Computer Vision and Image Understanding 117 (2013) 660–669

Contents lists available at SciVerse ScienceDirect

Computer Vision and Image Understanding

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

A spatio-temporal pyramid matching for video retrieval ☆

Jaesik Choi [a,1], Ziyu Wang [b], Sang-Chul Lee [b,2], Won J. Jeon [c,*]

[a] University of Illinois at Urbana-Champaign, 201 N. Goodwin Avenue, Urbana, IL 61801, USA
[b] Inha University, 1103 High-tech Center, Yonghyun-dong 253, Nam-gu, Incheon, Republic of Korea
[c] Samsung Research America - Silicon Valley, 75 West Plumeria Drive, San Jose, CA 95134, USA

ARTICLE INFO

Article history:
Received 3 October 2011
Accepted 14 February 2013
Available online 27 February 2013

Keywords:
Video retrieval
Query by video clip
High-activity videos
Sport videos
Pyramid matching
Spatio-temporal pyramid matching

ABSTRACT

An efficient video retrieval system is essential to search relevant video contents from a large set of video clips, which typically contain several heterogeneous video clips to match with. In this paper, we introduce a content-based video matching system that finds the most relevant video segments from video database for a given query video clip. Finding relevant video clips is not a trivial task, because objects in a video clip can constantly move over time. To perform this task efficiently, we propose a novel video matching called *Spatio-Temporal Pyramid Matching (STPM)*. Considering features of objects in 2D space and time, STPM recursively divides a video clip into a 3D spatio-temporal pyramidal space and compares the features in different resolutions. In order to improve the retrieval performance, we consider both static and dynamic features of objects. We also provide a sufficient condition in which the matching can get the additional benefit from temporal information. The experimental results show that our STPM performs better than the other video matching methods.

1. Introduction

The convenient access to networked multimedia devices and multimedia hosting services has contributed the huge increase in network traffic and data storage. The recent reports say that 34% of the current cell phone users do video recording [1] and video traffic is 40% of consumer Internet traffic [2]. In addition to the current video hosting services such as YouTube [3] and Vimeo [4], major IT companies such as Google [5] and Apple [6] have started to offer cloud audio/video storage services to customers.

Compared to the recent efforts and deployments of content-based *image* search such as automatic tagging based on face recognition [7,8], content-based *video* search is still under-developed. We have two main observations to explain its shortcomings.

First, The temporal information on videos adds more complexity of dimensions of data, so queries could be more complex than typical text-based ones. In addition, the representations of these queries generated by simple sketch tools [9,10] are so primitive or generic compared with text-represented queries, they would lead either wrong or diverse query results. More complex querying system (such as dynamical construction of hierarchical structures on targeting videos [11]) requires more elaboration on queries by users, which could be more error-prone.

Second, this has been assumed that the user does not have sample videos at hand for query, so additional querying tools are required. However, this assumption is no longer valid because mobile devices such as digital cameras, PDAs, and cell-phones with camera and solid-state memory enable instant image and video recording which can be used for a video query.

Taking advantage of this opportunity from mobile and ubiquitous multimedia, our content-based video query system takes a sample video clip[3] as a query and searches the collection of videos typically stored in multimedia portal service (such as YouTube, Vimeo, Google Video [12], Yahoo! Video [13], etc.), and suggests similar video clips from the database with relevance evaluation. As shown in Fig. 1, our system mainly performs the following two functionalities – (1) *offline population of our video database* for new video entries to database and (2) *online video matching* for a new query video. When a video is introduced in the database, it is partitioned into multiple clips by a clip boundary detection based on feature analysis and classification. The partitioned video clips are stored along with metadata information in the database. Next, for a new video from query process, it is analyzed and matched to the stored videos and the relevant scores are calculated by our spatio-temporal pyramid matching system.

☆ This paper has been recommended for acceptance by Chung-Sheng Li.

* Corresponding author.
E-mail addresses: jaesikchoi@lbl.gov (J. Choi), wangziyu@inha.edu (Z. Wang), sclee@inha.ac.kr (S.-C. Lee), won.jeon@samsung.com (W.J. Jeon).
[1] Present address: Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA 94720, USA.
[2] Co-corresponding author.

[3] In this paper, we define the terms as follows – a *scene* is an image frame, a *video clip (or clip)* is a set of image frames in which has continuous movement of objects, and a *video* is a set of clips.

