Computer Vision and Image Understanding 116 (2012) 698–714

Contents lists available at SciVerse ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

A model for the qualitative description of images based on visual and spatial features ☆

Zoe Falomir [a,b,*], Lledó Museros [a], Luis Gonzalez-Abril [c], M. Teresa Escrig [b], Juan A. Ortega [d]

[a] Department of Engineering and Computer Science, Universitat Jaume I, ES Tecnologia i Ciéncies Experimentals, Av. Vicent Sos Baynat s/n, E-12071 Castelló, Spain
[b] Cognitive Robots SL, Universitat Jaume I, Espaitec I, Av. Vicent Sos Baynat s/n, E-12071 Castelló, Spain
[c] Department of Applied Economics I, Universidad de Sevilla, Facultad de Ciencias Económicas, Avda. Ramón y Cajal, 1, E-41018 Sevilla, Spain
[d] Department of Languages and Computer Systems, Universidad de Sevilla, ETS Ingeniería Informática, Avda. Reina Mercedes s/n, E-41012 Sevilla, Spain

ARTICLE INFO

Article history:
Received 27 October 2010
Accepted 31 January 2012
Available online 16 February 2012

Keywords:
Computer vision
Qualitative shape
Qualitative colours
Qualitative orientation
Spatial description

ABSTRACT

An approach that provides a qualitative description of any image is presented in this paper. The main visual features (shape and colour) and the main spatial features (fixed orientation, relative orientation and topology) of each object within the image are described. This approach has been tested in two real scenarios that involve agents and human interaction: (i) images captured by the webcam of a mobile robot while it navigates, and (ii) images of tile compositions captured by an industrial camera used to select tile pieces to be used in assembling tile mosaics. In both scenarios, promising results have been obtained.

1. Introduction

Digital images are fully integrated within modern daily life. Digital cameras are used to take photographs of trips and holidays, mobile phone cameras allow users to capture any scene in everyday life, and webcams in laptops are used to communicate the images of users' surroundings instantly across the network. The digital images generated can be easily copied, deleted, edited, sent by email or multimedia messages, included in web pages, etc. and computer systems and programs have been developed to provide all these possibilities. However, there is still no system capable of describing a digital image cognitively, that is, in a similar way to how people do it.

Psychological studies carried out on how people describe images [1–4] explain that people find the most relevant content in the images and use words (qualitative tags) to describe it. Usually different colours/textures in an image indicate different objects/regions of interest to people [5]. Moreover, cognitive studies [6] explain that, although the retinal image of a visual object is a quantitative image in the sense that specific locations on the retina are stimulated by light of a specific spectrum of wavelengths and intensity, the knowledge about this image that can be retrieved from memory is qualitative because absolute locations, wavelengths and intensities cannot be retrieved from human memory. Thus, qualitative representations are similar in many ways to the *mental images* that people report when they describe what they have seen from memory or when they attempt to answer questions on the basis of visual memories [7,8].

Therefore, a cognitive description of any digital image must be a qualitative description that could be understood and interpreted by human users, which would allow the user-machine communication in many applications to be enhanced. For example, the qualitative description obtained may be used as the key search in image retrieval from data bases, and it may also be easily post-processed to produce a written narrative description of any image that may be included in a user-interface or read aloud by a speech synthesiser application for blind users to know what the image shows. Moreover, a cognitive description of a digital image must be able to describe any kind of object/region in the image by its features, regardless of whether it has been seen or unseen previously, as people can describe a scene or objects that they have never seen before or the names of which they cannot recall. Finally, it must be possible to extract and compute a cognitive description of any digital image automatically. It must also be independent of the image segmentation methods used to obtain the relevant regions in

☆ This paper has been recommended for acceptance by R. Bergevin.

* Corresponding author at: Department of Engineering and Computer Science, Universitat Jaume I, ES Tecnologia i Ciéncies Experimentals, Av. Vicent Sos Baynat s/n, E-12071 Castelló, Spain.

E-mail addresses: zfalomir@uji.es, zfalomir@gmail.com (Z. Falomir), museros@uji.es (L. Museros), luisgon@us.es (L. Gonzalez-Abril), escrigm@uji.es (M.T. Escrig), jortega@us.es (J.A. Ortega).

1077-3142/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
doi:10.1016/j.cviu.2012.01.007

the image, so that the system could apply a new more efficient method when it appears in the literature.

However, using computers to extract visual information from space and interpreting it in a meaningful way as human beings can do remains a challenge. Because digital images represent visual data numerically, most image processing has been carried out by applying mathematical techniques to obtain and describe image content. In the recent literature there are methods that describe and compare digital images numerically in order to obtain a degree of similarity between them [9–12]. All these approaches succeeded in the task they were designed for. However, they produce huge numerical file descriptions that cannot be interpreted or given a meaning without a correspondence of descriptions produced by other images of visualised scenes or objects previously stored in memory. Moreover, most of them need training or learning techniques. The main disadvantage of these methods is that a repository of all the possible images of scenes or objects existing in the world is still not possible. Therefore, they only succeed in specific delimited contexts. And they are not cognitive because they are not able to describe any feature of an object or scene that they have not seen before, that is, that have not been previously stored in memory.

