Understanding How Visual Context Influences Multimedia Content Analysis Problems

Phivos Mylonas

Image, Video and Multimedia Laboratory,
Department of Computer Science
School of Electrical and Computer Engineering
National Technical University of Athens
P.C. 15773, Athens, Greece
fmylonas@image.ntua.gr

Abstract. The importance of context in modern multimedia computing applications is widely acknowledged and has become a major topic of interest in multimedia content analysis systems. In this paper we focus on visual context, tackling it from the scope of its utilization within the above framework. We present a brief review of visual context modeling methods and identify and discriminate its useful types within multimedia applications, envisioning possible usage scenarios for contextual information. Finally, a representation of visual context modeling is reviewed, being suitable for aiding in the case of common multimedia analysis problems, such as object detection and scene classification.

Keywords: visual context, multimedia analysis, knowledge representation, context representation and analysis, context-driven multimedia analysis.

1 Introduction

The polysemy of the term *context* is widely acknowledged and currently there is no solid definition that covers its usage within most multimedia analysis efforts. In the field of computer science, the interest in contextual information is of great importance in fields like artificial intelligence, information search and retrieval, as well as image and video analysis [2]. Still, effective use of available contextual information within multimedia structures remains an open and challenging problem, although a categorization of context-aware applications according to subjective criteria has been tried out [10].

A fundamental problem tackled via access to and processing of contextual information is the bridging of two fundamental gaps in the literature; the *semantic* and *sensory gap* [4]. The *semantic gap*, an issue inherent in most multimedia applications, is described as the gap between high-level semantic descriptions humans ascribe to images and low-level features machines can automatically parse. The *sensory gap* is described as the gap between an object and the computer's ability to sense and describe this object. It is contextual knowledge that may enable computational systems to bridge both gaps. With the advent of all kind of new multimedia-enabled devices

I. Maglogiannis, V. Plagianakos, and I. Vlahavas (Eds.): SETN 2012, LNAI 7297, pp. 361–368, 2012.

and multimedia-based systems, new opportunities arise to infer media semantics and contextual metadata are capable of playing the important role of a "semantic mediator".

It is common knowledge, though, that context itself appears in various forms and modifications and researchers commonly emphasize distinctions between different types of context. This paper provides an overview on the definition of one basic aspect of context exploited within multimedia systems and applications, namely the aspect of context summarized in the term: *visual context*. It's efforts are directed towards the fields of scene classification and object detection in multimedia analysis, introducing envisioned usage scenarios in the area.

The rest of this paper is organized as follows: in Section 2, after underlining the importance of context identification, two useful types of context utilized within the scope of multimedia content-based systems are identified, namely *context of content analysis* and *context of use*. Section 3 deals with visual context in typical image analysis problems, such as scene classification and object detection/recognition. Section 4 addresses the problem of visual context modeling, whereas final comments on the topic and conclusions are drawn in Section 5.

2 Visual Context Identification

The task of suitable visual context definition and identification is very important, because all knowledge required for multimedia content analysis is thought to be context-sensitive, thus resulting in a specific need for formal definitions of context structures prior to any static or dynamic context analysis. The first objective formed within this task is the definition of the suitable aspect of context at hand, providing conceptual and audiovisual information. We may introduce two types of context: the *context of content analysis* and the *context of use*. *Context of content analysis* refers to the context during the initial content analysis phase. It is intended to be used to aid the extraction of semantic metadata both at the level of simple concepts and at the level of composite events and higher level concepts. For instance, during scene classification it is used to detect whether a picture or video clip represents *city* or *landscape* content, essentially aiding the analysis process. On the other hand, *context of use* is related to the use of content by search/retrieval and personalization applications. In this case, given the multimedia content and metadata, contextual information from external sources are utilized, consisting mainly of information about the particular user, network and client device.

In multimedia computing applications the aspects of context, which are thought to be the most suitable and appropriate for research and progress, are the ones described above. Therefore, from now on we shall present them under a common approach, summarized in the notion of *visual context*. *Visual context* forms a rather classical approach to context, tackling it from the scope of environmental or physical parameters in multimedia applications. Different architectures, conceptual approaches and models support dynamic and adaptive modeling of visual context. One of the main objectives in the field is the combination of context parameters extracted from low

level visual features with higher level concepts and interpretation (e.g. fuzzy set theory) to support additional knowledge processing tasks like reasoning. Specifically, it is wise for a context description to support fuzziness, in order for it to face the uncertainty introduced by content analysis or the lack of knowledge. Such a context representation also supports audiovisual information (e.g. lighting conditions, information about the environment, e.t.c.) and is separately handled by visual context models. The second objective is 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 for personalization purposes.

