USING REGION SEMANTICS AND VISUAL CONTEXT FOR SCENE CLASSIFICATION

Evaggelos Spyrou, Phivos Mylonas and Yannis Avrithis

Image, Video and Multimedia Laboratory,
National Technical University of Athens
Zographou Campus, PC 15773, Athens, Greece
{espyrou, fmylonas, iavr}@image.ntua.gr

ABSTRACT

In this paper we focus on scene classification and detection of high-level concepts within multimedia documents, by introducing an intermediate contextual approach as a means of exploiting the visual context of images. More specifically, we introduce and model a novel relational knowledge representation, founded on topological and semantic relations between the concepts of an image. We further develop an algorithm to address computationally efficient handling of visual context and extraction of mid-level region characteristics. Based on the proposed knowledge model, we combine the notion of visual context with region semantics, in order to exploit their efficacy in dealing with scene classification problems. Finally, initial experimental results are presented, in order to demonstrate possible applications of the proposed methodology.

Index Terms— scene classification, concept detection, visual context, region semantics

1. INTRODUCTION

Visual context [7] forms a rather classical approach to context, tackling it from the scope of environmental or physical parameters that are evident in multimedia applications. The discussed knowledge representation supports all visual information inherent in images. Our research objective deals with the, so called, visual context analysis, i.e. the way to take into account the extracted/recognized concepts during content analysis in order to identify the specific context, express it in a structural description form, and use it for improving or continuing the content analysis, indexing and searching procedures. In the following, we shall refer to the term *visual context*, by interpreting it as *all information related to the visual scene content of a still image that may be useful during its analysis phase*.

Visual context is strongly related to *scene classification*, one of the main problems of image analysis. Scene classification forms a *top-down* approach, where typically low-level visual features are employed to globally analyze the scene content and classify it in one of a number of predefined categories (e.g. indoor/outdoor, city/landscape). For instance, detection of a *green* region below an *azure* region in an image might imply an *outdoor field* scenery, or taking this a step further, detection of a *building* in the middle of an image might imply a *city* scene with higher probability.

An important step towards semantically analyzed and optimized results is to automate the process of semantic feature extraction and annotation of multimedia content objects, by enhancing image classification with semantic characteristics. Utilizing semantics in the form of detection of semantic features in still images or video sequences has been the ultimate task in earlier and current multimedia research efforts ([5], [12], [3]). Many approaches have been proposed, all sharing the common target and finally extracting high-level concepts from raw content. For instance, in [6], a multi-modal machine learning technique is used in order to model semantic concepts within video sequences. Region-based research approaches in content-based retrieval, like the one presented in [9] and which uses Latent Semantic Analysis (LSA), are common in the field. Moreover, a lexicon-driven approach is introduced in [4]. Finally, a mean-shift algorithm is used in [8], in order to extract low-level concepts, after the image is clustered.

So far and to the best of our knowledge, none of the current research efforts utilizes the herein proposed context in any form. This tends to be the main drawback of individual object and scene detectors, since they only examine isolated strips of pure object materials, without taking into consideration the context of the scene or individual objects themselves. This remark is very important and at the same time extremely challenging even for human observers. The notion of visual context is able to aid in the direction of scene classification methodologies, simulating the human approach to similar problems.

The structure of this paper is as follows: In Section 2, we present the proposed novel fuzzy knowledge representation, including some basic notation used throughout the paper. Section 3 is dedicated to the proposed contextual adaptation in terms of visual context algorithm optimization steps. Section 4 lists some preliminary experimental results derived from two different *beach* datasets and Section 5 concludes

briefly our work.

2. KNOWLEDGE REPRESENTATION

In principle, any kind of relation (semantic, topological, temporal, spatiotemporal) may be represented by an ontology [11]. However, herein we restrict the utilized relations' types to topological and semantic ones, as the latter are the most suitable relations to describe multimedia content. Based on these kinds of relations, we introduce a novel knowledge representation, package it in the form of a contextual ontology and utilize it, in order to semantically enhance the multimedia analysis process. The aforementioned relations are introduced in order to express in an optimal way the real-world relationships that exist between each image's participating concepts. In order for this ontology type to be highly descriptive, it must contain a representative number of distinct and even diverse relations among concepts, so as to scatter information among them and thus describe their context in a rather meaningful way. The utilized relations need to be meaningfully defined and even combined, so as to provide a view of the knowledge that suffices for context definition and estimation.

Additionally, 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. In the following we incorporate a rather classical subset of topological: { *adjacent*, *inside*, *above*, *below*, *left*, *right*}, as well as semantic: {*similarity*, *part*, *specification*} and *co-occurrence* relations among high-level concepts. While the notion of the first set of topological relations is straightforward, the second set of semantic relations is derived from the MPEG-7 semantic relations set [2], suitable for image analysis. We define both types of relations in a way to exploit and represent fuzziness, i.e. a *degree of confidence* is associated to each relation.

