Information Processing and Management 47 (2011) 676–691

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

A query expansion framework in image retrieval domain based on local and global analysis

M.M. Rahman *, S.K. Antani, G.R. Thoma

U.S. National Library of Medicine, National Institutes of Health, Bethesda, MD, USA

ARTICLE INFO

Article history:
Received 14 June 2010
Received in revised form 6 December 2010
Accepted 19 December 2010
Available online 22 January 2011

Keywords:
Image retrieval
Vector space model
Support vector machine
Relevance feedback
Query expansion

ABSTRACT

We present an image retrieval framework based on automatic query expansion in a concept feature space by generalizing the vector space model of information retrieval. In this framework, images are represented by vectors of weighted concepts similar to the keyword-based representation used in text retrieval. To generate the concept vocabularies, a statistical model is built by utilizing Support Vector Machine (SVM)-based classification techniques. The images are represented as "bag of concepts" that comprise perceptually and/or semantically distinguishable color and texture patches from local image regions in a multi-dimensional feature space. To explore the correlation between the concepts and overcome the assumption of feature independence in this model, we propose query expansion techniques in the image domain from a new perspective based on both local and global analysis. For the local analysis, the correlations between the concepts based on the co-occurrence pattern, and the metrical constraints based on the neighborhood proximity between the concepts in encoded images, are analyzed by considering local feedback information. We also analyze the concept similarities in the collection as a whole in the form of a similarity thesaurus and propose an efficient query expansion based on the global analysis. The experimental results on a photographic collection of natural scenes and a biomedical database of different imaging modalities demonstrate the effectiveness of the proposed framework in terms of precision and recall.

1. Introduction

To reduce the ambiguity due to the word mismatch problem in information retrieval and to thereby increase retrieval effectiveness, a variety of query modification and reformulation strategies have been investigated (Attar & Fraenkel, 1977; Crouch, 1990; Qiu & Frei, 1993; Salton & Buckley, 1999; Xu & Croft, 1996, 2000). There are three main approaches for query reformulation (Yates & Neto, 1999):

(1) Interactive techniques based on feedback information from the user, commonly known as *relevance feedback* (RF);
(2) Automated techniques based on the global information derived from the entire collection or corpus, commonly known as *thesaurus-based query expansion*, and
(3) Automated (might be interactive also in some cases) techniques based on local information from the top retrieved results, commonly known as *local feedback*.

* Corresponding author. Tel.: +1 301 435 3262.
E-mail address: rahmanmm@mail.nih.gov (M.M. Rahman).

0306-4573/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ipm.2010.12.001

The efficiency of image relevance computation has accelerated the development and use of RF in these three approaches (Salton & McGill, 1983; Rocchio, 1971; Ruthven & Lalmas, 2003) in content-based image retrieval (CBIR) applications (Zhou & Huang, 2003; Rui, Huang, & Mehrotra, 1997; Tong & Chang, 2001; Ishikawa, Subramanya, & Faloutsos, 1998). In CBIR, the access to any information is performed at a perceptual level based on automatically extracted visual features, such as color, texture, shape, and location (Smeulders, Worring, Santini, Gupta, & Jain, 2000). There is a great interest and a wealth of promise in CBIR as an emerging technology. Comprehensive surveys exist on both early low-level and recent semantic-level image retrieval techniques (Smeulders et al., 2000; Liua, Zhang, Lu, & Ma, 2007; Datta, Joshi, Li, & Wang, 2008). However, even after almost two decades of intensive research, CBIR systems still lag behind the best text-based search engines of today. One of the leading causes for this is the well known *"semantic gap"* problem (Smeulders et al., 2000), which is the mismatch between human understanding of image content consisting of semantic concepts (high level) and image representation through fairly rudimentary features (low level). To narrow the semantic gap, techniques have been developed over the last few years for semantic image classification and retrieval, which utilize various off-line and on-line machine learning-based methods (Liua et al., 2007; Datta et al., 2008; Bosch, Muoz, & Mart, 2007; Zhou & Huang, 2003; Rui et al., 1997).

