Neurocomputing 76 (2012) 50–60

Contents lists available at SciVerse ScienceDirect

Neurocomputing

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

Multimodal representation, indexing, automated annotation and retrieval of image collections via non-negative matrix factorization

Juan C. Caicedo [a], Jaafar BenAbdallah [b], Fabio A. González [a,*], Olfa Nasraoui [b]

[a] Computer Systems and Industrial Engineering Department, National University of Colombia, Cra 30 45 - 03, Ciudad Universitaria, Edif. 453, Of. 114. Bogotá, Colombia
[b] Department of Computer Engineering and Computer Science, University of Louisville, Louisville KY, USA

ARTICLE INFO

Available online 11 August 2011

Keywords:
Image collection analysis
Multimodal indexing
Multimodal representation
Image annotation
Content-based image retrieval
Non-negative matrix factorization
Latent semantic indexing

ABSTRACT

Massive image collections are increasingly available on the Web. These collections often incorporate complementary non-visual data such as text descriptions, comments, user ratings and tags. These additional data modalities may provide a semantic complement to the image visual content, which could improve the performance of different image content analysis tasks. This paper presents a novel method based on non-negative matrix factorization to generate multimodal image representations that integrate visual features and text information. The proposed approach discovers a set of latent factors that correlate multimodal data in the same representation space. We evaluated the potential of this multimodal image representation in various tasks associated to image indexing and search. Experimental results show that the proposed method highly outperforms the response of the system in both tasks, when compared to multimodal latent semantic spaces generated by a singular value decomposition.

1. Introduction

Most of the effort on web-content mining has concentrated on textual (and hyper-textual) data. However, visual information is an important component of the web content nowadays. In particular, the advent of the Web 2.0 has been accompanied by an explosion of multimedia content. Specialized sites, such as Flickr and Picassa, host billions of pictures uploaded by users. Other types of sites that allow users to upload visual content include: social networking sites, such as Facebook and MySpace, community-generated content, such as Wikipedia, and individual-generated content, such as in the blogosphere and Twitter.

The most salient characteristic of web image collections is that they come with a wide variety of associated data, such as text descriptions, tags, ratings and user comments. The availability of different sources of information brings the possibility to involve semantic evidence during the analysis of visual content in image collections, which is especially useful when considering the semantic gap [1], i.e., the discrepancy between visual features and semantic interpretations. Therefore, the combination of these data sources together with visual characteristics of images has received increasing attention from the research community in multimedia processing. The main problem is to take advantage of different data modalities to enable computer systems with the ability to make appropriate decisions according to the high-level task, which is known in the literature as *multimodal fusion* [2].

In this paper, we consider the problem of building a multimodal image representation that combines two data modalities: visual patterns extracted from images and text terms extracted from attached text descriptions. The proposed strategy mines the relationships between these two modalities to construct a unified representation based on Latent Semantic Analysis (LSA) principles. We propose a solution based on non-negative matrix factorization (NMF) to construct a latent-factor-based representation that can be spanned using text terms or visual features. We formulate a set of NMF-based algorithms for multimodal image analysis, which generates a joint visual–textual representation that is useful to approach different image analysis tasks.

The main contribution of our work is an NMF-based model to index multimodal data. In this work, multimodal collections are composed of images and some associated text descriptions. These image collections can be built from many different web sources, including Flickr and Picassa, in which several text descriptions for images can be identified using information extraction techniques. More details about the representation and preprocessing of each separate data modality are given in Section 3. Then, given a database of images with the corresponding text annotations, the multimodal analysis is performed using NMF algorithms as depicted in Fig. 1.

* Corresponding author. Tel.: +571 3165000x14077, +571 3165000x14011.
E-mail addresses: jccaicedoru@unal.edu.co (J.C. Caicedo), j0bena01@louisville.edu.(J. BenAbdallah)fagonzalezo@unal.edu.co (F.A. González), olfa.nasraoui@louisville.edu (O. Nasraoui).

0925-2312/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2011.04.037

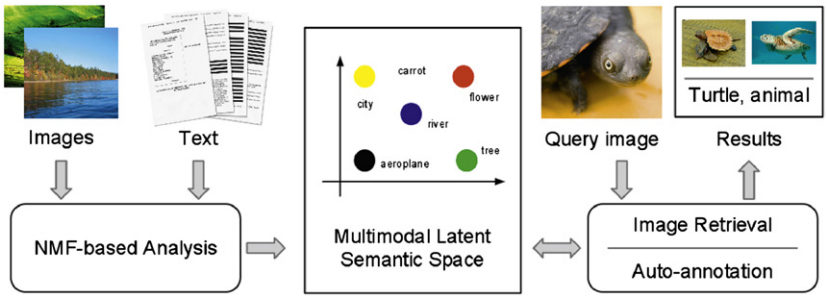

Fig. 1. Overview of the proposed approach. NMF-based algorithms process image and text features to generate a multimodal latent semantic space, in which both data modalities are represented together. Image retrieval and auto-annotation methods exploit the latent representation to find semantically related images and valid automatic annotations, respectively.

