Computer Vision and Image Understanding 140 (2015) 58–67

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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

A data-driven approach for tag refinement and localization in web videos☆

Lamberto Ballan[a,c,*], Marco Bertini[a], Giuseppe Serra[b], Alberto Del Bimbo[a]

[a] Media Integration and Communication Center (MICC), Università degli Studi di Firenze, Viale Morgagni 65, 50134 Firenze, Italy
[b] Dipartimento di Ingegneria "Enzo Ferrari", Università degli Studi di Modena e Reggio Emilia, Via Vignolese 905/b, 41125 Modena, Italy
[c] Computer Science Department, Stanford University, 353 Serra Mall, Stanford, CA 94305, United States

ARTICLE INFO

Article history:
Received 2 July 2014
Accepted 30 May 2015
Available online 5 June 2015

Keywords:
Video tagging
Web video
Tag refinement
Tag localization
Social media
Data-driven
Lazy learning

ABSTRACT

Tagging of visual content is becoming more and more widespread as web-based services and social networks have popularized tagging functionalities among their users. These user-generated tags are used to ease browsing and exploration of media collections, e.g. using tag clouds, or to retrieve multimedia content. However, not all media are equally tagged by users. Using the current systems is easy to tag a single photo, and even tagging a part of a photo, like a face, has become common in sites like Flickr and Facebook. On the other hand, tagging a video sequence is more complicated and time consuming, so that users just tag the overall content of a video. In this paper we present a method for automatic video annotation that increases the number of tags originally provided by users, and localizes them temporally, associating tags to keyframes. Our approach exploits collective knowledge embedded in user-generated tags and web sources, and visual similarity of keyframes and images uploaded to social sites like YouTube and Flickr, as well as web sources like Google and Bing. Given a keyframe, our method is able to select "on the fly" from these visual sources the training exemplars that should be the most relevant for this test sample, and proceeds to transfer labels across similar images. Compared to existing video tagging approaches that require training classifiers for each tag, our system has few parameters, is easy to implement and can deal with an open vocabulary scenario. We demonstrate the approach on tag refinement and localization on DUT-WEBV, a large dataset of web videos, and show state-of-the-art results.

1. Introduction

Over the past recent years social media repositories such as Flickr and YouTube have become more and more popular, allowing users to upload, share and tag visual content. Tags provide contextual and semantic information which can be used to organize and facilitate media content search and access. The performance of current social image and video retrieval systems depends mainly on the availability and quality of tags. However, these are often imprecise, ambiguous and overly personalized [1]. Tags are also very few (typically three tags per image, on average) [2], and their use may change over time, following the creation of new folksonomies created by users. Another issue to be considered is the 'web-scale' of data, that calls for efficient and scalable annotation methods.

Many efforts have been done in the past few years in the area of content-based tag processing for social images [3,4]. The main focus of these works has been put on three aspects: *tag relevance* (or *ranking*) [5], *tag refinement* (or *completion*) [6] and *tag-to-region localization* [7]. Among the others, nearest-neighbor based approaches have attracted much attention for image annotation [8–11], tag relevance estimation [12] and tag refinement [13]. Here the key idea is that if different users label similar images with the same tags, these tags truly represent the actual visual content. So a simple voting procedure may be able to transfer annotations between similar images. This tag propagation can be seen as a lazy local learning method in which the generalization beyond the training data is deferred until test time. A nice property of this solution is that it naturally adapts to an open vocabulary scenario in which users may continuously add new labels to annotate the media content. In fact, a key limitation of the traditional methods in which classifiers are trained to label images with the concept represented within, is that the number of labels must be fixed in advance. More recently, some efforts have been made also to design methods to automatically assign the annotated labels at image level to those derived semantic regions [7,14,15]. A relevant example is the work of Yang et al. [14] in which the encoding

☆ This paper has been recommended for acceptance by C.V. Jawahar.

* Corresponding author at: Media Integration and Communication Center (MICC), Università degli Studi di Firenze, Viale Morgagni 65, 50134 Firenze, Italy.
E-mail addresses: lamberto.ballan@unifi.it (L. Ballan), marco.bertini@unifi.it (M. Bertini), giuseppe.serra@unimore.it (G. Serra), alberto.delbimbo@unifi.it (A.D. Bimbo).

http://dx.doi.org/10.1016/j.cviu.2015.05.009
1077-3142/

Fig. 1. Example of video tag localization: *top)* YouTube video with its related tags; *bottom)* localization of tags in keyframes.

ability of group sparse coding is reinforced with spatial correlations among regions.

The problem of *video tagging* so far has received less attention from the research community. Moreover, typically it has been considered the task of assigning tags to whole videos, rather than that of associating tags to single relevant keyframes or shots. Most of the recent works on web videos have addressed problems like: *i) near duplicate detection*, applied to IPR protection [16,17] or to analyze the popularity of social videos [18]; *ii) video categorization*, e.g. addressing actions and events [19,20], genres [21] or YouTube categories [22]. However, the problem of video tagging "in the wild" remains open and it might have a great impact in many modern web applications.

In this paper, the proposed method aims at two goals: to extend and refine the video tags and, at the same time, associate the tags to the relevant keyframes that compose the video, as shown in Fig. 1. The first goal is related to the fact that the videos available on media sharing sites, like YouTube, have relatively few noisy tags that do not allow to annotate thoroughly the content of the whole video. Tackling this task can be viewed also as an application of image tag refinement to video keyframes [4,6]. The second goal is related to the fact that tags describe the global content of a video, but they may be associated only to certain shots and not to others. Our approach takes inspiration from the recent success of nonparametric data-driven approaches [8,23–25]. We build on the idea of nearest-neighbor voting for tag propagation, and we introduce a temporal smoothing strategy which exploits the continuity of a video. Compared to existing video tagging approaches in which classifiers are trained for each tag, our system has few parameters and does not require a fixed vocabulary. Although the basic idea has been previously used for image annotation, this is the first attempt to extend this idea to video annotation and tag localization.

Our contributions can be summarized as follows:

- We propose an automatic approach that locates the temporal positions of tags in videos at keyframe level. Our method is based on a lazy learning algorithm which is able to deal with a scenario in which there is no pre-defined set of tags.
- We show state-of-the-art results on DUT-WEBV, a large dataset for tag localization in web videos. Moreover, we report an extensive experimental validation about the use of different web sources (Flickr, Google, Bing) to enrich and reinforce the video annotation.
- We show how the proposed approach can be applied in a real-world scenario to perform open vocabulary tag annotation. To evaluate the results, we collected more than 5000 frames from 40 YouTube videos and three individuals to manually verify the annotation.

2. Related work

Probably the most important effort in semantic video annotation is TRECVID [26], an evaluation campaign with the goal to promote progress in content-based retrieval from digital video archives. Recently, online videos have also attracted the attention of researchers [22,27–30], since millions of videos are available on the web and they include rich metadata such as title, comments and user tags.