Expert Systems with Applications 42 (2015) 488–502

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Temporal segmentation and keyframe selection methods for user-generated video search-based annotation

Iván González-Díaz *, Tomás Martínez-Cortés, Ascensión Gallardo-Antolín, Fernando Díaz-de-María

Department of Signal Theory and Communications, Universidad Carlos III de Madrid, Leganés 28911, Madrid, Spain

ARTICLE INFO

Article history:
Available online 17 August 2014

Keywords:
User Generated Video
Video annotation
Video temporal segmentation
Camera motion analysis
Keyframe selection

ABSTRACT

In this paper we propose a temporal segmentation and a keyframe selection method for User-Generated Video (UGV). Since UGV is rarely structured in shots and usually user's interest are revealed through camera movements, a UGV temporal segmentation system has been proposed that generates a video partition based on a camera motion classification. Motion-related mid-level features have been suggested to feed a Hierarchical Hidden Markov Model (HHMM) that produces a user-meaningful UGV temporal segmentation. Moreover, a keyframe selection method has been proposed that picks a keyframe for fixed-content camera motion patterns such as *zoom*, *still*, or *shake* and a set of keyframes for varying-content *translation* patterns.

The proposed video segmentation approach has been compared to a state-of-the-art algorithm, achieving 8% performance improvement in a segmentation-based evaluation. Furthermore, a complete search-based UGV annotation system has been developed to assess the influence of the proposed algorithms on an end-user task. To that purpose, two UGV datasets have been developed and made available online. Specifically, the relevance of the considered camera motion types has been analyzed for these two datasets, and some guidelines are given to achieve the desired performance-complexity tradeoff. The keyframe selection algorithm for varying-content *translation* patterns has also been assessed, revealing a notable contribution to the performance of the global UGV annotation system. Finally, it has been shown that the UGV segmentation algorithm also produces improved annotation results with respect to a fixed-rate keyframe selection baseline or a traditional method relying on frame-level visual features.

1. Introduction

The amount of multimedia content that is generated daily has dramatically grown during recent years. This is particularly true in the case of User Generated Content (UGC), due to the massive access of users to mobile devices with recording capabilities (Cricri, Dabov, Curcio, Mate, & Gabbouj, 2011). Consequently, algorithms providing automatic content annotation and content-based search are more and more demanded by both multimedia hosting services and users.

Although the automatic annotation problem has been traditionally posed as that of object/concept recognition (Deng et al., 2009; Everingham, Van Gool, Williams, Winn, & Zisserman,; Smeaton, Over, & Kraaij, 2006), this approach has not yet reached a suitable solution due to the large amount of visual concepts to detect, including not only general visual categories such as car, street, or chair, but also particular places, people, artworks, and other objects of special interest for users.

Alternatively, the problem of content annotation can be approached by taking advantage of valuable user-provided metadata (tags, titles, and descriptions) that are available through online repositories such as Panoramio,[1] Flickr[2] or Picasa.[3] Such a vast amount of (noisy) annotated contents opens the possibility of annotating a particular image or video by propagating tags from visually similar content. This approach has been referred to as search-based annotation in the literature (Wang, Zhang, Jing, & Ma, 2006; Wang, Zhang, Liu, Li, & Ma, 2010; Wang, Zhang, & Ma, 2012b).

Most successful methods make use of some kind of contextual information to preselect a candidate set of images/videos that show some aspect in common with the query content. In Soderberg and Kakogianni (2010) a set of tags is suggested by

* Corresponding author.
E-mail addresses: igonzalez@tsc.uc3m.es (I. González-Díaz), tmcortes@tsc.uc3m.es (T. Martínez-Cortés), gallardo@tsc.uc3m.es (A. Gallardo-Antolín), fdiaz@tsc.uc3m.es (F. Díaz-de-María).

[1] http://www.panoramio.com/.
[2] http://www.flickr.com/.
[3] http://www.picasa.com/.

http://dx.doi.org/10.1016/j.eswa.2014.08.001
0957-4174/

combining the context in which the photo was captured with prior knowledge about popular annotation concepts. In Moxley, Kleban, and Manjunath (2008) GPS coordinates are used together with image features to propose labels that are chosen by considering both geographical distances and visual similarities. Similarly, in Lee, Yang, and Wang (2011) a Fuzzy ARTMAP network was used to map images and their visual features to geographic nouns. The main disadvantage of these methods is that they are not applicable to non-geolocated contents, which has resulted in the creation of mechanisms for automatic geotagging (Sevillano, Valero, & Alias, 2012; Schindler, Krishnamurthy, Lublinerman, Liu, & Dellaert, 2008).

Although the initial approaches for search-based annotation were restricted to images, in the last few years some effort has been directed towards video content and, in particular, towards the development of methods that exploit redundancy among videos. In Siersdorfer, San Pedro, and Sanderson (2009) a system combining video copy detection and tag propagation techniques used redundancy between videos as a key to annotate new ones. The work in Shang, Yang, Wang, Chan, and Hua (2010) focused on real-time video retrieval over large-scale web datasets by developing efficient spatio-temporal features. In Li et al. (2011a), the authors proposed a system that relied on user global tags to further analyze the video content at shot level. The approach in Tang, Sukthankar, Yagnik, and Fei-Fei (2013) went beyond and, besides identifying the particular segments associated with a tag, also generated spatio-temporal segmentations of the object representing the tag. The work in Li et al. (2011a) was later extended in Wang et al. (2012a) to detect events and automatically generate video summaries. Finally, Ulges, Schulze, Keysers, and Breuel (2008) proposed a system that identified relevant frames in a video from its global tags and then used these frames to train concept detectors.

