Information Sciences 302 (2015) 122–131

Contents lists available at ScienceDirect

Information Sciences

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

Localizing web videos using social images

Liujuan Cao [a], Xian-Ming Liu [b], Wei Liu [c], Rongrong Ji [a,*], Thomas Huang [b]

[a] School of Information Science and Engineering, Xiamen University, PR China
[b] Beckman Institute, University of Illinois at Urbana-Champaign, USA
[c] IBM T.J. Watson Research Center, USA

ARTICLE INFO

Article history:
Received 23 February 2013
Received in revised form 23 July 2014
Accepted 3 August 2014
Available online 16 September 2014

Keywords:
Web video analysis
Cross-domain
Social media
Landmark recognition
Classification

ABSTRACT

While inferring the geo-locations of web images has been widely studied, there is limited work engaging in geo-location inference of web videos due to inadequate labeled samples available for training. However, such a geographical localization functionality is of great importance to help existing video sharing websites provide location-aware services, such as location-based video browsing, video geo-tag recommendation, and location sensitive video search on mobile devices. In this paper, we address the problem of localizing web videos through transferring large-scale web images with geographic tags to web videos, where near-duplicate detection between images and video frames is conducted to link the visually relevant web images and videos. To perform our approach, we choose the trustworthy web images by evaluating the consistency between the visual features and associated metadata of the collected images, therefore eliminating the noisy images. In doing so, a novel transfer learning algorithm is proposed to align the landmark prototypes across both domains of images and video frames, leading to a reliable prediction of the geo-locations of web videos. A group of experiments are carried out on two datasets which collect Flickr images and YouTube videos crawled from the Web. The experimental results demonstrate the effectiveness of our video geo-location inference approach which outperforms several competing approaches using the traditional frame-level video geo-location inference.

1. Introduction

With the explosive growth of social networks, massive and heterogeneous multimedia data available in the Web provides a unique opportunity to bridge digital corpus and our physical world, which has become the mainstream of the ongoing multimedia research. To facilitate accurate recommendations by exploiting social media knowledge, web media tagging is also becoming an emerging important research direction [27,26,25,28,9].

Despite the widely studied semantic tagging scenario, annotating geographical locations of social media recently arises to be a major trend, which includes plenty of novel research subjects such as landmark retrieval and recognition, visual tour guide, geography-aware image search, mobile device localization, and aiding virtual reality.

Geographically tagging user-contributed images has been investigated in recent years [16,13]. However, tagging web videos has been less explored so far, in which inferring geo-locations of web videos is especially vital to location-based video services. The key difficulty of video geo-location inference is twofold. On the device level, current mobile phones and digital

* Corresponding author.

http://dx.doi.org/10.1016/j.ins.2014.08.017
0020-0255/

(a) An YouTube video clip about the "Golden State Bridge", which is split by the scenes of interviews.

(b) The thumbnails of the returned YouTube videos searched by the query words placed under them

Fig. 1. Examples of the challenging videos collected from YouTube for the evaluation of web video localization. (a) The video clip contains multiple scenes of irrelevant visual content. (b) A large proportion of video frames are usually irrelevant to the video tag.

cameras typically do not record the geo-information when shotting videos; on the other hand, there are limited geographic tags available for web videos, so it may be infeasible to train accurate classifiers or search engines in order to geographically tag the web videos.

The previous attempts of web video geo-location inference almost resorted to using metadata of social networks, such as titles, descriptions, and comments [17]. To improve the inference accuracy, extra evidence, *e.g.*, visual and acoustic features, is incorporated [15]. However, limited work has successfully exploited visual content to assist the geo-location inference of web videos, mainly due to insufficient training examples that cause a challenge in effectively modeling the landmark appearances. This challenge originates from two aspects: (1) the low quality of web videos, which results in a limited amount of SIFT features being extracted and therefore compromises the accuracy of near-duplicate visual content matching; (2) the difficulty to diversify different scenes in the same video, which prevents assigning the correct geographic tags to the corresponding locations or physical scenes. We illustrate the challenge in Fig. 1.

In this paper, we propose to tackle this challenge from a novel transfer learning perspective, *i.e.*, transferring an accurate and easy-to-learn video geo-tagging model from the image domain. Nowadays, there is an increasing amount of geo-tagged images available on the Web. Such massive image data prompts us to "transfer" the geo-tags from web images to web videos.

To do knowledge transfer across the image and video domains, we need to address two major issues: first, the tags of web images are usually noisy; second, the visual features of images and videos appear quite different due to their large variations. We address the first issue by proposing a *web image trustworthy* measurement to remove the "untrustworthy" web images. Afterwards, we perform a view-specific spectral clustering over the images of a given landmark to diversify different "views" of a single location. To build an effective video location inference model, a novel search-based transfer learning algorithm is proposed by constructing an AdaBoost [6] classifier in each view, and the outputs of multi-view AdaBoost classifiers are then combined into an overall landmark recognition model. In addition, we incorporate the temporal consistency to further improve the inference accuracy, which leverages the fact that temporally related video frames within the same video shot are more likely to be captured from the same view of a landmark or location.

To verify the proposed video geo-location inference model, we collect more than 50,000 geo-tagged images from Flickr[1] and 2000 video clips from YouTube,[2] respectively. Experimental comparisons on the two datasets show that our method achieves remarkable improvements over various competing methods. Besides, the proposed method can easily be integrated into the applications involving geo-location based web video browsing.

The rest of this paper is organized as follows: Section 2 briefly introduces the related work. The overall framework of the proposed method is presented in Section 3. Section 4 builds the effective landmark model using social images, which is further transferred to the video domain as described in Section 5. We conduct the experimental validations in Section 6. Finally, we conclude the paper and discuss the future work in Section 7.

2. Related work

Geography related image analysis has attracted intensive research attention from both academia and industry. One of the pioneering work comes from Kennedy et al. [16], which analyzed geographic tags (*e.g.*, landmarks) of millions of Flickr

[1] www.flickr.com.
[2] www.youtube.com.