In terms of knowledge-assisted content analysis and processing, a set of core visual context functionalities of the multimedia application requires to be defined, regarding the way such a system is expected to execute knowledge-assisted image analysis functions automatically or in a supervised mode, so as to either detect or to recognize parts of content. Additionally, context is thought to generate or assist end-users classify their contents and metadata, through suggestions or sorting being performed in a sophisticated way, making quite naturally implicit use of its analysis functionalities. For example, in a face recognition scenario, visual clues may help the system detect the right person. Issues relating more to the automatic creation of metadata even after analysis, e.g. through inference, make use of context, as different sources of information (different analysis modules, textual inputs) may also be integrated.

As far as retrieval is concerned, a set of core visual context functionalities of a multimedia search and retrieval system need to be also defined; there are many distinct aspects suggested and commented by users, regarding the way of performing searches, the type of searches they expect to have and the constraints they imagine. Organizing multimedia data into meaningful categories marked by end-users as being important, could exploit contextual information. Retrieval is especially related to context, when tackling textual query analysis, search by semantic, visual or metadata similarity, semantic grouping, browsing and rendering of retrieved content, personalization and relevance feedback. However, user browsing capabilities, together with retrieval capabilities, suppose detection of common metadata, which is not considered to be related to the notion of visual context discussed herein. Another form of context, dealing mostly with the semantic part of the analysis would be more useful in this case [14]. In any case, research efforts focusing on search by visual similarity [3] may definitely benefit from the use of visual context information, as in the case of scene classification and object detection discussed in the following.

3 Visual Context in Image Analysis

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. Although image analysis deals with several well-known research problems, visual

context is mostly related to *scene classification* and *object detection*. *Scene classification* 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 so on. On the other hand, *object detection/recognition* 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. These 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 favor the recognition of a building vs. a tree. Solution to the above problem can be dealt through modeling of visual concept descriptors in one or more context domain ontologies and ontology learning/visual concept detection techniques that would utilize visual context information.

Attempts worth mentioning in the area include the one proposed in [7], where a list of semantic objects is used in a framework for semantic indexing and retrieval of video. As expected, colour has also been one of the central features of existing work on natural object detection. For example, in [9] *colour classification* is utilized in order to detect sky. In the context of content-based image retrieval, Smith and Li [12] assumed that a blue extended patch at the top of an image is likely to represent clear sky. An exemplar-based approach is presented more recently that uses a combination of colour and texture features to classify sub-blocks in an outdoor scene as sky or vegetation, assuming correct image orientation [13]. The latter brings up the issue of utilizing context orientation information in object class detection algorithms, a task that is generally avoided due to the fact that such contextual information is not always available and the performance of the algorithms is more than adequate despite this shortcoming.

However, none of the above methods and techniques utilizes visual context in the form we defined it herein. This tends to be the main drawback of these individual object 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 is very important and also extremely challenging even for human observers. 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. In this manner, it is useful in many cases to reduce the ambiguity among conflicting detectors and eliminate improbable spatial configurations in object detection.

3.1 The Role of Spatial Context

An important variation of visual context is *spatial context*; *spatial context* is associated to spatial relationships between objects or regions in a still image or video sequence. One may identify two types of spatial contextual relationships:

- Relationships that exist between co-occurrence of objects in natural images.
- Relationships that exist between spatial locations of certain objects in an image.

The definition of spatial context is an important issue for the notion of visual context in general. In order to be able to use context in applications, a mechanism to sense the current context - when thought as location, identities of nearby people or objects and changes to those objects - and deliver it to the application is crucial and must be present. A significant distinction exists between methods trying to determine location in computing applications and research fields. On the one hand, most of the existing approaches tend to restrict themselves, trying to infer the location where the image was taken (i.e., camera location); inferring the location of what the image was taken of (i.e., image content location) is a rather difficult and more complex task tackled by much less approaches. In [2], this challenge is addressed by leveraging regularities in a given user's and in a community of users' photo taking behaviors. Suitable weights, based on past experience and intuition, are chosen in order to assist in the process of location-determining features and then adjusted through a process of trial and error. An example describing the notion behind the method considers the following: it seems rather intuitive that if two pictures are being taken in the same location within a certain time frame (e.g., a few minutes for pedestrian users), they are probably in or around the same location.

