Neurocomputing 153 (2015) 278–285

Contents lists available at ScienceDirect

Neurocomputing

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

Social images tag ranking based on visual words in compressed domain

Jing Zhang*, Xin Liu, Li Zhuo, Chao Wang

Signal and Information Processing Laboratory, Beijing University of Technology, 100 Ping Le Yuan, Chao Yang District, Beijing 100124, China

ARTICLE INFO

Article history:
Received 12 July 2014
Received in revised form
9 October 2014
Accepted 14 November 2014
Communicated by: Dr. M. Wang
Available online 24 November 2014

Keywords:
Social images
Tag ranking
Visual words
Compressed domain
Neighbor voting

ABSTRACT

With the introduction of many image compression standards, the social images are stored and transmitted in compressed formats such as JPEG. For large-scale image database, tag ranking must fully decompress the compressed data to predict tag relevance based on visual content. In order to improve the accuracy of tag ranking and further reduce the ranking time, social images tag ranking based on visual words in compressed domain is proposed in this paper, which includes three steps: (1) low-resolution social images are constructed from the compressed image data; (2) visual words are created according to extracted SIFT descriptors in low-resolution social image; (3) the neighbor voting model is utilized to rank the image tags after matching the similarity based on visual words of an image. In order to evaluate the performance of the proposed method, average *NDCG* (normalized discounted cumulative gain) and tag ranking time are compared. Experimental results show that the proposed method can significantly reduce the time of image tag ranking under ensuring the ranking accuracy of social image tags.

1. Introduction

With the rapid development of digital images and social media sharing web sites, people have witnessed an explosion of social media in many image sharing websites like Flickr. Many users are allowed to upload images and tag them with tags. One of the advantages is the websites have created a framework to significantly enhance user's ability to understand social images.

Multimedia analysis and retrieval has attracted a lot of attention recently, and extensive research efforts have been dedicated to topics related to web image search in the past years [1–6]. Different from these web images, social images can be indexed with user contributed tags. However, tags in the social image are known to be ambiguous, limited in terms of completeness, and overly personalized [7]. It reveals that many tags provided by Flickr users are imprecise and there are only around 50% tags actually related to the image [8], which limits the application of tags in web image search and browsing. Thus, it is very important to learn the relevance between textual tags and the visual image in order to accurately rank image tags, which can be applied to further narrow the semantic gap in content based image annotation and retrieval.

In recent years, the technology of social images tag ranking based on visual words has attracted more and more attention. Inspired by the text content analysis, researchers have viewed an image as a word combination (visual words), and applied text analysis method to image classification, scene analysis and image search, which has achieved remarkable results [1–5,7,9–13]. And among them, hypergraph learning has achieved excellent performance in many applications. In Ref. [7], a social image hypergraph is constructed, in which the visual and text terms are combined to represent its hyperedges. After hypergraph learning, the tag's relevance estimation and tag-based social image search can be achieved.

For accurately ranking image tags, many tag ranking methods have been explored to automatically predict tag's relevance according to the visual content. And most of works adopt supervised machine learning methods [14–18]. In general, the methods heavily rely on learning a mapping between low-level visual features and high-level semantic concepts. Since the number of training examples is limited by the supervised methods, which are not scalable to cover the unlimited concepts existing in social tagging. Moreover, uncontrolled visual content contributed by users has significant diversity in visual concept. Thus, the scarcity of training examples and the significant diversity in visual concept might make the learned models difficult to realize [19]. Intuitively, if different individuals tag visually similar images using the same tags, these tags are likely to reflect objective content of the visual image. Based on this idea, the Ref. [19]

* Corresponding author. Tel.: +86 10 67392799; mobile: +86 15911139568.
E-mail address: zhj@bjut.edu.cn (J. Zhang).

http://dx.doi.org/10.1016/j.neucom.2014.11.027
0925-2312/

proposed an algorithm that learns tag relevance by accumulating votes from visually similar neighbors, which effectively improves the accuracy of image tags. Meanwhile, the model doesn't require making any model training.

With the introduction of many image compression standards, the network images are stored and transmitted in compressed formats such as JPEG. Since traditional image processing method must fully decompress the compressed data, it will be time consuming in the pixel domain [20]. Especially for such a large-scale image database, tag ranking must fully decompress the compressed data to predict tag relevance based on visual content. Obviously, tag-ranking in pixel domain will be time-consuming. An effective method is to process image data in compressed domain, which can make full use of image compression processing and characteristics of the compressed data [21]. Suresh [22] presented a technique to index an image using DC and AC coefficients to extract features from chrominance and luminance planes separately. In Ref. [23], a method to resize images was presented in compressed domain. The method utilized the spatial relationship of DCT coefficients between a block and its sub-blocks. In our previous work, much research about compressed domain has been explored [20–21,24–26]. Based on these studies, considering the compressed social images, social images tag ranking based on visual words in compressed domain is proposed in order to reduce the tag ranking time under ensuring the accuracy of social image tags.

