IRM PRESS
701 E. Chocolate Avenue, Suite 200, Hershey PA 17033-1240, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.irm-press.com
ITB11369

This chapter appears in the book, *Managing Multimedia Semantics*, edited by Uma Srinivasan and Surya Nepal

Chapter 2

From Classification to Retrieval: Exploiting Pattern Classifiers in Semantic Image Indexing and Retrieval

Joo-Hwee Lim, Institute for Infocomm Research, Singapore

Jesse S. Jin, The University of Newcastle, Australia

ABSTRACT

Users query images by using semantics. Though low-level features can be easily extracted from images, they are inconsistent with human visual perception. Hence, low-level features cannot provide sufficient information for retrieval. High-level semantic information is useful and effective in retrieval. However, semantic information is heavily dependent upon semantic image regions and beyond, which are difficult to obtain themselves. Bridging this semantic gap between computed visual features and user query expectation poses a key research challenge in managing multimedia semantics. As a spin-off from pattern recognition and computer vision research more than a decade ago, content-based image retrieval research focuses on a different problem from pattern classification though they are closely related. When the patterns concerned are images, pattern classification could become an image classification problem or an object recognition problem. While the former deals with the entire image

as a pattern, the latter attempts to extract useful local semantics, in the form of objects, in the image to enhance image understanding. In this chapter, we review the role of pattern classifiers in state-of-the-art content-based image retrieval systems and discuss their limitations. We present three new indexing schemes that exploit pattern classifiers for semantic image indexing, and illustrate the usefulness of these schemes on the retrieval of 2,400 unconstrained consumer images.

INTRODUCTION

Users query images by using semantics. For instance, in a recent paper by Enser (2000), he gave a typical request to a stock photo library, using broad and abstract semantics to describe the images one is looking for:

"Pretty girl doing something active, sporty in a summery setting, beach — not wearing lycra, exercise clothes — more relaxed in tee-shirt. Feature is about deodorant so girl should look active — not sweaty but happy, healthy, carefree — nothing too posed or set up — nice and natural looking."

Using existing image processing and computer vision techniques, low-level features such as color, texture, and shape can be easily extracted from images. However, they have proved to be inconsistent with human visual perception, let alone the incapability to capture broad and abstract semantics as illustrated by the example above. Hence, low-level features cannot provide sufficient information for retrieval. High-level semantic information is useful and effective in retrieval. However, semantic information is heavily dependent upon semantic image regions and beyond, which are difficult to obtain themselves. Between low-level features and high-level semantic information, there is a so- called "semantic gap." Content-based image retrieval research has yet to bridge this "gap between the information that one can extract from the visual data and the interpretation that the same data have for a user in a given situation" (Smeulders et al., 2000).

In our opinion, the semantic gap is due to two inherent problems. One problem is that the extraction of complete semantics from image data is extremely hard, as it demands general object recognition and scene understanding. This is called the *semantics extraction problem*. The other problem is the complexity, ambiguity and subjectivity in user interpretation, that is, the *semantics interpretation problem*. They are illustrated in Figure 1. We think that these two problems are manifestation of two one-to-many relations.

In the first one-to-many relation that makes the *semantics extraction problem* difficult, a real world object, say a face, can be presented in various appearances in an image. This could be due to the illumination condition when the image of the face is being recorded; the parameters associated with the image capturing device (focus, zooming, angle, distance, etc.); the pose of the person; the facial expression; artifacts such as spectacles and hats; variations due to moustache, aging, and so forth. Hence, the same real-world object may not have consistent color, texture and shape as far as computer vision is concerned.