All the mentioned methods focus on similar aspects of the annotation system such as the development of robust and efficient visual search methods for near-duplicate contents, the application of these search methods to the more challenging tasks of video segment alignment and matching, the design of methods for tag propagation, or the deployment of systems for large-scale datasets with even billions of images. We have found, however, that little or no effort has been devoted to study how the video temporal segmentation and the subsequent keyframe selection affect the video annotation performance. In fact, all aforementioned methods employ very basic techniques to select the frames being analyzed: they run a shot-boundary detector to identify abrupt cuts in videos and then represent each shot by means of one keyframe, normally sampled at the middle of the temporal segment.

As we will discuss in the section devoted to related work, although several methods can be found in the literature proposing smart techniques for video segmentation, all of them have been assessed just in terms of segmentation quality, thus obviating how they may influence subsequent end-user tasks, such as video content annotation.

In this paper, we present a video segmentation algorithm that analyzes the camera motion using a Hierarchical Hidden Markov Model (HHMM) and provides a fine-grain temporal segmentation of the video content. Moreover, a strategy for keyframe selection is proposed that considers a camera motion-based model of the interests of the person who is recording the video. Finally, we embed these subsystems on a complete system for automatic annotation of User Generated Video (UGV) and prove that our model for video segmentation not only achieves successful segmentation results, but also contributes to improve the performance of a high-level tasks such as specific object/place recognition and search-based video annotation.

Furthermore, as a by-product of our experimental evaluation, two video datasets for specific object/place recognition have been developed and made publicly available, which also becomes an important contribution of our work, and hopefully will help future developments in the field.

The rest of the paper is organized as follows. Section 2 introduces related work on temporal video segmentation and keyframe selection. Section 3 explains in detail the proposed method for automatic video segmentation and keyframe selection. Section 4 describes the complete system for video annotation. Section 5 is devoted to the experimental results, assessing both the segmentation performance and its impact on a higher-level task. Finally, Section 6 summarizes our conclusions and outlines some future work directions.

2. Related Work

Temporal video segmentation aims to split a video sequence into homogeneous subsequences, in such a manner that the properties of each subsequence are different enough from those of its temporal neighbors. When dealing with edited video, most temporal segmentation techniques rely on shot boundary detection, which entails detecting both abrupt or gradual changes in the video and/or audio signal properties (Smeaton, Over, & Doherty, 2010; Yuan et al., 2007).

By contrast, User Generated Videos are usually continuous recordings taken with a mobile phone or a digital camcorder, where (frequently) only one shot is present. Thus, in order to divide UGVs into meaningful semantic units, segmentation must be performed at sub-shot level. According to the definition given by Petersohn (2009), a sub-shot is *an unbroken sequence of frames within a shot only having a small variation in visual content*. Some sub-shot detection methods are based on the comparison of color histograms between video frames (Cahuina & Camara Chavez, 2013; Petersohn, 2009). However, since the type of camera motion (such as pan, tilt, or zoom) can be an indicator of the user's interests in the scene, and therefore of the video content, recently, several temporal video segmentation techniques have been proposed which use features derived from the camera motion information. In this sense, (Abdollahian, Taskiran, Pizlo, and Delp, 2010) define the so-called *camera view* as the basic unit of UGV.

Camera motion-based segmentation approaches involve, as first stage, the extraction of a set of features that allow for discriminating among the different types of camera motion considered. Typically used features include region-based correlation between consecutive frames (Aggarwal, Prakash, & Sofat, 2008), parameters derived from a 2D affine motion model (Bouthemy, Gelgon, & Ganansia, 1999; Mei, Tang, Tang, & Hua, 2013), motion vectors (Abdollahian et al., 2010), or even parameters provided by auxiliary motion-sensors, such as accelerometers (Cricri et al., 2011).

Once the relevant features have been extracted, the temporal segmentation can be approached in different ways. In some works a simple thresholding method is used to detect the different camera motions (Bouthemy et al., 1999; Luo, Papin, & Costello, 2009; Mei et al., 2013). The main drawback of this method being the difficulty to find threshold values suitable for all kinds of video sequences. On the contrary, supervised machine learning methods, such as Support Vector Machines (SVM) or Hidden Markov Models, do not require any threshold adjustments. A SVM-based segmentation method was proposed in Abdollahian et al. (2010), where binary SVMs are used to classify the camera motion of each video frame, and the final segmentation is obtained by grouping together neighboring frames that exhibit the same type of camera motion. HMMs has been used for shot detection and segmentation due to their ability for modeling time varying sequences (Bae, Jin, & Ro, 2004; Zhang, Lin, Chen, Huang, & Liu, 2006). Nevertheless, the complexity of multimedia data might make HMM not suitable