Among these, a number of RF-based approaches have been proposed in the CBIR domain, such as: query point movement, feature re-weighting, active learning, etc. (Zhou & Huang, 2003; Rui et al., 1997; Tong & Chang, 2001; Ishikawa et al., 1998). The majority of these approaches estimate the ideal query parameters from the low-level image features at a global level. However, for the high-level concepts or semantics that cannot be adequately represented by low-level features, these systems often do not return many relevant results even with a large number of feedback iterations. To address this limitation, some recent systems incorporate the RF approach on both low-level content and high-level context (keyword) based feature spaces for Web image retrieval (Zhang, Chen, Li, & Su, 2003; Hu, Zhu, Zhang, & Yang, 2000). Although the RF algorithms provide a performance boost, one of the major drawbacks is that the users do not always provide enough feedback information to make the methods perform effectively. It has often been found to be a complex mechanism with different levels of feedback, such as the relevance levels in the MARS (Rui et al., 1997) and the goodness scores in the Mindreader (Ishikawa et al., 1998) systems.

The automatic query reformulation based on term co-occurrence or term similarity also has been widely investigated in the text retrieval domain with varying degrees of success (Xu & Croft, 2000; Xu & Croft, 1996; Crouch, 1990; Attar & Fraenkel, 1977; Qiu & Frei, 1993). These approaches have a significant advantage over interactive RF as they require no effort on the part of the user. The techniques that have so far been investigated can be described as being based on either global or local analysis (Xu & Croft, 2000; Xu & Croft, 1996). In a global analysis approach, all the documents in the collection are analyzed to determine a global thesaurus-like structure that defines certain term relationships. This structure is then used to select additional terms for the query expansion. In local analysis, the top retrieved documents for a query are examined (usually without any assistance from the user) at query time to determine the terms for the query expansion (Xu & Croft, 1996; Crouch, 1990).

Due to the nature of the low-level continuous feature representation, it is difficult to perform query reformulation or expansion based on term co-occurrence or the term similarity in the CBIR domain. The obvious choice for terms in text documents are the words that are directly related to the semantic contents. However, the images contain information in a very dense and continuous form where the units of an image at a low level are pixel values having no direct semantic meaning. In order to perform an automatic query expansion based on term-term correlations, the images should have a similar feature type and distribution as the ones from the text. Therefore, the number of possible image features should be high, and the number of image features per image should be low. Recently, the ongoing success of the *"bag-of-words"* methodology in the text retrieval domain have motivated researchers to explore analogous techniques mainly for scene classification and retrieval (Bosch et al., 2007). The models are applied to images by using a visual analogue of a word by automatically extracting different predominant color or texture patches or semantic patches, such as, water, sand, sky, cloud, etc. in natural photographic images. This intermediate semantic level representation is introduced as a first step to deal with the semantic gap between low-level features and high-level concepts.

Recent works have shown that local features represented by *"bags-of-words"* are suitable for scene classification, yielding impressive levels of performance (Zhu, Zhang, Rao, & Srihari, 2002; Lim, 2001; Jing, Li, Zhang, & Zhang, 2004; Vogel & Schiele, 2007; Shi, Feng, Chua, & Lee, 2004). For example, a framework to automatically generate the visual terms (''keyblock''), proposed in Zhu et al. (2002), applies a vector quantization or clustering technique. It represents images similar to the *"bags-of-words"* representation in correlation-enhanced feature spaces based on the *n*-gram and bi-gram models (Yates & Neto, 1999). For the reliable identification of image elements, the work in Lim (2001) manually identifies visual patches (''visual keywords'') from the sample images. In Jing et al. (2004), a compact and sparse representation of images is proposed based on using a region codebook generated by a clustering technique. A semantic modeling approach, investigated in Vogel and Schiele (2007) for a small collection of images, is based on the binary classification of semantic patches of local image regions. Similarly, in Shi et al. (2004) a binary support vector machine (SVM) is employed for the model generation of 23 selected concepts. At the testing stage, unlabeled regions are fed into all the models, and the concept from the model giving the highest positive result is associated with the region. A learning-based unified image retrieval framework to represent images in local visual and semantic concept-based feature spaces is presented in our previous work (Rahman, Bhattacharya, & Desai, 2009). In this framework, a visual concept vocabulary (codebook) is automatically constructed by utilizing a self-organizing map (SOM), and statistical models are built for local semantic concepts using a probabilistic multi-class support vector machine (SVM). A detailed review and accuracy comparison of some of the most commonly used scene classification approaches has recently been presented in Bosch et al. (2007). Experimental evaluation has shown that approaches using intermediate semantic concepts are more appropriate to reduce the gap between low and high-level.