A Bayesian Image Retrieval Framework

Rui Zhang, Ryerson University, Canada
Ling Guan, Ryerson University, Canada

ABSTRACT

Conventional approaches to content-based image retrieval exploit low-level visual information to represent images and relevance feedback techniques to incorporate human knowledge into the retrieval process, which can only alleviate the semantic gap to some extent. To further boost the performance, a Bayesian framework is proposed in which information independent of the visual content of images is utilized and integrated with the visual information. Two particular instances of the general framework are studied. First, context which is the statistical relation across the images is integrated with visual content such that the framework can extract information from both the images and past retrieval results. Second, characteristic sounds made by different objects are utilized along with their visual appearance. Based on various performance evaluation criteria, the proposed framework is evaluated using two databases for the two examples, respectively. The results demonstrate the advantage of the integration of information from multiple sources.

Keywords: Bayesian Image Retrieval, Integration of Content and Context, Integration of Audio and Visual Information

I. INTRODUCTION

The ever-lasting growth of multimedia information has been witnessed and experienced by human beings since the beginning of the information era. An immediate challenge resulting from the information explosion is how to intelligently manage and enjoy the multimedia databases. In the course of the technological development of multimedia information retrieval, various approaches have been proposed with the ultimate goal of enabling semantic-based search and browsing. Among those intensively explored topics, content-based image retrieval (CBIR), born at the crossroad of computer vision, machine learning and database technologies, has been studied for more than a decade, yet still remaining difficult (Smeulders, Worring, Santini, Gupta, & Jain, 2001; Datta, Joshi, Li, & Wang, 2008). In a nutshell, the content-based approaches to image retrieval primarily rely on the pictorial information, a.k.a. low level visual features such as color, texture, shape and layout, which can be automatically extracted from images for similarity measure. The essential challenge is that the low level visual features accurately characterizing the semantic meaning of images are difficult to discover. Therefore, semantically relevant images may be located far away from each other in the space of the
pictorial information, which is referred to as the semantic gap. To reduce the semantic gap, human knowledge was utilized to help refine the representation of the semantic meaning in a user’s query. To this end, the relevance feedback (RF), a technique originally proposed for traditional document retrieval, was adapted to solve the problem of image retrieval (Cruceanu, Ferecatu, & Boujemaa, 2004; Zhou & Huang, 2003). A common aspect of most RF techniques is that the learned knowledge will not be propagated forward to the retrieval in the future and hence can be considered as the short-term relevance feedback (STRF). STRF techniques alleviate the semantic gap by incorporating human users’ knowledge into the process of labeling training samples yet still suffering from the problem of sample sparseness, as average users are normally willing to select only a few relevant and irrelevant images. In addition, as irrelevant images may be distinct from the relevant ones in many different ways, there is a good chance that training samples of the two categories in the context of STRF are imbalanced. Along with the demand for the real-time performance of a practical search engine, the above-mentioned problems can be considered as the major factors leading to the performance bottleneck.

To further boost the performance of a retrieval system, we propose a Bayesian framework to tackle the semantic gap of image retrieval by: 1) incorporating information independent of the low-level visual content; 2) integrating such information with the visual content of images based on the Bayes’ theorem. To study the effectiveness of the framework, we consider two particular examples, in which the contextual and audio information are integrated with the visual features, respectively. To be specific, the context is defined as the statistical relation across the images in the database. The Bayesian image retrieval framework utilizing content and context is motivated by the following thoughts. It has been recognized that information extracted within only a single retrieval session is not sufficient and learning the representation of high level semantics associated with images is a long-term task. Hence, a retrieval framework capable of utilizing the information learned from both the past and the present retrieval is desired. The advantages of the Bayesian image framework using content and context include: 1) when the contextual information is not available or sufficient, the content-based component of the system can help accumulate the training samples, with which the context component can be incrementally updated; 2) the contextual information, when available, can help reduce the semantic gap of the content-based component yet still with the flexibility that a user can refine the query on his/her own preference using STRF. As opposed to the STRF, the estimation of the context model can be regarded as long-term relevance feedback (LTRF) because information assisting the retrieval in the future is learned from the past. Since the contextual information is learned from past retrieval results with relevance feedback from many different users, the approach is also referred to as collaborative Bayesian image retrieval (CLBIR). In addition, the Bayesian framework for image retrieval using both visual and audio features is motivated by the fact that most of the real-world objects make their own representative sound. In terms of practical applications, it is also driven by the increasing demand of users for on-the-go information retrieval resulting from the rocketing popularity of camera-integrated mobile phones. A potential application is searching for images of animals using a mobile phone while traveling around in a wildlife park. Since the audio information is not available for most of the image databases, we only collect training samples for a single object class rather than each individual image. The learned audio information is then propagated to the images by classifying them into the categories for which the audio samples are collected. The advantage of this framework lies in its ability to utilize information from multiple modalities. To evaluate the performance, we experiment with the Corel database with 10000 images and a database of 4400 images featuring 44 kinds of animals and audio files containing the animals’ sound collected from the Internet,