Another factor to be considered is the intersection of spatial metadata in determining the location of image content. For example, patterns of being in certain locations at certain times with certain people will help determine the probability of which building in an area a user might be in. Information on whether this particular building is the place he/she works in can also be derived in such a case. Rule-based constraint and inference engines can also be used to aid reasoning, as well as machine learning algorithms to learn from past performance to optimize and adjust the relative importance of the various location-determining features. Taking the process a step further into the field of context modeling, transforms the problem into how to represent the contextual information in a way that can help bridging the gap between applications using contextual information and the deployment of context-aware services. The development of such applications requires tools that are based on clearly defined models of context. A simple approach is to use a plain model with context being maintained by a set of environment variables.

3.2 The Role of Scene Context

Visual context information may also be derived from the overall description of the entire scene; the so-called *scene context*. In a number of studies the context provided by a real-world scene has been claimed to have a mandatory, perceptual effect on the

identification of individual objects in such a scene. This claim has provided a basis for challenging widely accepted data-driven models of visual perception. The so far discussed visual context, defined by normal relationships among the locations of different materials in the scene without knowing exactly what the scene type is, is referred to as spatial context, and it the one that is going to be used mostly in a multimedia system application. In the sequel, visual context analysis is discussed in relation to the problems of scene classification and object detection.

Given the increase in the number and size of digital archives and libraries, there is a clear need for automated, flexible, and reliable image search and retrieval algorithms, as well as for image and video database indexing. Scene classification provides solutions in the means of suitable applications for all of these problems. The ultimate goal is to classify scenes based on their content. However, scene classification remains a major open challenge. Most solutions proposed so far, such as those based on colour histograms and local texture statistics [1][11], lack the ability to capture a scene's global configuration, which is critical in perceptual judgments of scene similarity. On the other hand, common standard approaches to object detection usually look at local pieces of an image in isolation when deciding if the object is present or not at a particular location. Of course, this is suboptimal and can be easily illustrated in the following example: consider the problem of finding a table in an office. A table is typically covered with other objects; indeed almost none of the table itself may be visible, and the parts that may be visible, such as its edge, are fairly generic features that may occur in many images. However, the table can be identified using contextual cues of various kinds. Of course, this problem is not restricted to tables, or occluded objects: almost any object, when seen at a large enough distance, becomes impossible to recognize without using visual context.

However, most techniques utilized in the field have usually positive results only in case of objects which have well-defined boundaries. Consequently, such strategies are not well suited for complex scenes, especially those which consist mostly of natural objects. The main difference between scene classification and object recognition techniques relies in the latter statement. Given these difficulties inherent in individual object recognition, scene classification approaches usually classify scenes without first attempting to recognize their components. Also, efforts have been made in using scene classification to facilitate object detection, and vice versa [5].

As already discussed in this section, several approaches of analyzing the content of images exist in the literature and many aspects of context are identified aiding in the process of image analysis. One of the main goals in the field is the effective combination of local and global information, towards implementing robust methods to use in typical image analysis problems. It should be clear by now, that visual context can play a key role in the procedure of combining this information; context should actually stand in the middle, being able to handle both types of information and providing the means to achieve better coherence and reliable research results. In order to achieve the latter, appropriate visual context models should be selected and designed in a straightforward and productive manner, utilizing the variations of the particular aspects of visual context.

4 Visual Context Modeling

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 actually model the relationships between the information and not the information themselves, with respect to the level of details present in each relationship. In this manner, at least two types of meaningful visual (spatial) contextual relationships are identified in digital images: a) relationships exist between co-occurrence of certain objects in the image; e.g. detection of *snow* with high probability would imply low *grass* probability, b) relationships exist between spatial locations of certain objects within an image; e.g. *grass* tends to occur below *sky*, *sky* above *snow*. The goal is, of course, to develop a non-scene specific method for generating spatial context models useful for general scene understanding problems.

In general, spatial context modeling refers to the process of building relationship models that define the spatial arrangement and distribution of objects of interest in a scene. There has been prior work on using high-level scene models for spatial context-based material detection [8]. However, the main limitation of such techniques is the need for constructing a different model for each scene type, thus restricting its applicability to a general scene understanding application. Other researchers propose different approaches for spatial context modeling, e.g. configuration-based scene modeling, targeted towards content-based indexing and retrieval applications [4]. In this work, the qualitative and photometric relationships between various objects in a scene are modeled in a spatial sense, and these relationships are used to retrieve other scenes with semantically similar content.