2.1. Relations among high-level concepts

To begin, we define some fundamental sets, necessary for the definition of more specific sets and relations. More specifically, let $\mathcal{C}$ be the set of all high-level concepts, $\mathcal{P}$ be the set of all images of the training set and $\mathcal{S}$ be the set of all regions of all images. Within each image p, we define:

- $\mathcal{C} = \{c_p\}, \quad p = 1, 2, \ldots, N_c$ be the set of all high-level concepts within the domain(s) of interest. The high-level concepts are determined by a domain expert. An applicable subset of all these possible concepts is selected from the ontology user/developer.
- $\mathcal{S} = \{s_p\}, \quad p = 1, 2, \ldots, N_s$ be the set of all regions (segments), of all images, as extracted by a specific segmentation tool.
- $C_p = \{c_k^p\},\ k = 1, 2, \ldots, N_c^p,\ p \in \mathcal{P}$, be the set of all high-level concepts present in image p. As obvious, $C_p \subset \mathcal{C}$. C_p is determined by the provided annotation for the training set of images.
- $D_p = \{d_k^p\},\ k = 1, 2, \ldots, N_d^p,\ p \in \mathcal{P}$, be the set of all initial detector values of image p. The initial detector values of an image result from the application of appropriate high-level feature detectors.

Letting $R_1(e_1, e_2)$ be a binary relation between concepts c_1 and c_2 and $R_2(c_1, c_2)$ be the opposite relation (e.g. "above" is the opposite relation of "below"), we define the *inverse* relation as: $\mathbf{R^{-1}}$: $R_1^{-1}(c_1, c_2) = R_1(c_2, c_1)$ and the *opposite* relation as: $\neg\mathbf{R}$: $\neg R_1(c_1, c_2) = R_2(c_1, c_2)$. The cardinality of a set is denoted by $|\cdot|$.

2.1.1. Semantic relations

In order to acquire a meaningful set of semantic relations suitable for image analysis problems, we extend a subset of the well-known MPEG-7 semantic relations [2] and re-define them in a way to represent fuzziness, i.e. a degree of confidence is associated to each relation. Let sem denote any considered semantic relation between any given pair of concepts defined by an expert:

$$R_{cc}^{sem} = \{r_{c_1,c_2}^{sem}\}, \quad r_{c_1,c_2}^{sem} = sem(c_1, c_2),\ c_1, c_2 \in \mathcal{C} \quad (1)$$

As already indicated, sem belongs to one of either three possible semantic relation types defined within the MPEG-7 standard, namely: $sem \in \{sim, part, spec\}$. The first one, sim, denotes the semantic *Similarity* amongst any pair of concepts (e.g. *automobile*/*car*), $part$ is the MPEG-7 *PartOf* semantic relation (e.g. *Sydney*/*Australia*) and $spec$ is the *Specialization* relation between high-level concepts c_1 and c_2 (e.g. *cow*/*animal*).

2.1.2. Topological relations

Apart from the above presented semantic relations, in order to fine-tune the analysis process, we also define utilize a set of fuzzy topological relations. Thus, let top denote any topological relation between any given concepts c_1 and c_2:

$$R_{cc}^{top} = \{r_{c_1,c_2}^{top}\} = \{top(c_1, c_2)\},\ c_1, c_2 \in \mathcal{C} \quad (2)$$

where $top \in \{adj, ins, ab, bel, left, rgt\}$. Each one of these six relations is explained in the Table 1:

Finally, a very important relation to be taken into consideration is the *co-occurence* relation, which is defined statistically on the training set data. We define:

$$R_{cc}^{co} = \{r_{c_1,c_2}^{co}\} = \{co(c_1, c_2)\},\ c_1, c_2 \in \mathcal{C} \quad (3)$$

where:

$$co(c_1, c_2) = \frac{|\{p \in \mathcal{P} : c_1 \in C_p \wedge c_2 \in C_p\}|}{|\{p \in \mathcal{P} : c_1 \in C_p \vee c_2 \in C_p\}|} \quad (4)$$

Table 1. Proposed fuzzy topological relations.

