Content-Based Image Retrieval

Alan Wee-Chung Liew
Griffith University, Australia

Ngai-Fong Law
The Hong Kong Polytechnic University, Hong Kong

INTRODUCTION

With the rapid growth of Internet and multimedia systems, the use of visual information has increased enormously, such that indexing and retrieval techniques have become important. Historically, images are usually manually annotated with metadata such as captions or keywords (Chang & Hsu, 1992). Image retrieval is then performed by searching images with similar keywords. However, the keywords used may differ from one person to another. Also, many keywords can be used for describing the same image. Consequently, retrieval results are often inconsistent and unreliable.

Due to these limitations, there is a growing interest in content-based image retrieval (CBIR). These techniques extract meaningful information or features from an image so that images can be classified and retrieved automatically based on their contents. Existing image retrieval systems such as QBIC and Virage extract the so-called low-level features such as color, texture and shape from an image in the spatial domain for indexing.

Low-level features sometimes fail to represent high level semantic image features as they are subjective and depend greatly upon user preferences. To bridge the gap, a top-down retrieval approach involving high level knowledge can complement these low-level features. This articles deals with various aspects of CBIR. This includes bottom-up feature-based image retrieval in both the spatial and compressed domains, as well as top-down task-based image retrieval using prior knowledge.

BACKGROUND

Traditional text-based indexes for large image archives are time consuming to create. A domain expert is required to examine each image scene and describe its content using several keywords. The language-based descriptions, however, can never capture the visual content sufficiently because a description of the overall semantic content in an image does not include an enumeration of all the objects and their properties. Manual text-based annotation generally suffers from two major drawbacks: (i) content mismatch, and (ii) language mismatch. A content mismatch arises when the information that the domain expert ascertains from an image differs from the information that the user is interested in. When this occurs, little can be done to recover the missing annotations. On the other hand, a language mismatch occurs when the user and the domain expert use different languages or phrases to describe the same scene. To circumvent language mismatch, a strictly controlled set of formal vocabulary or ontology is needed, but this complicates the annotation and the query processes. In text-based image query, when the user does not specify the right keywords or phrases, the desired images cannot be retrieved without visually examining the entire archive.

In view of the deficiencies of text-based approach, major research effort has been spent on CBIR over the past 15 years. CBIR generally involves the application of computer vision techniques to search for certain images in large image databases. "Content-based" means that the search makes use of the contents of the images themselves, rather than relying on manually annotated texts.

From a user perspective, CBIR should involve image semantics. An ideal CBIR system would perform semantic retrievals like "find pictures of dogs" or even "find pictures of George Bush." However, this type of open-ended query is very difficult for computers to perform because, for example, a dog's appearance can vary significantly between species. Current CBIR systems therefore generally make use of low-level features like texture, color, and shape. However, biologically-inspired vision research generally suggests two processes in visual analysis: bottom-up image-based analysis and top-down task-related analysis (Navalpakkam & Itti, 2006). Bottom-up analysis consists of memoryless stimulus-centric factors such as low-level image features. Top-down analysis uses prior domain knowledge to influence bottom-up analysis. An effective image retrieval system should therefore combine both the low-level features as well as the high level knowledge so that images can be classified automatically according to their context and semantic meaning.

EXISTING CBIR SYSTEMS AND STANDARDS

The best-known commercial CBIR system is the QBIC (Query by Image Content) system developed by IBM (Flickner et al., 1995). Image retrieval is achieved by any combination of color, texture or shape as well as by keyword. Image queries can be formulated by selection from a palette, specifying an example image, or sketching a desired shape on the screen. The other well-known commercial CBIR systems are Virage (Gupta & Jain, 1997) which is used by AltaVista for image searching, and Excalibur (Feder, 1996) which is adopted by Yahoo! for image searching. Photobook (Pentland, Picard, & Sclaroff, 1996) from MIT Media Lab is the representative research CBIR system. Like QBIC, images are represented by color, shape, texture and other appropriate features. However, Photobook computes information-preserving features, from which all essential aspects of the original image can be reconstructed.

In 1996, the Moving Picture Experts Group (MPEG)–a working group (JTC1/SC29/WG11) of the International Organization for Standardization (ISO) and the International Electrotechnical Commission (IEC)—decided to start a standardization project called MPEG-7 (Manjunath, Salembier, & Sikora, 2002). The aim is to provide a quick and efficient identification and management of multimedia content so that audio-visual information can be easily searched.

The MPEG-7 specifies the descriptors, description schemes and a description definition language. The descriptors is a representation of features at different levels of abstraction, ranging from low-level visual features like shape, texture and color to high level semantic information such as abstract concept and genres. The descriptor defines the syntax and semantics of the feature representation.

The description schemes specify the structure and semantics of the relationships between its components such as descriptors. The scheme provides a solution to model and describe content in terms of structures and semantics. The description definition language allows the creation of new description schemes as well as descriptors. This allows extension and modification of existing description schemes.

The MPEG-7 standard has eight parts. Of these, Part 3 specifies a set of standardized low-level descriptors and description schemes for visual content which includes shape descriptor, color descriptor, texture descriptor and motion descriptor. Note that the MPEG-7 specifies the descriptors only, and their extraction are not specified as part of the MPEG-7 standard.

CBIR METHODOLOGY

A CBIR system has three key components: feature extraction, efficient indexing and user interface:

- **Feature extraction:** Image features include *primitive features* and *semantic features*. Examples of primitive features are *color*, *texture*, and *shape.* Primitive features are usually quantitative in nature and they can be extracted automatically from the image. Semantic features are qualitative in nature and they provide abstract representations of visual data at various levels of detail. Typically, semantic features are extracted manually. Once the features have been extracted, image retrieval becomes a task of measuring similarity between image features.
- **Efficient indexing:** To facilitate efficient query and search, the image indices needed to be organized into an efficient data structure. Because image features maybe interrelated, flexible data structures should be used in order to facilitate storage/retrieval. Structures such as *k-d-tree*, *R-tree*, *R*-tree*, *quad-tree*, and *grid file* are commonly used.
- **User interface:** In visual information systems, user interaction plays an important role. The user interface consists of a query processor and a browser to provide an interactive environment for querying and browsing the database. Common query mechanisms provided by the user interface are: *query by keyword*, *query by sketch*, *query by example*, *browsing by categories*, *feature selection*, and *retrieval refinement*.

In *query by example*, the user specifies a query image (either supplied by the user or chosen from a random set), and the system finds images similar to it based on various low-level criteria. In *query by sketch*, the user draws a rough sketch of the image he/she is looking for, for example, with blobs of color at different locations, and the system locates images whose layout matches the sketch. In either case, features are first extracted automatically from this query image to form a query image signature. A matching with all other images in the archive is performed by measuring the similarity between their signatures. It can be seen that the matching result is heavily influenced by the choice of features.

BOTTOM-UP CONTENT-BASED RETRIEVAL IN THE SPATIAL DOMAIN

Content-based retrieval makes use of low level image features computed from the image itself for matching. Commonly used