1077-3142/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.cviu.2013.02.003

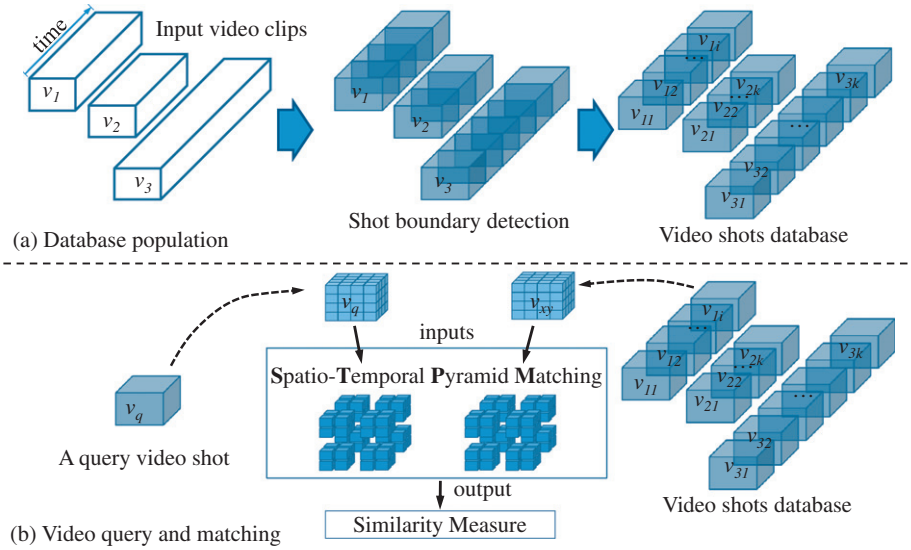

Fig. 1. Overview of our video matching system: (a) video partitioning for new video entries to database, (b) video searching for a given query video clip using STPM.

The rest of this paper is organized as follows. In Section 2, related work on image and video retrieval is discussed. Our spatio-temporal pyramid matching system is presented in Section 3. We also analyze formal conditions where our spatio-temporal pyramid matching system gets benefits from temporal information in Section 4. The experimental results are presented in Section 5, and finally we conclude our paper.

2. Related work

The challenges and characteristics of content-based image and video query systems are well discussed in [14]. A significant amount of research on automatic image annotation [15–20] has been done, and recently researchers are more focusing on automatic video annotation [21–24]. Especially, Ulges et al. [25] and Ando et al. [26] discussed video tagging and scene recognition problems, which have similar goals to ours but takes different approaches. Dynamic models such as Hidden Markov Models (HMM) have been used to track and summarize moving objects in video [27,28,26]. Compared to models with HMMs, our pyramid matching kernel has a simpler representation of features in time domain, therefore is faster to calculate the score of relevance feedback for video query.

Recently content-based video retrieval for sports video [29] has been widely discussed. The work in [30] and [31] focused on the framework and personalization of generic sports videos, whereas others target particular sports such as baseball [32,33], soccer [34], basketball [35], etc.

Different techniques of finding similarity of subsets of video streams have been discussed in [36–40]. Pyramid matching [41–43] is known as one of the best matching algorithms for image retrieval and recognition. Our spatio-temporal pyramid matching system [29] extends the spatial pyramid matching [42] to accommodate time domain for efficient video matching and query. Recently, our spatio-temporal pyramid matching is also adapted and improved further [44–47] and shown to be more effective than previous state-of-the-art methods for action retrievals in videos.

Temporally aligned pyramid matching (TAPM) [48] computes the similarity of video clips by dividing a video clip into a set of images, aligning the images independently, and matching the images hierarchically by using the agglomerative clustering [49]. Our STPM is similar to TAPM in a sense that we divide temporal domain hierarchically. However, there is a fundamental difference between STPM and TAPM in a sense that STPM incorporates temporal domain and spatial domains together more tightly. TAPM matches whole images in the lowest level in spatial domain, followed by the temporal hierarchy. Meanwhile, our STPM divides temporal domain with spatial domain. Thus, in the lowest (coarsest) level, we only see local spatio-temporal features, which are shown to be effective in [44]. The size of grid becomes larger along the level of pyramid hierarchy. Moreover, STPM is efficient and simple to implement because STPM does not require alignments.

A temporally-binned model [50], a local spatio-temporal action detector, and Space–time interest points (STIP) [51] are also similar to our STPM. The main difference is that our STPM uses a weighted, multi-scale pyramid (e.g., from the local patch-level to to the global video-level).

3. System design

Given a video clip as a new entry to the database, the clip boundary detection in our system divides it into multiple video clips and they are stored in our video matching database. Once the system receives a query video clip, the similarity between

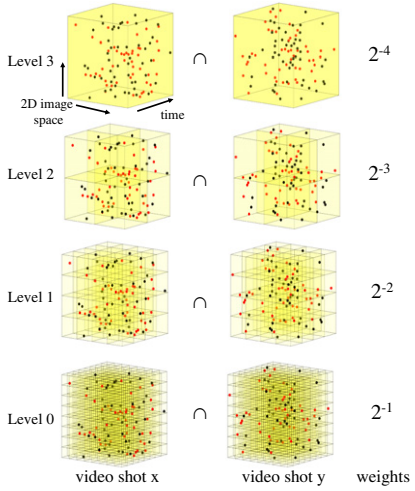

Fig. 2. Hierarchical structure of spatio-temporal pyramid matching.