Available online at www.sciencedirect.com
ScienceDirect
Information Processing and Management 43 (2007) 515–530
INFORMATION PROCESSING & MANAGEMENT
www.elsevier.com/locate/infoproman

Object identification and retrieval from efficient image matching. Snap2Tell with the STOIC dataset

Jean-Pierre Chevallet [a,b,*], Joo-Hwee Lim [b], Mun-Kew Leong [b]

[a] IPAL-Centre National de la Recherche Scientifique (CNRS) Laboratory, UMI 2955, France
[b] Institute for Infocomm Research (I^2R), 21 Heng Mui Keng Terrace, Singapore 119613, Singapore

Received 23 June 2006; accepted 25 July 2006

Abstract

Traditional content based image retrieval attempts to retrieve images using syntactic features for a query image. Annotated image banks and Google allow the use of text to retrieve images. In this paper, we studied the task of using the content of an image to retrieve information in general. We describe the significance of object identification in an information retrieval paradigm that uses image set as intermediate means in indexing and matching. We also describe a unique Singapore Tourist Object Identification Collection with associated queries and relevance judgments for evaluating the new task and the need for efficient image matching using simple image features. We present comprehensive experimental evaluation on the effects of feature dimensions, context, spatial weightings, coverage of image indexes, and query devices on task performance. Lastly we describe the current system developed to support mobile image-based tourist information retrieval.

Keywords: Image retrieval; Object identification; Mobile information retrieval; Test collection building

1. Introduction

The primary difference between text and non-text IR is that text IR attempts to retrieve relevant documents based on "semantic" content whereas traditional non-text IR (e.g., content based image retrieval (CBIR)) attempts to retrieve images based on "syntactic" (i.e., low-level) features. If we considered, for the sake of illustration, that an image is analogous to a printed page of a document, on such an analogy, traditional CBIR which is feature based, would roughly be similar to retrieving text documents based on their font size, their layout, the colour of the ink, etc. (i.e., physical characteristics of the document) rather than on their meaningful content.

* Corresponding author. Address: Institute for Infocomm Research (I^2R), 21 Heng Mui Keng Terrace, Singapore 119613, Singapore. Tel.: +65 68748526; fax: +65 6775 501.
E-mail addresses: viscjp@i2r.a-star.edu.sg, Jean-Pierre.Chevallet@imag.fr (J.-P. Chevallet), joohwee@i2r.a-star.edu.sg (J.-H. Lim), mkleong@i2r.a-star.edu.sg (M.-K. Leong).

0306-4573/$ - see front matter
doi:10.1016/j.ipm.2006.07.017

In general, image retrieval systems are only of the following types: using image to retrieve images, using non-image (usually text) to retrieve images, or using image to retrieve non-images (information in general). Traditional CBIR is of the first variety, annotated image banks and Google are of the second, and there are only a few efforts in the third. In this paper we explore the possibility of non-image retrieval based on the content of images. Such an approach roughly requires an *object identification* phase to generate[1] the semantic content followed by whatever suitable actions based on those semantics.

There are many applications for successful object identification systems. We are particularly interested in two of them. The first is for homeland security or image monitoring in general. In our discussions with intelligence agencies, they tell us that they need a way to filter images in the same way that text is filtered. Traditional CBIR does not work for their scenario which needs to work at the semantic object level and not at the syntactic feature level. Another strong reason is that it is easy to give a codeword to replace a name or definite description. So terrorists may use *FOO* in email and chat to refer to, say, the *Subic Bay Naval Base* thus defeating keyword spotting algorithms, but if they were to exchange an image, it would be very difficult to code it. Note that they cannot just encrypt their conversations or images since encrypted data is a red flag in monitoring scenarios.

The second application follows an important trend in mobility; this is the increasing prevalence of cameras on mobile phones. During an industry panel at the Consumer Electronics Show in January 2005[2] it was estimated that 700 million mobile handsets will be sold in 2005 and 2/3 of them will have cameras. A significant number of pictures taken on such cameras are likely to be ''throw-away'' images, i.e., pictures taken to serve a function and which has no value once that function is served. Scenarios mooted include taking a picture of a dress to get an opinion from a friend, or as an illustration to a message. But the scenario which we are interested in, is taking a picture to find out more information. So a tourist takes a picture of an unknown landmark, sends it to a server, and gets back useful information. Or a health-conscious consumer takes a picture of his dinner, sends it to a server and gets back nutritional information.

In Section 2, we emphasize the significance of a few aspects in our work including the task, the indexing and retrieval paradigm, a unique image dataset, and the requirement for fast query processing. Then we describe important applications related to the object identification task followed by related approaches. Section 4 is devoted to the description of the current prototype on mobile image-based tourist directory. Our experimental evaluation on our unique STOIC dataset is given in Section 5.

2. Significance

There are four key aspects in our work. First, we look at object identification as an important genre of image search. Second, images are used as intermediate means to retrieve information about an object or location. Third, we introduce a new type of image dataset with associated queries and relevance judgments. Conventional image datasets are not designed or evaluated at the semantic level. Last but not least, we show that simple image feature matching is sufficient for good object identification if we provide a sufficient image set for object description. Such efficient techniques are necessary for very large scale critical applications such as homeland security monitoring or the limited processing capacity of the ubiquitous camera phones.

2.1. Object identification

The object identification task may be described as follows: given an image, determine the referent[3] for the most salient object in the image. For example, all three images in Fig. 1 are of the same object, albeit from different perspectives, scales, and colour. The referent is the Merlion statue in One Fullerton in Singapore. The most salient object in the image may be the image in its entirety, e.g., the skyline image in Fig. 2 but which also includes the Merlion in it.

[1] We use the term *generate* rather than *extract* since the semantics we want is often not intrinsic to the image.

[2] See ''DH: Digital Cameras Get Competition'', 5th January 2005, http://www.cesweb.org.

[3] Different words or phrases may be used to described objects or experiences. The *referent* is that which is designated by those words. So, ''George Washington'' and ''the first president of the United States'' both have the same referent.