ELSEVIER

Internet cross-media retrieval based on deep learning ☆

CrossMark

Bin Jiang [a], Jiachen Yang [a,d,*], Zhihan Lv [b,*], Kun Tian [c], Qinggang Meng [d], Yan Yan [e]

[a] School of Electrical Automation and Information Engineering, Tianjin University, Tianjin, PR China
[b] Dept. of Computer Science, University College London, London WC1E 6EA, UK
[c] National Key Laboratory of Science and Technology on Aerospace Intelligence Control, Beijing, PR China
[d] Department of Computer Science, School of Science at Loughborough University, UK
[e] Department of Information Engineering and Computer Science, University of Trento, Italy

ARTICLE INFO

ABSTRACT

With the development of Internet, multimedia information such as image and video is widely used. Therefore, how to find the required multimedia data quickly and accurately in a large number of resources, has become a research focus in the field of information process. In this paper, we propose a real time internet cross-media retrieval method based on deep learning. As an innovation, we have made full improvement in feature extracting and distance detection. After getting a large amount of image feature vectors, we sort the elements in the vector according to their contribution and then eliminate unnecessary features. Experiments show that our method can achieve high precision in image-text cross media retrieval, using less retrieval time. This method has a great application space in the field of cross media retrieval.

1. Introduction

With the growth of Internet services, multimedia technology has a continuous development, which brings a tremendous increase in the amount of Internet data. At the same time, the number of digital images per day is also growing exponentially at an alarming rate, with the popularity of digital cameras or other devices, as shown in Fig. 1. How to make effective content retrieval and analysis from such a large number of images or texts, become the focus of scientific research [1].

Image retrieval technology is divided into the text based image retrieval (TBIR) technology and content based image retrieval (CBIR) technology [2,3]. The traditional image retrieval system is based on the images with manual text annotation, and it is keyword search. This kind of system consumes huge manpower and material resources on the massive image data sets, so the retrieval precision is greatly disturbed by the subjective factors [4,5]. The content based image retrieval technology is different, most of the content based systems use low-level visual features of the image itself, such as color, texture and shape. Some similarity evaluation methods are used to match the images using visual features, then the query image and database image are matched. Without manual intervention, speed and precision of the whole process have been greatly improved. The application field is very wide and can be used for the search of a particular picture on the network for common users, and can also be used for various types of professional organizations [6].

In this paper, we explore the cross media retrieval technology, texts are used to search related image information, images are used to search related text information, as shown in Fig. 2.

Cross media retrieval technology is a new technology field, it is related to the multidisciplinary cross and comprehensive, some fields are considered: pattern recognition, machine learning, image processing, video processing, speech recognition, data mining technology, agent of artificial intelligence and natural language processing.

However, unlike single media learning, cross media learning is still a new research direction of salary. At present, the relevant research is still relatively preliminary, there are still a lot of challenges and difficulties in the following areas need further research.

- Consistency description of cross media data
  Due to the different modes of cross media data, expression with different dimensions and different attributes cannot directly be the computing correlations. In addition to the heterogeneous characteristics, between the bottom of the content and high-level semantic gap, characteristics by the traditional method

