on Artificial Intelligence, Samos, Greece, May 2004.

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

[12] Smeaton, A. F., Over, P., and Kraaij, W., *Evaluation campaigns and TRECVid*, In Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval, Santa Barbara, California, USA, October 26 - 27, 2006.

[13] Smeulders, A. W. M., Worring, M., Santini, S., Gupta, A. and Jain, R.: Content-Based Image Retrieval at the End of the Early Years, IEEE Trans. on Pattern Analysis and Machine Intelligence, **22**: 1349–1380 (2000)

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

[15] E. Spyrou and Y. Avrithis, *A Region Thesaurus Approach for High-Level Concept Detection in the Natural Disaster Domain*, In Proc. of the 2nd International Conference on Semantics And digital Media Technologies (SAMT), 2007.

[16] E. Spyrou, G. Tolias, Ph. Mylonas and Y. Avrithis, *A Semantic Multimedia Analysis Approach Utilizing a Region Thesaurus and LSA*, 9th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS 2008).

[17] Staab, S., and Studer, R., *Handbook on Ontologies*, International Handbooks on Information Systems, Springer-Verlag, Heidelberg, 2004.

[18] A. Torralba, *Contextual influences on saliency*, Neurobiology of attention, Ac. Press, London, 2005.

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

[20] J. Yuan, J. Li and B. Zhang, *Exploiting spatial context constraints for automatic image region annotation*, In Proc. of the 15th international conf. on Multimedia, pp. 595–604, Augsburg, Germany, 2007.

[21] W3C, *RDF Reification*, http://www.w3.org/TR/rdf-schema/#ch_reificationvocab