Name	Symbol	Properties
adjacent	adj	$adj\,(c_1, c_2) = adj^{-1}\,(c_1, c_2)$
inside	ins	$ins\,(c_1, c_2) \neq ins^{-1}\,(c_1, c_2)$
above	ab	$ab\,(c_1, c_2) \neq ab^{-1}\,(c_1, c_2)$ $\neg ab\,(c_1, c_2) = bel\,(c_1, c_2)$
below	bel	$bel\,(c_1, c_2) \neq bel^{-1}\,(c_1, c_2)$ $\neg bel\,(c_1, c_2) = ab\,(c_1, c_2)$
left	$left$	$left\,(c_1, c_2) \neq left^{-1}\,(c_1, c_2)$ $\neg left\,(c_1, c_2) = rgt\,(c_1, c_2)$
above	rgt	$rgt\,(c_1, c_2) \neq rgt^{-1}\,(c_1, c_2)$ $\neg rgt\,(c_1, c_2) = left\,(c_1, c_2)$

2.1.3. Knowledge formalization

All the above knowledge may be integrated into a single, "fuzzified" version of an ontology described by $\mathcal{O}$:

$$\mathcal{O} = \left\{\mathcal{C}, \mathcal{P}, \mathcal{S}, \mathcal{R}_{c_i,c_j}\right\}, \quad i, j = 1, \ldots m, \quad i \neq j \tag{5}$$

where $\mathcal{C}$ represents the set of all high-level concepts, $\mathcal{P}$ be the set of all images of the training set and $\mathcal{S}$ be the set of all regions of all images and

$$\mathcal{R}_{c_i,c_j} = F(R_{c_i,c_j}) = \{R_{cc}^{sem}, R_{cc}^{top}, R_{cc}^{co}\} \tag{6}$$

denotes a fuzzy relation amongst two concepts c_i, c_j.

A meaningful combination of these relations

$$\mathcal{Z}_{c_i,c_j} = (\underset{i,j}{\cup} \mathcal{R}_{c_i,c_j}{}^{p_{ij}}), \quad p_{ij} \in \{-1, 0, 1\} \tag{7}$$

may then be used to form the abstract *contextual knowledge model* formed herein and ready to be used during the analysis phase. The value of p_{ij} is determined by the semantics of each relation $\mathcal{R}_{c_i,c_j}$ used in the construction of $\mathcal{Z}_{c_i,c_j}$. More specifically:

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

3. EMPLOYING VISUAL CONTEXT

In the following, we introduce an entity-based methodology, founded on the knowledge representation presented in the previous Section 2. We utilize a set of semantic concepts of the image, as well as the set of the fuzzy relations $\mathcal{Z}$ between them. The core functionality of the visual context algorithm is the meaningful interpretation of the initial concept detectors' confidence values $d_{c_i} \in D_p$ associated to a region $s_p \in \mathcal{S}$ of an image. These initial values may be obtained from any kind of image segmentation and analysis module. The novelty introduced herein deals with the extraction of the contextualized value of each detector f_{c_i}, based on the fuzzy contextual relationships $\mathcal{Z}_{c_i,c_j} \in [0, 1]$ evident in the image and the domain of interest. In other words, the concept's context refers to the overall relevance of each concept to the related domain, as well as the rest of the concepts that are present in the image under consideration.

The general structure of the proposed algorithm for the simplified case of one image (i.e. we omit the ^{p} index) is as follows.

1. For each available concept $c_i \in C_p, \quad i = 1 \ldots N_c$, obtain its initial concept detector value $d_{c_i} \in D_p$. N_c is the cardinality of set C.
2. For each concept c_i, obtain its fuzzy relationships $\mathcal{Z}_{c_i,c_j} \in [0, 1]$ to any other concept c_j in the knowledge, where $j = 1 \ldots N_c$ and $\mathcal{Z}_{c_i,c_i} = 1$.
3. Calculate the contextualized concept detector value f_{c_i} of each concept c_i, using the function:

   $$f_{c_i} \quad = \quad (d_{c_i}, t_{c_i})$$

   where:

   $$t_{c_i} \quad = \quad \frac{\sum_{c_i \neq c_j} d_{c_j} \cdot \mathcal{Z}_{c_i,c_j}}{\sum_{c_i \neq c_j} d_{c_j}}$$

   It is worth noting that this function incorporates both globally available (d_{c_i}) and locally available (i.e. within the particular image under consideration) (t_{c_i}) information in the calculation of f_{c_i}. In case the confidence of the initial concept detector is extremely high or low, contextualization of the value is not taken into consideration, i.e. if $d_{c_i} = \{0, 1\} \Rightarrow f_{c_i} = d_{c_i}$, whereas for $d_{c_i} = 0.5 \Rightarrow f_{c_i} = t_{c_i}$.
4. We use the linear combination:

   $$f_{c_i} \quad = \quad h(d_{c_i}) \cdot t_{c_i} + (1 - h(d_{c_i})) \cdot d_{c_i}$$

   to express the calculation of the new value f_{c_i}, where $h(d_{c_i})$ represents the triangular function

   $$\begin{aligned} d_{c_i} \quad &= \quad tri(x) \\ &= \quad \begin{cases} 1 - |x|; & 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

The above algorithm extracts the underlying contextual knowledge and estimates the contextualized value of the concept detectors f_{c_i}, based on the fuzzy contextual relationships $\mathcal{Z}_{c_i,c_j}$ and the initial concept detectors' confidence values d_{c_i}, associated to the regions $s_p \in \mathcal{S}$ of the image.

