Pattern Recognition 48 (2015) 522–533

Contents lists available at ScienceDirect

Pattern Recognition

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

Video summarization via minimum sparse reconstruction

Shaohui Mei [a,*], Genliang Guan [b], Zhiyong Wang [b], Shuai Wan [a], Mingyi He [a], David Dagan Feng [b]

[a] School of Electronics and Information, Northwestern Polytechnical University, Xi'an, Shaanxi 710129, China
[b] School of Information Technologies, The University of Sydney, NSW 2006, Australia

ARTICLE INFO

Article history:
Received 24 February 2014
Received in revised form
27 May 2014
Accepted 2 August 2014
Available online 15 August 2014

Keywords:
Video summarization
Keyframe extraction
Sparse reconstruction
Dictionary selection

ABSTRACT

The rapid growth of video data demands both effective and efficient video summarization methods so that users are empowered to quickly browse and comprehend a large amount of video content. In this paper, we formulate the video summarization task with a novel minimum sparse reconstruction (MSR) problem. That is, the original video sequence can be best reconstructed with as few selected keyframes as possible. Different from the recently proposed convex relaxation based sparse dictionary selection method, our proposed method utilizes the true sparse constraint L_0 norm, instead of the relaxed constraint $L_{2,1}$ norm, such that keyframes are directly selected as a sparse dictionary that can well reconstruct all the video frames. An on-line version is further developed owing to the real-time efficiency of the proposed MSR principle. In addition, a percentage of reconstruction (POR) criterion is proposed to intuitively guide users in obtaining a summary with an appropriate length. Experimental results on two benchmark datasets with various types of videos demonstrate that the proposed methods outperform the state of the art.

1. Introduction

Videos are visual data captured along time to depict the dynamic processes of events or activities. Traditionally, videos are represented as a sequence of consecutive frames, each of which corresponds to a constant time interval. As a result, it generally takes the same amount of time to watch them. Nowadays, video data are created in a rapidly increasing rate. For example, as reported by Youtube Statistics [1], 100 h of video are uploaded to Youtube every minute, which is equivalent to about 16 years of new videos in just one day. As a result, traditional access to video data presents significant limitations for the emerging new multimedia services such as content-based search, retrieval, navigation, video browsing, and semantic annotation. Therefore, video summarization (VS) is proposed for effective and efficient access to video content by extracting the essential information of a video to produce a compact and informative version.

VS has been extensively studied and there exist a large number of methods in the literature [2–4]. Generally, there are two basic forms of VS: keyframe extraction and video skimming. Keyframe extraction selects a collection of salient images from the underlying source, while video skimming selects a collection of video segments (and corresponding audio). Video skims can be created from keyframes by joining fixed-size segments, subshots, or the whole shots that enclose them [2,5]. In addition, keyframes provide a quicker view of video content and are helpful to reduce lots of computational complexity for various video analysis and retrieval applications. Therefore, in this paper, keyframe based VS is considered.

Keyframe extraction from a video is actually a ranking process of individual frames in terms of their representativeness to the video. As a result, most of the existing keyframe extraction methods can be organized into two categories: local representative methods and global representative methods, where the former category derives representativeness of a frame from its neighborhood segment and the latter category emphasizes more on the representativeness of a frame with regard to the whole video. The assumption of local representative methods is that a frame is representative if it differs significantly from its adjacent frames. For example, shot boundary based algorithms detect shot boundaries firstly and then select keyframes in each shot for VS [6,7]. The sufficient content change based methods select a frame as the keyframe only if its visual content significantly differs from previously extracted keyframes [8]. While local representative methods work well for some videos such as actions and

* Corresponding author. Tel.: +86 29 88431250.
E-mail addresses: meish@nwpu.edu.cn (S. Mei), genliang.guan@sydney.edu.au (G. Guan), zhiyong.wang@sydney.edu.au (Z. Wang), swan@nwpu.edu.cn (S. Wan), myhe@nwpu.edu.cn (M. He), feng@it.usyd.edu.au (D. Dagan Feng).

http://dx.doi.org/10.1016/j.patcog.2014.08.002
0031-3203/

procedural instructions where temporal order is important, the actual summary output may not be concise enough when there are repetitive segments. As a result, taking a global perspective could produce more concise video summaries for such videos.

By taking a global viewpoint, a video is viewed as a set of unordered frames and keyframes are assumed to be visually representative. Therefore, many clustering based VS approaches have been proposed in the literature [9–12]. The basic idea is to produce the summary by clustering similar frames/shots and select a limited number of frames per cluster (usually, one frame per cluster). However, it is usually very difficult to extract all clusters due to large intraclass and low interclass visual variance. In addition, existing works in keyframe extraction often fail to adequately evaluate the results of the clustering process [2]. Moreover, some clustering based algorithms are of high computational complexity. For example, in [10], the computation of the summaries takes around 10 times the video length. That is, the time for summarizing an 1-min video would be about 10 min.

