Context modeling for multimedia analysis[1]

Phivos Mylonas, Yannis Avrithis

Image, Video and Multimedia Laboratory,
National Technical University of Athens
15773 Athens, Greece
{fmylonas, iavr}@image.ntua.gr
http://www.image.ntua.gr/~fmylonas

Abstract. Context is of great importance in a wide range of computing applications and has become a major topic in multimedia content search and retrieval systems. In this paper we focus our research efforts on visual context, a part of context suitable for multimedia analysis and usage. We introduce our efforts towards the scope of clarifying context in the fields of object detection and scene classification during multimedia analysis. We also present a method for visual context modelling, based on spatial object and region-based relations, to use in content-based multimedia search and retrieval systems.

1 Introduction

Unquestionably, the term context can take on many meanings and there is no definition that is felt to be satisfactory; the term has a long history in artificial intelligence, information retrieval and image and video analysis [1]. The use of context is especially important in applications dominated by rapid changes in the user's context, such as handheld and ubiquitous computing [4]. Researchers commonly emphasize distinctions between different types of context and illustrate how little each type has to do with the others [2].

This paper provides an integrated view on the contextual aspect exploited within multimedia systems and applications, namely the aspect of context summarized in the term *visual context*. Its efforts are directed to the fields of scene classification and object detection in multimedia analysis in the framework of aceMedia [3], which focuses on knowledge discovery embedded into media content. The notion of visual context is used as the vehicle towards the achievement of this goal, as it forms the basic framework used for the next steps of our work in multimedia analysis, namely segmentation, object detection and scene classification. In particular, in section 2, a definition for visual context utilized within the scope of multimedia content-based systems is provided, as well as novel ideas are presented regarding visual context exploitation in multimedia analysis. Section 3 tackles visual context modelling issues, whereas conclusions and future work initiatives are drawn in section 4.

[1] This work was supported by the European Commission under contract FP6-001765 aceMedia (URL: http://www.acemedia.org).

2 Visual context in multimedia analysis

Visual context forms a rather classical approach to context, tackling it from the scope of environmental or physical parameters in multimedia applications. The discussed context representation supports audiovisual information (e.g. lighting conditions, environmental information, etc) and is separately handled by visual context models. Research objectives in the field include visual context analysis, i.e. to take into account the extracted/recognized concepts during content analysis in order to find the specific context, express it in a structural description form, and use it for improving or continuing the content analysis, indexing and searching procedures, as well as personalization aspects.

By visual context in the sequel we will refer to all information related to the visual scene content of a still image or video sequence that may be useful for its analysis. Visual context is related to two problems in image analysis. *Scene classification,* which forms a top-down approach where low-level visual features are employed to globally analyse the scene content and classify it in one of a number of pre-defined categories, e.g. indoor/outdoor, city/landscape and on the other hand, *object detection/recognition*, which is a bottom-up approach that focuses on local analysis to detect and recognise specific objects in limited regions of an image, without explicit knowledge of the surrounding context, e.g. recognise a building or a tree. The above two major fields of image analysis actually comprise a chicken-and-egg problem, as, for instance, detection of a building in the middle of an image might imply a picture of a city with a high probability, whereas pre-classification of the picture as "city" would favour the recognition of a building vs. a tree.

In *content-based image search and retrieval*, more and more researchers are looking beyond low-level colour, texture, and shape features in pursuit of more effective searching methods. Natural object detection in indoor or outdoor scenes, i.e. identifying key object types such as sky, grass, foliage, water and snow, can facilitate content-based applications, ranging from image enhancement to coding or other multimedia applications. However, a significant number of misclassifications usually occur because of the similarities in colour and texture characteristics of various object types and the lack of context information, which is a major limitation of individual object detectors.

So far, none of the above methods and techniques utilize context in any form, which tends to be their main drawback, since they only examine isolated strips of pure object materials, without taking into consideration the context of the scene or individual objects. The notion of visual context is able to aid in the direction of natural object detection methodologies, simulating the human approach to similar problems. Many object materials can have the same appearance in terms of colour and texture, while the same object may have different appearances under different imaging conditions (e.g. lighting, magnification). However, one important trait of humans is that they examine all the objects in the scene before making a final decision on the identity of individual objects. The use of visual context forms the key for this unambiguous recognition process, as it refers to the relationships among the location of different objects in the scene. It may be either spatial or temporal; *spatial context* is associated to spatial relationships between objects or regions in a still image or video sequence, while *temporal context* to temporal relationships between objects, regions or scenes in video sequences. In the sequel, discussion will be restricted to spatial context analysis.