Extending this technique in the case of scene classification, the approach is different in the concept that a top-down technique is necessary, since information is not available in the form of objects, but in the form of regions. Towards fulfilling the ultimate goal of this task, i.e. , classification of images or video sequences based on their content, most of the strategies implemented use aggregate measures of an image's colour and texture as a signature for the image and compare these signatures afterwards in order to achieve levels of similarity between the images. However, this is not adequate in the case images' components vary significantly in colour distribution, texture, illumination or even spatial layout. In such cases the aid of visual context is more than evident as depicted in [6].

5 Conclusions and Future Work

Structures and techniques for representing and exploiting visual contextual information are necessary preconditions for the smooth operability of multimedia analysis. In this work we attempted to introduce suitable types of visual context and context models; we identified two types of context, namely context of content analysis and context of use. We observed why visual context information may be extremely helpful in knowledge extraction, especially when handling typical multimedia analysis problems like scene classification and object detection.

As part of a future avocation scene classification and/or object detection techniques may benefit from available visual contextual information, in order to provide information about indoor/outdoor scenery at the metadata level. Also information about the mood of depicted persons or of the depicted scene as a whole may be tackled by means of analyzing visual context. The latter can also aid towards satisfying simple user requests such as the orientation of a multimedia item. As far as retrieval is concerned, closest match search capabilities together with image search by visual similarity depict clearly possible future benefits from exploiting visual context information parameters. In the field of a multimedia system's content adaptation, the task of correction of image orientation or even general enhancements is tackled by methods dealing to a great degree with visual context.

References

1. Ashley, J., Flickner, M., Lee, D., Niblack, W., Petkovic, D.: Query by image content and its applications. IBM Research Report, RJ 9947 (87906) Computer Science/Mathematics (March 1995)
2. Davis, M., Good, N., Sarvas, R.: From Context to Content: Leveraging Context for Mobile Media Metadata (2004)
3. Kalantidis, Y., Tolias, G., Avrithis, Y., Phinikettos, M., Spyrou, E., Mylonas, P., Kollias, S.: VIRaL: Visual Image Retrieval and Localization. Multimedia Tools and Applications 51(2), 555–592 (2011)
4. Lipson, P., Grimson, E., Sinha, P.: Configuration based scene classification and image indexing. In: IEEE International Conference on Computer Vision & Pattern Recognition (1997)
5. Murphy, K., Torralba, A., Freeman, B.: Using the Forest to See the Trees: A Graphical Model Relating Features, Objects, and Scenes. In: NIPS 2003 (2003)
6. Mylonas, P., Spyrou, E., Avrithis, Y., Kollias, S.: Using Visual Context and Region Semantics for High-Level Concept Detection. IEEE Transactions on Multimedia 11(11), 229–243 (2009)
7. Naphade, M., Huang, T.S.: A factor graph framework for semantic indexing and retrieval in video. In: CVPR Workshop on Content-based Image and Video Retrieval (2000)
8. Ohta, Y.: Knowledge-based interpretation of outdoor natural color scenes. Pitman Advanced Publishing Program, Boston (1983)
9. Saber, E., Tekalp, A.M., Eschbach, R., Knox, K.: Automatic image annotation using adaptive colour classification. CVGIP: Graphical Models and Image Processing 58, 115–126 (1996)
10. Schilit, B., Adams, N., Want, R.: Context-Aware Computing Applications. In: IEEE Workshop on Mobile Computing Systems and Applications, Santa Cruz, CA (1994)
11. Smith, J.R., Chang, S.: Local color and texture extraction and spatial query. In: IEEE International Conference on Image Processing (1996)
12. Smith, J.R., Li, C.-S.: Decoding image semantics using composite region templates. In: IEEE Int. Workshop on Content-based Access of Image & Video Database (1998)
13. Vailaya, A., Jain, A.: Detecting sky and vegetation in outdoor images. In: SPIE, vol. 3972 (January 2000)
14. Wallace, M., Akrivas, G., Mylonas, P., Avrithis, Y., Kollias, S.: Using context and fuzzy relations to interpret multimedia content. In: 3rd International Workshop on Content-Based Multimedia Indexing (CBMI), IRISA, Rennes, France (September 2003)