Recently, Cong et al. [13] formulated VS from a new point of view, in which VS is transformed into a sparse dictionary (SD) selection problem and a relaxed constraint based on $L_{2,1}$ is imposed to ensure sparsity. However, it was not able to directly select keyframes based on the sparse dictionary since the $L_{2,1}$ norm based constraint cannot ensure sparsity directly. Instead, keyframes were selected by identifying local maximums of an importance curve generated according to the norm of reconstruction coefficients, which makes the SD based method [13] to be a *local representative method* for VS. As a result, the selected keyframes are not the optimal subset of frames for the defined model since the frames with least reconstruction coefficients, which also contribute to reducing reconstruction errors, may not be selected. In addition, the optimization algorithm for dictionary selection is computationally expensive, which makes it not suitable for real-time applications.

In this paper, with the inspiration from the SD based algorithm [13], VS is formulated as a problem selecting the minimum number of keyframes to reconstruct the entire video as accurate as possible. That is, we aim to select keyframes from a global point of view by adopting a real sparse constraint based on L_0 norm, instead of the relaxed constraint based on $L_{2,1}$ norm in [13]. A selection matrix is proposed to model keyframe selection from the original video, according to which the L_0 norm of this selection matrix is proposed to ensure selecting as few keyframes as possible. Specifically, two computationally efficient VS algorithms based on the minimum sparse reconstruction (MSR) principle, including an off-line version and an on-line version, are proposed to extract keyframes for VS. As a result, keyframes are selected directly for VS in the proposed algorithm, which makes our proposed methods to be global representative ones for VS. In addition, a percentage of reconstruction (POR) criterion is proposed to summarize video sequences with different lengths, enabling the proposed MSR based VS algorithms adaptive to different kinds of videos. Finally, experiments on two benchmark datasets are conducted to demonstrate the effectiveness of the proposed algorithms.

The main contributions of this paper reside in three aspects:

1. An MSR based VS model is formulated by utilizing a selection matrix, such that VS is performed by utilizing minimum number of keyframes to reconstruct the entire video as accurate as possible. An L_0 norm based constraint is imposed to ensure real sparsity such that keyframes are selected directly according to the selection matrix. Different from the traditional local representative SD method [13], the proposed MSR based VS method selects keyframes from a global point of view.
2. Two efficient and effective MSR based VS algorithms are proposed for off-line and on-line applications, respectively.
3. A scalable strategy is designed to provide flexibility for practical applications. The proposed POR criterion can be tuned to extract a keyframe set at different levels of reconstruction of the original video sequence.

The remainder of this paper is organized as follows. Section 2 reviews existing VS algorithms by organizing them in two categories: local representative methods and global representative ones. Section 3 explains the MSR model for VS. Section 4 presents our solution to the MSR based VS model and how it is employed to devise the off-line and on-line VS algorithms. Section 5 reports and discusses experimental results on two benchmark datasets. Finally, conclusion and comments on promising future research are stated in Section 6.

2. Related work

Video summarization, also known as video abstracting, has been well researched for decades. In [2], Truong and Venkatesh categorized existing researches into keyframe based and video skim based approaches in terms of the forms of video summaries. In [3], Money and Agius reviewed existing research with three categories: internal (i.e., video signals), external (e.g., audio and subtitles) and hybrid (i.e., a combination of internal and external information), in terms of how different information sources are incorporated to produce a final summary. Recently, many approaches focus on bringing high level semantics into video summarization, such as event [14–16], which often deal with multiple videos, and objects [17]. In this section, rather than providing a comprehensive coverage, we review a number of representative keyframe based summarization methods with two categories: local representative methods and global representative methods, in terms of whether temporal information is considered or not.

2.1. Local representative methods

Assuming that keyframes are different with their neighboring frames, local representative algorithms exploit redundancy within a temporal window. One of the simplest ideas is to detect shot boundaries first [6,7], and then to select keyframes in each shot for VS. For example, the first, last frame, or several key frames separated by a fixed time distance of each shot were selected as the keyframes of the shot [18–20]. Although such strategy is efficient for stationary shots, it is not adequate for dynamic shots since the dynamic visual content of a shot is not taken into account at all. Therefore, many algorithms have been proposed to account for the dynamics of a video shot. For example, three iso-content principles, namely Iso-Content Distance, Iso-Content Error, and Iso-Content Distortion [20], were utilized to extract keyframes in each shot so that the selected keyframes are equidistant after each shot is characterized with visual feature vectors and projected into the feature space.

Without explicitly detecting the shots, the Sufficient Content Change based methods selected a frame as the keyframe only if its visual content significantly differs from previously extracted keyframes. A variety of metrics have been proposed to measure the content change in the algorithm, including the histogram difference [21,22], the accumulated energy function computed from image-block displacements across two successive frames [23], and changes in the number or geometric properties of extracted video objects [24]. As a variant of the Sufficient Content Change method, the Equal Temporal Variance method selected keyframes under