Furthermore, extracting semantic information from images is still an on-going area of research in computer vision. The association of meaning with the representations of objects obtained by robotic systems, also known as the *symbol-grounding problem*, is a prominent issue within the field of Artificial Intelligence [13]. A qualitative description of images can contribute in this topic because it would be understandable not only by people but also by intelligent agents. The advantage of a description based on qualitative tags is that a semantic meaning can be assigned to them by means of ontologies. Therefore, the knowledge of any agent able to describe images qualitatively would be increased, i.e. a software agent could 'know' the content of the images on the Web or a physical agent (i.e. a mobile robot) could 'know' the features of all the objects in the images captured by its webcam even if these objects have not been seen before.

In this paper, the contribution presented is the automatic extraction of a qualitative description of any digital image based on the description of the visual and spatial information of the relevant regions within it, which are extracted independently of the segmentation method applied to the image. Specifically, the Qualitative Image Description approach (QID approach) uses qualitative models of shape and colour to describe the visual features of each relevant region in the image, and the qualitative models of topology [14] and orientation [15,16] to describe their spatial features.

The QID approach is intended to contribute to the solution of the open research issues previously presented. In this paper, the QID approach is applied to two real-world scenarios that involve agents and human interaction: (i) images captured by the webcam of a mobile robot while it navigates, and (ii) images of tile compositions captured by an industrial camera used to select tile pieces to be used in assembling tile mosaics. In addition, here the QID approach is generalised to show its flexibility in all kind of images and also its adaptability to other scenarios (i.e. medical image description, geographical image description, etc.).

This flexibility opens the way to future applications that involve extending the QID approach for: (1) extracting meaning from images and improving the understanding of those images by web agents by translating the QID to a description logics-based ontology; (2) obtaining a semantic similarity measure between the meaning of two ontological descriptions, where this similarity would calculate the resemblance between the different instances generated; (3) measuring the similarity of two qualitative image descriptions from the point of view of human thinking; (4) using that similarity measure for visual image/scene recognition and retrieval from a cognitive point of view (applicable to design automation processes, psychological research, and other fields involving imitation and study of human perception); (5) generating a human-language description for each processed image using a context free grammar which may produce a written paragraph describing the scene and which may be read aloud for a speech-synthesiser application for blind users to understand; etc.

The remainder of this paper is organised as follows. Section 2 presents the related work. Section 3 presents the QID approach and its implementation is explained in Section 4. Section 5 details the two scenarios where the experimentation was carried out and the results obtained. Finally, conclusions are drawn in Section 6.

2. Related work

Similar approaches that extract qualitative or semantic information from images representing scenes have appeared in the literature [17–20,10]. Socher et al. [17] provided a robotic manipulator system with a verbal description of an image so that it can identify and pick up an object that has been previously modelled geometrically and then categorised qualitatively by its type, colour, size and shape. The spatial relations between the predefined objects detected in the image are also described qualitatively. Lovett et al. [18] proposed a qualitative description for sketch image recognition, which described lines, arcs and ellipses as basic elements and also the relative position, length and orientation of their edges. Qayyum and Cohn [19] divided landscape images using a grid for their description so that semantic categories (grass, water, etc.) could be identified and qualitative relations of relative size, time and topology could be used for image description and retrieval in data bases. Oliva and Torralba [20] obtained the *spatial envelope* of complex environmental scenes by analysing the discrete Fourier transform of each image and extracting perceptual properties of the images (naturalness, openness, roughness, ruggedness and expansion) that enable the images to be classified in the following semantic categories: coast, countryside, forest, mountain, highway, street, close-up and tall building. Quattoni and Torralba [10] proposed an approach for classifying images of indoor scenes in semantic categories such as book store, clothing store, kitchen, bathroom, restaurant, office, classroom, etc. This approach combined global spatial properties and local discriminative information (i.e. information about objects contained in places) and used learning distance functions for visual recognition.

All the studies described above provide evidence for the effectiveness of using qualitative information to describe images. However, in the approach developed by Socher et al. [17], a previous object recognition process is needed before it becomes possible to give a qualitative description of the image of the scene the robot manipulator has to manage, whereas the QID approach is able to describe the image of the scene in front of the robot without this prior process. The approach of Lovett et al. [18] is applied to sketches, while the QID approach is applied to digital images captured from the real robot environment. Qayyum and Cohn [19] used a grid to divide the image and to describe what is inside each grid square (grass, water, etc.), which is suitable for their application. However, the objects are divided into an artificial number of parts that depend on the size of the cell, while the QID approach extracts complete objects. Oliva and Torralba's [20] approach is useful for distinguishing between outdoor environments. However, as this approach does not take into account local object information, it will obtain similar *spatial envelopes* for similar images corresponding to the indoor environments where our robot navigates, such as corridors in buildings. Quattoni and Torralba's [10] approach performs well for recognising indoor scenes, although it uses a learning distance function and, therefore, it must be