Information Sciences 432 (2018) 319–331

Contents lists available at ScienceDirect

Information Sciences

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

A salient dictionary learning framework for activity video summarization via key-frame extraction

Ioannis Mademlis*, Anastasios Tefas, Ioannis Pitas

Department of Informatics, Aristotle University of Thessaloniki, Thessaloniki, Greece

ARTICLE INFO

Article history:
Received 2 October 2017
Revised 24 November 2017
Accepted 10 December 2017
Available online 13 December 2017

Keywords:
Video summarization
Key-frame extraction
Column subset selection problem
Video saliency
Genetic Algorithm

ABSTRACT

Recently, dictionary learning methods for unsupervised video summarization have surpassed traditional video frame clustering approaches. This paper addresses static summarization of videos depicting activities, which possess certain recurrent properties. In this context, a flexible definition of an activity video summary is proposed, as the set of key-frames that can both reconstruct the original, full-length video and simultaneously represent its most salient parts. Both objectives can be jointly optimized across several information modalities. The two criteria are merged into a "salient dictionary" learning task that is proposed as a strict definition of the video summarization problem, encapsulating many existing algorithms. Three specific, novel video summarization methods are derived from this definition: the Numerical, the Greedy and the Genetic Algorithm. In all formulations, the reconstruction term is modeled algebraically as a Column Subset Selection Problem (CSSP), while the saliency term is modeled as an outlier detection problem, a low-rank approximation problem, or a summary dispersion maximization problem. In quantitative evaluation, the Greedy Algorithm seems to provide the best balance between speed and overall performance, with the faster Numerical Algorithm a close second. All the proposed methods outperform a baseline clustering approach and two competing state-of-the-art static video summarization algorithms.

1. Introduction

Massive amounts of digital visual media data are publicly available nowadays, accelerating the transformation of global culture into a vision-dominated one [38]. Thus, the need for compact and succinct visual data presentation has arisen. It is a problem of broad interest in domains where large-scale video footage must be stored, archived, analysed or visualized, that typically demands tedious human intervention and manual effort.

Automated *video summarization* offers one solution to the video presentation problem, by generating concise versions of a video stream that only retain its most informative and representative content. Relevant algorithms are expected to meticulously strike a balance between summary compactness, conciseness, enjoyability and content coverage. *Static summarization* typically extracts a set of representative video frames, i.e., *key-frames*, that in a sense summarize the entire video content. When editing cuts between clearly separated video shots are discernible (e.g., in television or film content [29]), a shot cut/boundary detector [4] is typically employed before key-frame extraction, to facilitate the summarization process by operating independently at each shot [16].

* Corresponding author.
E-mail address: imademlis@aiia.csd.auth.gr (I. Mademlis).

https://doi.org/10.1016/j.ins.2017.12.020
0020-0255/

Several other possibilities exist for video summarization, such as *dynamic summarization* (also called *skimming*) [8], *video synopsis* [37] or *temporal video segmentation* exploiting semantic activity cues [39]. Despite their advantages, they may only be suitable for specific applications (for instance, the use of synopses is limited to cases where the video frames are not visually crowded and retaining the original content is not a requirement), or even require key-frame extraction as a pre-processing / post-processing step. Thus, this paper focuses on key-frame extraction, used hereafter synonymously with video summarization.

In most of the relevant literature, video summarization is implicitly defined as a video frame sampling problem, constrained by an attempt to simultaneously satisfy several intuitive heuristic criteria, such as representativeness (extraction of key-frames that are jointly indicative of the original video content), compactness (lack of redundancy in the selected key-frames), outlier inclusion (selection of atypical key-frames) and content coverage (representation of the entire original video in the produced summary) [25]. Additionally, the summary should be as concise (i.e., short in length) as possible, or as desired by the end-user.

The traditional summarization method based on the constrained video frame sampling philosophy is video frame clustering, where frames closest to the estimated cluster centroids, or medoids, are selected as key-frames [44]. Thus, the video summarization problem is simply cast as a distance-based data partitioning task, with all semantic content description offloaded solely to the underlying video frame description/representation algorithm.

The modern alternative route is extracting a key-frame set as a dictionary of representative video frames that can linearly reconstruct the entire original video stream. This is an effective approach, supported by a sound theoretical background, that does not depend on shot cut/boundary detection or temporal video segmentation and formalizes the representativeness, compactness and content coverage criteria. Under a reasonable linear representatives assumption [11], i.e., all original video frames can be approximately reconstructed as linear combinations of a representative subset of them, it can be argued that such methods, when supported by appropriate underlying video frame description/representation schemes, are able to incorporate scene semantics into the summarization algorithm itself. The reason is that the extracted key-frames will inherently tend to depict disjoint subsets of visually important scene objects, spatial segments, activities etc. In contrast, with a distance-based clustering approach, such a semantically meaningful partitioning of the key-frames will only be a serendipitous outcome.

However, dictionary-of-representatives approaches do not guarantee outlier inclusion. A related issue is that the reconstructive advantage conveyed by a video frame (i.e., the sole factor typically considered) cannot be the only criterion for its inclusion in the extracted key-frame set. The reason is that this leads to the extraction of unimportant video frames that happen to depict common but uninteresting visual building blocks of the entire video (e.g., the background), at the expense of engaging video frames, containing atypical visual elements which do not contribute much to the reconstruction.

Since most human activities can easily be decomposed into specific combinations of elementary actions [1], activity videos tend to satisfy the linear representatives assumption. However, the problems of dictionary-of-representatives approaches are especially pronounced when summarizing activity videos, due to their characteristic properties: static camera, static background, heavy inter-frame visual redundancy and lack of editing cuts. This paper, which integrates and extends preliminary work [30,31], introduces a framework for activity video summarization that attempts to overcome the above issues. Its contributions are four-fold.

First, video summarization is explicitly formalized under a flexible, multimodal framework that can accommodate several existing algorithms as special cases. The proposed framework generalizes most current dictionary methods, which only consider the reconstructive ability, and conceptually places video summarization at the crossroad between video saliency estimation and video dictionary learning, thus defining it as a "salient dictionary" learning task. Second, three key-frame extraction algorithms are formulated in this context, where the video reconstruction term is modeled as a Column Subset Selection Problem (CSSP) [2]. This guarantees summary conciseness, favors summary compactness and had not been utilized for key-frame extraction before the preliminary work that this paper extends [30]. Third, the saliency terms for the Numerical and for the Greedy Algorithms are novel, video-oriented modifications of state-of-the-art image saliency algorithms ([10,26] respectively). Fourth, a novel metric, called Independence Ratio (IR), is proposed as an objective performance indicator of activity video key-frame extraction.

The three presented video summarization methods are compared against a baseline clustering approach [8] and two competing, dictionary-of-representatives state-of-the-art static video summarization algorithms [7,32]. Each of the proposed algorithms is evaluated in five separate variants, characterized by a different balance between the reconstruction and the saliency term. All of the above algorithms, including both the proposed and the competing ones, are different, specific formulations of the presented salient dictionary learning framework for static video summarization.

2. Related work

Current static video summarization methods can be broadly classified into supervised and unsupervised learning methods. Supervised approaches have surfaced lately [42,43], in the wake of the success of deep learning. Such methods do not rely on heuristic summarization criteria, but attempt to implicitly learn them from human-created manual video summaries. However, due to the subjectiveness inherent in the problem (different persons may produce widely differing summaries from the same video source) and the lack of manual activity video summaries readily available for training in most use-cases, this paper is focused solely on unsupervised algorithms.