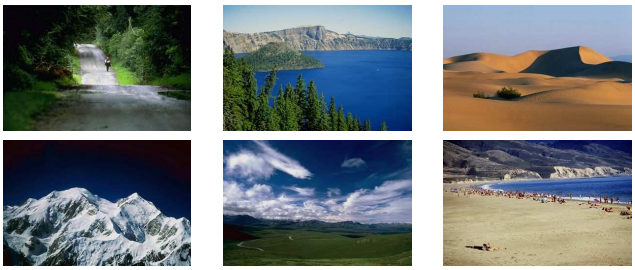

Fig. 1. Representative Images

4. RESULTS

In this section we provide experimental results of the proposed algorithm. We carried out experiments utilizing 750 images and 6 high-level concepts, acquired from the Corel image collection. A small sample of this set is depicted on Fig. 1. For the initial (i.e., without contextual knowledge) detection of the high-level concepts, the approach of [10] was applied. We utilized 525 images to train and 225 images to test the 6 separate concept detectors.

In Table 2, we present both the initial detection results and those refined by the contextual approach presented herein. We may observe that the precision was improved for all 6 concepts. We should note that concepts *road* and *sand* had a smaller number of positive examples than the others, thus it was rather difficult to train reliable detectors. However, contextual knowledge also improved those "weak" results.

Table 2. Precision/Recall per concept **before** and **after** the application of the visual context algorithm.

	Before		**After**		**Difference**	
Concepts	**P**	**R**	**P**	**R**	**P**	**R**
road	0.22	0.25	0.43	0.21	+95%	-16%
sand	0.38	0.33	0.55	0.28	+45%	-15%
sea	0.78	0.71	0.89	0.68	+14%	-4%
sky	0.81	0.72	0.91	0.67	+12%	-7%
snow	0.48	0.58	0.72	0.45	+50%	-22%
vegetation	0.74	0.62	0.87	0.54	+18%	-13%
Overall	**0.57**	**0.54**	**0.73**	**0.47**	**+28%**	**-13%**

5. CONCLUSIONS

This paper presented an approach towards more efficient high-level detection in images. Its contributions are the set of relations that are defined and the visual context algorithm that refined the initial detection results. It is shown that the existing concept relations improve the precision of the results not only for already well-trained and effective detectors, but also for weak and non-effective. Future work will concentrate on the definition of similar relations for the image regions and region types from which an image is consisted of, and the contextualized inter-relations between different semantic entities.

6. ACKNOWLEDGEMENT

The research leading to this paper was partially supported by the European Commission's 6th and 7th Framework Programmes FP6/2002-2006 and FP7/2007-2013, under grants Nr. 027685 - MESH and Nr. 215453 - WeKnowIt.

7. REFERENCES

[1] Th. Athanasiadis, Ph. Mylonas, Y. Avrithis and S. Kollias, *Semantic Image Segmentation and Object Labeling*, IEEE Trans. on Circuits and Systems for Video Technology, 17(3), 298-312, March 2007.

[2] A. B. Benitez, D. Zhong, S.-F. Chang and J. R. Smith, *MPEG-7 MDS Content Description Tools and Applications*, LNCS, vol. 2124, pp. 41-52, 2001.

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

[4] D. C. K. Cees, G. M. Snoek, M. Worring and A. W. Smeulders, *Learned lexicon-driven interactive video retrieval*, 5th Int. Conference on Image and Video Retrieval (CIVR), Tempe, Arizona, USA, July 2006.

[5] J.M. Henderson, A. Hollingworth,. *High level scene perception*, Annu. Rev. Psychol., vol. 50, pp. 243-271, 1999

[6] IBM, *Marvel: Multimedia analysis and retrieval system*, `http://mp7.watson.ibm.com/`

[7] Ph. Mylonas and Y. Avrithis, *Context modelling for multimedia analysis*, 5th International and Interdisciplinary Conference on Modeling and Using Context (CONTEXT '05), Paris, France, July 2005.

[8] B. Saux and G.Amato, *Image classifiers for scene analysis*, Int. Conference on Computer Vision and Graphics (ICCVG), Warsaw, Poland, September 2004.

[9] F. Souvannavong, B. Merialdo and B. Huet, *Region-based video content indexing and retrieval*, 4th International Workshop on Content-Based Multimedia Indexing (CBMI), Riga, Latvia, June 2005.

[10] 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.

[11] S. Staab and R. Studer, *Handbook on ontologies*, Springer Series on Handbooks in Information Systems, Heidelberg, Springer-Verlag, 2004.

[12] A. Torralba, *Contextual influences on saliency*, *Neurobiology of attention*, Eds. L. Itti, G. Rees, J. Tsotsos, Academic Press Inc. (London) Ltd, 2005