Artificial Intelligence in Medicine 66 (2016) 1–13

Contents lists available at ScienceDirect

Artificial Intelligence in Medicine

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

Scalable gastroscopic video summarization via similar-inhibition dictionary selection

Shuai Wang[a,d,*], Yang Cong[a], Jun Cao[b], Yunsheng Yang[e], Yandong Tang[a], Huaici Zhao[c], Haibin Yu[f]

[a] State Key Laboratory of Robotics, Shenyang Institute of Automation, Chinese Academy of Sciences, Nanta Street 114, Shenyang 110016, China
[b] Department of Computer Science, Arizona State University, 1711 South Rural Road, Tempe, AZ 85287, USA
[c] Key Laboratory of Image Understanding and Computer Vision, Shenyang Institute of Automation, Chinese Academy of Sciences, Nanta Street 114, Shenyang 110016, China
[d] University of Chinese Academy of Sciences, No. 19A Yuquan Road, Beijing 100049, China
[e] Department of Gastroenterology and Hepatology, Chinese PLA General Hospital, 28 Fuxing Road, Beijing 100000, China
[f] Key Laboratory of Networked Control Systems, Shenyang Institute of Automation, Chinese Academy of Sciences, Nanta Street 114, Shenyang 110016, China

ARTICLE INFO

Article history:
Received 7 February 2015
Received in revised form 19 June 2015
Accepted 7 August 2015

Keywords:
Video summarization
Key frame
Similar-inhibition dictionary selection
Image attention prior
Gastroscopic video

ABSTRACT

Objective: This paper aims at developing an automated gastroscopic video summarization algorithm to assist clinicians to more effectively go through the abnormal contents of the video.
Methods and materials: To select the most representative frames from the original video sequence, we formulate the problem of gastroscopic video summarization as a dictionary selection issue. Different from the traditional dictionary selection methods, which take into account only the number and reconstruction ability of selected key frames, our model introduces the similar-inhibition constraint to reinforce the diversity of selected key frames. We calculate the attention cost by merging both gaze and content change into a prior cue to help select the frames with more high-level semantic information. Moreover, we adopt an image quality evaluation process to eliminate the interference of the poor quality images and a segmentation process to reduce the computational complexity.
Results: For experiments, we build a new gastroscopic video dataset captured from 30 volunteers with more than 400k images and compare our method with the state-of-the-arts using the content consistency, index consistency and content-index consistency with the ground truth. Compared with all competitors, our method obtains the best results in 23 of 30 videos evaluated based on content consistency, 24 of 30 videos evaluated based on index consistency and all videos evaluated based on content-index consistency.
Conclusions: For gastroscopic video summarization, we propose an automated annotation method via similar-inhibition dictionary selection. Our model can achieve better performance compared with other state-of-the-art models and supplies more suitable key frames for diagnosis. The developed algorithm can be automatically adapted to various real applications, such as the training of young clinicians, computer-aided diagnosis or medical report generation.

1. Introduction

More and more people are suffering from stomach diseases, and the trend is rising [1]. As an effective technique to show the interior of a stomach directly, gastroscopy has been widely used for clinical examination, especially for the early detection of gastric cancer. Usually, the entire procedure lasts approximately 20 min, and a video containing approximately 15,000 frames is captured. However, the visual inspection of such a large number of frames is a challenging task, even for the most experienced clinicians. To more easily browse through such a video archive, a clinician records approximately 20–50 images manually during the examination for diagnosis and later generates a medical report. Nonetheless, the manual annotation may have the following shortcomings:

- Because of the need to perform multiple tasks simultaneously, clinicians may miss some important information for the final diagnosis.

* Corresponding author at: State Key Laboratory of Robotics, Shenyang Institute of Automation, Chinese Academy of Sciences, Nanta Street 114, Shenyang 110016, China. Tel.: +86 24 23970421; fax: +86 24 23970021.
E-mail address: shuaiwang@sia.cn (S. Wang).

http://dx.doi.org/10.1016/j.artmed.2015.08.006
0933-3657/

- Due to the lack of enough experience, some junior clinicians cannot guarantee accuracy when analyzing the massive data continuously. Especially when the operation is not timely, clinicians may select poor quality images.
- After the completion of the manual operation, the number of selected frames is fixed, which cannot meet the needs of different scenarios and may increase the time cost for re-analysis.