The remainder of this paper is organized as follows. Section 2 introduces the proposed social images tag ranking based on visual words in compressed domain. Section 3 presents the experimental results. Finally, conclusions and future work are given in Section 4.

2. Social images tag ranking based on visual words in compressed domain

Firstly, we will briefly introduce the image processing in the compressed domain, which refers to process the compressed data directly without full decoding. The image processing in JPEG compressed domain is shown in Fig. 1, in which "Non-decompressed process" corresponds to the entropy decoder, "The partly decompressed process" is after entropy decoding or before dequantizing.

In this paper, compressed domain refers to construct low-resolution (LR) images from the compressed data directly before inverse Discrete Cosine Transform (IDCT) as denoted with the partly decompressed process in Fig. 1, and to perform extraction of SIFT descriptor from low-resolution images, see our previous work [21].

The proposed social images tag ranking based on visual words in compressed domain is mainly divided into two parts: the first part is visual words creation in compressed domain; the second part is tag ranking based on neighbor voting. Fig. 2 shows the system diagram.

2.1. Visual words creation for social images in compressed domain

Visual words are commonly generated by clustering a large amount of image local features such as SIFT descriptor. After that each cluster center is taken as a visual word, and a corresponding visual word database is generated. The proposed visual word creation in compressed domain consists of three steps: (1) constructing LR images from compressed data; (2) extracting SIFT descriptors; (3) K-means clustering to create visual words. Fig. 3 illustrates the general process of visual word creation.

2.1.1. Low resolution social images construction in compressed domain

In order to reduce decoding time of JPEG image, low resolution (LR) images are constructed with the compressed data before Inverse Discrete Cosine Transform (IDCT).This method can obtain a $1/2 \times 1/2$ version of the original image.

The first sixteen DCT coefficients of zigzag sequence in every 8×8 DCT block are kept firstly, then be used for constructing 4×4 dimensional matrix A_n (n is the total number of blocks) after inverse quantization in the decoder. Then transfer it to the 4×4 dimensional matrix I_n by (1), and finally combine this calculated matrix I_n to construct $1/2 \times 1/2$ LR images [24].

$$I_n = CA_nC^T \tag{1}$$

where

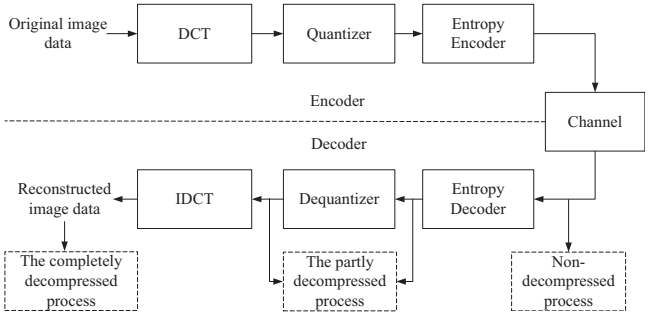

Fig. 1. Image processing in JPEG compressed domain.

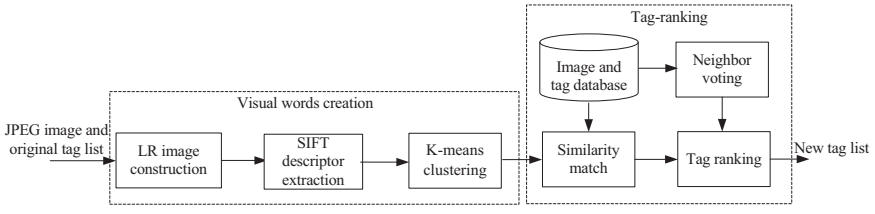

Fig. 2. Proposed tag ranking in compressed domain.

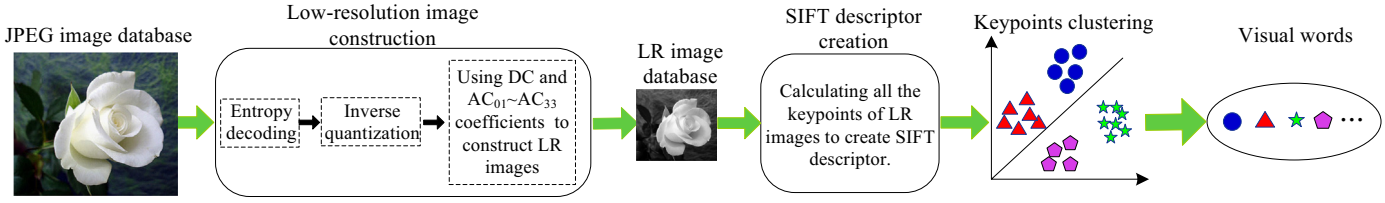

Fig. 3. An illustration of the visual words creation.