The NMF-based strategy generates the latent semantic space using both data modalities. The goal is to find a set of latent factors that explain the underlying structure of the collection and the relationships between multimodal features. The latent representation is computed using a training data set composed of objects exhibiting both modalities. New objects can later be projected to the latent semantic space even if they do not have both data modalities. In consequence, the multimodal latent semantic model can deal with images without text or text without images, which is particularly useful to address several image analysis and retrieval tasks. The proposed models and their properties are presented in Section 4.

We evaluate the proposed strategy on two different tasks to demonstrate the potential of the multimodal representation: image indexing and automatic image annotation. Our method projects the input visual features to the multimodal latent semantic space to allow the subsequent analysis. Thus, in addition, we develop a set of algorithms for image search and image auto-annotation that are presented in Section 5.

An experimental evaluation was conducted using two image collections: Corel 5k [3], a collection of photographs with several tags and categories, and MIRFlickr 25000 [4], a data set of images downloaded from Flickr.com with the corresponding user generated tags and some additional labels provided as ground truth. The experimental setup and results are presented in Section 6, which shows that our proposed model outperforms baseline strategies. The final discussion and concluding remarks are presented in Section 7. Portions of this work have been previously reported in [5,6].

2. Relation to previous work

The use of multiple data modalities for multimedia analysis has become an important research topic during the last years. A comprehensive survey of the many research aspects of multimodal fusion for automatic multimedia analysis can be found in [2], which includes applications in audio, image and video processing using multiple data sources to achieve semantic decisions. Also, in the particular field of image retrieval, Datta et al. [7] discussed the importance of multimodal fusion for image indexing. The construction of systems that make semantic decisions using heterogeneous data sources is the ultimate goal of multimodal fusion.

Two main strategies can be considered for combining multimodal information: late fusion and early fusion. Late fusion, also known as rank aggregation or fusion at a decision level, consists in processing each data source separately during the indexing phase, with the multimodal integration taking place during the query phase. The work of Ah-Pine et al. [8] is an example of similarity combination to achieve multimodal access in image collections, using pseudo-relevance feedback to re-rank images in different applications. On the other hand, early fusion, or fusion at a feature level, consists in modeling feature relationships to create a new multimodal representation, so that during the decision phase, the only task to do is usually analyzing multimodal features [9,10]. Our work is categorized as an early fusion strategy for multimodal image analysis.

Latent topic analysis has been used to model the relationships between multimodal data, specifically images and text annotations. A set of generative models that use latent variables have been proposed to predict missing captions given unlabeled images [11,12]. These works are based on extensions of the latent Dirichlet allocation (LDA) model, in which a set of hidden factors are assumed to explain the associations between the two data types. Later, Monay and Gatica-Perez [13] proposed a simplified aspect model based on probabilistic latent semantic analysis (PLSA) to index and annotate images by jointly processing visual features and text data.

More recent works follow a latent topic analysis using matrix factorization approaches. Hare et al. [14] proposed a linear algebraic technique based on singular value decomposition (SVD) to learn a semantic space for image features and textual descriptions. This method is a multimodal extension of latent semantic indexing (LSI) for image retrieval that results in a semantic space suitable for image search. Latent topic analysis using matrix factorization has recently drawn of wide interest in information retrieval and image analysis. In particular, NMF algorithms have been used to analyze visual data to discover object classes [15] and to find correlations between image tags [16]. Other applications of NMF decompositions for visual data include [17–19].

All past works are different from ours since they are focused on processing either visual features or text annotations rather than exploiting multimodal interactions between both data types. Our work is the first one, according to our knowledge, that addresses the multimodal indexing problem using an NMF-based algorithm. In [5,6], we addressed the problems of multimodal image indexing and automatic image annotation, respectively. The present paper builds upon these works by proposing a unified method for solving both problems, and performing a systematic and extended experimental evaluation.

3. Multimodal image collections

Assume an image collection with attached unstructured text annotations. An excerpt of text may be identified from the source document for each image using information extraction techniques to locate captions, to parse image names and tool-tips, among others [20]. After the information extraction step, each image has an associated unstructured set of text terms. In our model, it is