Available online at www.sciencedirect.com
ScienceDirect
Information Processing and Management 43 (2007) 488–502
INFORMATION PROCESSING & MANAGEMENT
www.elsevier.com/locate/infoproman

Integrating textual and visual information for cross-language image retrieval: A trans-media dictionary approach

Wen-Cheng Lin, Yih-Chen Chang, Hsin-Hsi Chen *

Department of Computer Science and Information Engineering, National Taiwan University, No. 1, Sec. 4, Roosevelt Road, Taipei 106, Taiwan

Received 17 May 2006; accepted 25 July 2006
Available online 11 October 2006

Abstract

This paper explores the integration of textual and visual information for cross-language image retrieval. An approach which automatically transforms textual queries into visual representations is proposed. First, we mine the relationships between text and images and employ the mined relationships to construct visual queries from textual ones. Then, the retrieval results of textual and visual queries are combined. To evaluate the proposed approach, we conduct English monolingual and Chinese–English cross-language retrieval experiments. The selection of suitable textual query terms to construct visual queries is the major issue. Experimental results show that the proposed approach improves retrieval performance, and use of nouns is appropriate to generate visual queries.

Keywords: Cross-language image retrieval; ImageCLEF; Language translation; Medium transformation; Trans-media dictionary

1. Introduction

Today multimedia data grows explosively. The Internet, for example, contains millions of images, videos, and music. Finding the requested information from large amounts of multimedia data is challenging. Two types of approaches, i.e., content-based and text-based, are usually adopted in image retrieval (Goodrum, 2000). Content-based image retrieval (CBIR) uses low-level visual features such as color, texture, and shape to represent images. Users can employ example images as queries, or directly specify the weight of low-level visual features to retrieve images. Images that are visually similar to an example image or contain the specified visual features are returned.

* Corresponding author. Tel.: +886 2 33664888x311; fax: +886 2 23628167.
E-mail addresses: denislin@nlg.csie.ntu.edu.tw (W.-C. Lin), ycchang@nlg.csie.ntu.edu.tw (Y.-C. Chang), hhchen@csie.ntu.edu.tw (H.-H. Chen).

0306-4573/$ - see front matter
doi:10.1016/j.ipm.2006.07.015

In text-based approaches, text is used to describe images and formulate queries. Because images and image representations are in different types of media, media transformation is required. Images are transformed into text, and a text retrieval system is used to index and retrieve images. Textual features can also be derived from the text accompanying an image such as a caption or the surrounding text. Text-based approach encounters the following problems:

(1) Image captions are usually short. The short annotation cannot represent the image content completely.
(2) Image captions are not always available. Manually assigning captions to images is time consuming and costly.
(3) Some visual properties cannot be described directly in captions. For example, the styles of images, e.g., warm, cold, dark, sharp, or blurry, are usually not specified in captions.
(4) Users' queries may have different levels of semantics. Users may search for images at a higher semantic level or at a primitive level.

Since images are produced by people familiar with their own language, they can be annotated in different languages. In this way, text-based image retrieval has a multilingual nature. In addition, images are understandable by different language users. They can resolve the major argument in cross-language information retrieval, i.e., users that are not familiar with the target language cannot understand the retrieved documents. In such a situation, cross-language image retrieval has attracted researchers' attentions recently and is organized as one of evaluation tasks in the Cross-Language Evaluation Forum (CLEF) (Clough, Sanderson, & Müller, 2005). In addition to media transformation, language translation is also necessary to unify the language usages in queries and documents in cross-language image retrieval.

Textual and low-level visual features have different semantic levels. Textual feature is highly semantic, while low-level visual feature is less semantic and is more emotive. These two types of features are complementary and provide different aspects of information about images. In this paper, we explore the integration of textual and visual information in cross-language image retrieval. An approach that automatically transforms textual queries into visual representations is proposed. The generated visual representation is treated as a visual query to retrieve images. The retrieved results using textual and visual queries are combined to generate the final result.

The rest of this paper is organized as follows. Section 2 introduces the proposed model. The integration of textual and visual information is illustrated. Section 3 models the relationships between text and images. How to generate visual representation of a textual query is introduced. Section 4 specifies the experimental materials. Section 5 shows the experiment designs. Here, the selection of suitable textual query terms to construct visual queries is the major issue. In addition, three types of experiments are evaluated, including monolingual image retrieval, cross-language image retrieval and ideal visual queries. Finally, we conclude our work in Section 6.

2. Integrating textual and visual information

Several hybrid approaches that integrate visual and textual information have been proposed. A simple approach conducts text- and content-based retrieval separately and merges the retrieval results of the two runs (Besançon, Hède, Moellic, & Fluhr, 2005; Jones et al., 2005; Lin, Chang, & Chen, 2005). In contrast to this parallel approach, a pipeline approach employs textual or visual information to perform initial retrieval, and then uses the other feature to filter out the irrelevant images (Lowlands Team, 2001). In the above two approaches, users have to issue two types of queries, i.e., textual and visual. In these approaches, sometimes it is not intuitive to find an example image or to specify low-level visual features.

We take another approach as shown in Fig. 1 with our cross-language image retrieval system. This system automatically transforms textual queries into visual representations. First, the relationships between text and images are mined from a set of images annotated with text descriptions. A trans-media dictionary which is similar to a bilingual dictionary is set up from the training collections. When a user issues a textual query, the system automatically transforms the textual query into a visual one using the trans-media dictionary. The generated visual representation is treated as a visual query and is used to retrieve images. In this way, we have both textual and visual queries.