3 Visual context modelling

Focusing our efforts in providing a robust context model capable of handling both local and global information in image analysis, resulted in the ascertainment that the only way to achieve this is to model the relationships between the information and not the information themselves, with respect to the level of the details present in each relationship. In this manner, at least two types of meaningful visual (spatial) contextual relationships are identified in natural images. First, relationships exist between *co-occurrence* of certain objects in natural images. For example, detection of snow with high probability would imply low grass probability. Second, relationships exist between *spatial locations* of certain objects within an image: grass tends to occur below sky, sky above snow, etc. The ultimate goal is to develop a non-scene specific method for generating spatial context models useful for general scene understanding problems. Subsequently, spatial context constraints are used to reduce the number of false positives by constraining the initial beliefs to conform to the spatial context models.

In general, spatial context modelling refers to the process of building relationship models that define the spatial arrangement and distribution of the objects of interest in a scene. Depending on the requirements of the application, the set of spatial relationships can be rich (many spatial relationships with minor differences between each) or sparse (fewer distinct relationships). The spatial relations define the absolute or relative spatial information between objects. Various spatial arrangements for two regions can be defined and the mapping of these spatial arrangements to semantic spatial relationships can also be constructed. For example, a rather complete set of spatial relationships in such an application can be modeled as: *above*, *far_above*, *below*, *far_below*, *beside*, *enclosed*, *enclosing*. Suitable thresholds are used to discriminate between *above/below* and *far_above/far_below*. Consequently, numerous relationships between two objects can be defined, such as: *Connectivity*, *Position*, *Depth*, *Partonomic, Size* and *Shape relations* .

In the case of scene classification, for instance, where information is not available in the form of objects, but in the form of regions, a top-down technique is necessary. Towards fulfilling the ultimate goal of this task, i.e. , classification of images or video sequences based on their content, contextual information can be taken into advantage in the form of the spatial layout of regions in an image. For example (**Figure 1**), a class of images representing a sunny beach seaside may be described as having three perceptually salient regions: (i) a blue region representing the sea, (ii) a yellow region representing the sand and (iii) a lighter blue region representing the cloudless sky. In all cases, regions (i) and (ii) are always below region (iii), whereas in the first four, region (ii) lies below region (i) and in the fifth image, region (i) lies below region (ii); regions (i) and (ii) may exchange their spatial positions in some shots, according to the perspective used. The above example suggests that the desirable classification of a scene may remain valid as long as the relative spatial contextual interregional relationships between the image regions remain the same, even though absolute region values may change. Again, numerous relationships between two regions in the scene can be defined, in the same manner as between objects.

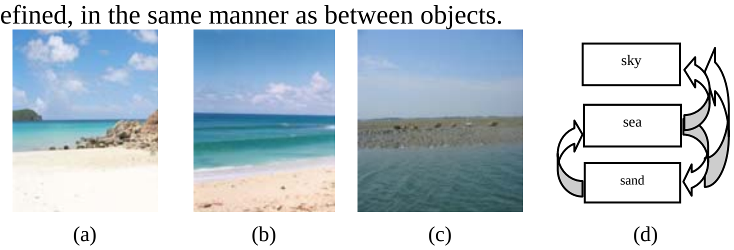

Figure 1. (a) – (c) Examples of images representing a sunny beach. (d) Interregional spatial-contextual relationships

4 Conclusions and future work

This work introduced a novel type of context and context model suitable for use in image analysis and retrieval in the form of visual context. Specifically, it has introduced an approach dealing with visual context in the task of knowledge-assisted image and video analysis, adopted for use within the aceMedia system. Its efforts concluded that visual context information significantly aids in knowledge extraction, when handling scene classification and object detection problems. The latter are also gaining benefit from available visual contextual information, in order to provide information about indoor/outdoor and city/landscape scenery problems.

Finally, the herein presented effort forms a small piece of work at the beginning of a research on knowledge-assisted image/video analysis. It places itself in the process, as it relates to object identification and image classification and will be utilized in the form of driving the analysis process of our work by selecting suitable algorithms, detectors and classifiers. Future work will include all above mentioned issues, along with experimental results indicating its benefits and contributing to the overall usage of context in multimedia analysis.

References

1. M. Davis, N. Good and R. Sarvas, "From Context to Content: Leveraging Context for Mobile Media Metadata", 2004.
2. B. Edmonds, "The Pragmatic Roots of Context", Proceedings of the 2nd International and Interdisciplinary Conference on Modeling and Using Context (CONTEXT-99). LNAI, vol. 1688, Berlin: Springer, 1999.
3. I. Kompatsiaris, Y. Avrithis, P. Hobson and M.G. Strinzis, "Integrating Knowledge, Semantics and Content for User-Centred Intelligent Media Services: the aceMedia Project", Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS '04), Lisboa, Portugal, April 21-23, 2004.
4. M. Weiser, "Some computer science issues in ubiquitous computing", Special Issue, Computer-Augmented Environments, CACM, 36(7):74–83, July 1993.