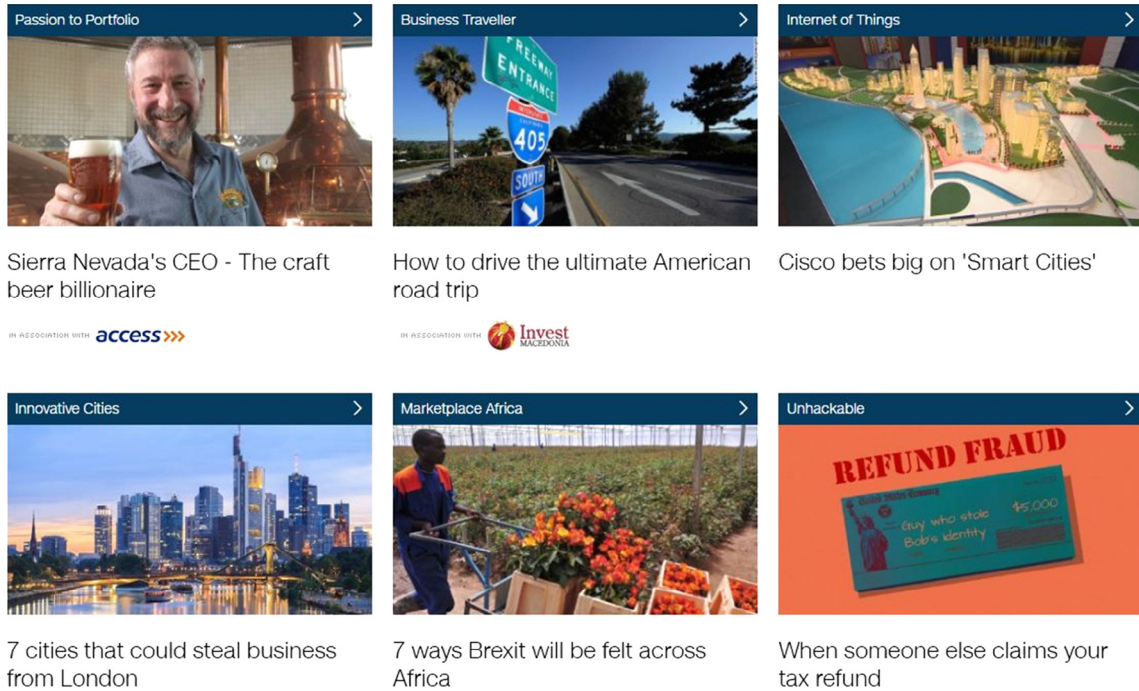B. Jiang et al. / J. Vis. Commun. Image R. 48 (2017) 356–366

357

Fig. 1. Internet Cross-media: we can see that there will be a lot of texts around the images.
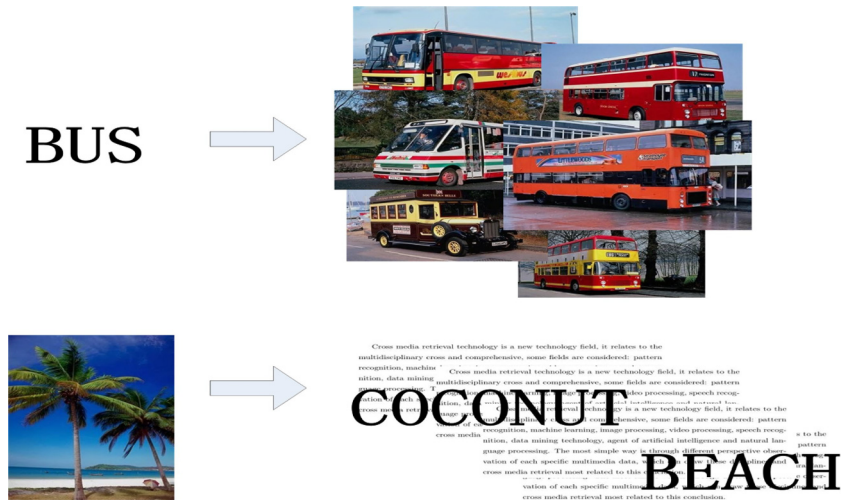

Fig. 2. Text query images and image query texts.

cannot be applied to the study of cross media for learning. Therefore, it is very necessary to excavate the consistency of cross media data and to realize the flexibility of different modes.

- Incremental learning across media

 The vast majority of cross media learning methods are involved in the singular value decomposition for the input matrix. Therefore, once to insert the new data to the original data set, the algorithm will spend more computation time is recalculated augmented the input data matrix singular value decomposition. Therefore, in order to avoid the high computational complexity, it is very important to propose an efficient incremental updating algorithm for singular value decomposition [7,8].

- The lack of modal complement

 With the rapid development of information technology, more and more cross media data appear in many applications, such as medical diagnosis, web page classification and cross media analysis. However, due to the high cost of data acquisition, the lack of authenticity and rejection and other reasons, these applications are faced with the problem of lack of mode. It is inevitable that the application of the full modality of these requirements can be handled with a very limited amount of data. So, the cross media applications need a cross media data deletion modal complement method [9,10].

- Correspondence between different modal descriptions

 In practice, there is noise in the description of some modes, which can destroy the corresponding relationship. The vast majority of applications in the real world (such as cross media retrieval and pattern recognition) need complementary descriptions from different models to obtain more accurate and robust estimation. So there is an urgent need to find a cross media denoising method, to re-establish the correspondence between the cross media description.