In fact, the above process is a typical video summarization procedure, i.e., selecting some frames with the most important and meaningful semantic content from a full-length video sequence [2–5]. Therefore, in this paper, we intend to design a computer-aided gastroscopic video summarization algorithm to overcome these problems and assist clinicians to more effectively go through the abnormal contents of the video. The computer-aided system based on our algorithm can be adopted in real applications, such as the training of young clinicians, computer-aided diagnosis or medical report generation.

For video summarization [6–10], most state-of-the-art methods mainly focus on the summarization of structured videos, such as sports, cartoons or surveillance videos. In comparison, the automatic summarization of unstructured data, e.g., gastroscopic videos, is much more challenging. First, gastroscopic videos contain deformable and low-texture context, which makes it more difficult to extract semantic information. Second, due to the complexity of the inner human cavity and arbitrary movement of the camera, some gastroscopic images are of poor quality, which makes an accurate video summarization difficult. Finally, the objective of gastroscopic video summarization is for diagnosis, so the result of video summarization should highlight the suspected regions. Some previous models, e.g., the group sparsity dictionary selection model [2] in our previous work, cannot handle the above challenges very well. For gastroscopic videos, the result cannot encompass all video content, and some similar frames are also frequently selected as key frames. Therefore, we design a new similar-inhibition dictionary selection model by adopting the similar-inhibition constraint to select elements with more diversity between each other. Based on the similar-inhibition constraint, the video structure information will be taken into account to cover as much video content as possible in comparison with traditional sparse dictionary selection models. Furthermore, we also integrate an attention prior into the group sparsity term to reduce the gap between low-level features and high-level concepts. The main contributions of this paper reside in three aspects:

- We design a new dictionary selection model by adopting the similar-inhibition constraint, which reinforces the diversity of the selected subset.
- By taking into account the attention prior, we propose a scalable gastroscopic video summarization algorithm via similar-inhibition dictionary selection, which can select key frames with the most semantic information efficiently.
- We collect and build a new gastroscopic video summarization dataset from 30 volunteers with approximately 432,000 frames, and we annotate the ground truth for evaluation as well. To the best of our knowledge, this dataset is the first gastroscopic video summarization dataset.

The rest of this paper is arranged as follows. Section 2 discusses the related works. In Section 3, we present the formulation of the problem. Section 4 describes the implementation of our video summarization. Section 5 presents various experiments and comparisons. Finally, Section 6 concludes the paper.

2. Related works

The problem of video summarization has attracted significant attention, especially over the past few years, and [9,11] propose detailed reviews of existing techniques. To capture the content changes in a video sequence, most existing approaches first segment the whole video into shots using shot detection methods and then select key frames from each shot [12,13]. The simplest method is to select the first/middle/last frame of each shot as key frames. Shahraray and Gibbon [14] propose a content-based sampling method to select key frames. Panagiotakis et al. [15] first choose the first and last frames of each shot as key frames and then compute the remaining key frames with the maximum equidistant in the sense of the Iso-Content. Taking the pre-determined number of key frames as a constraint, Lee and Kim [16] adjust the positions and time-intervals of key frames by reducing the distortion iteratively. These algorithms can be called sequential algorithms and produce acceptable results for videos with simple structures, such as movies and news, where the presence of a shot indicates new content.

Another approach intends to group the frames into visually similar clusters and selects frames relevant to the cluster centers as key frames [17,18]. Zhuang et al. [19] group the frames in clusters and selects the key frames from the largest clusters. Jiang et al. [20] use the k-means method to cluster the data points in Laplacian subspace and select the key frames from a three layer video structure. To reduce the redundancy of video for personal video recorders, Gao et al. [21] remove redundant video content to select key frames using the hierarchical agglomerative cluster method. Lee et al. [22] first use novel egocentric saliency cues to train a category-independent regression model for predicting how likely an image region belongs to an important person or object. Then, a new video is partitioned into events by clustering scenes with similar global appearance. For each cluster, the region with the highest importance score is selected as its representative. Although clustering techniques have been quite effective, the loss of semantic details is almost inevitable, leading to a significant semantic gap.

There are also many optimization-based algorithms. By solving a specifically designed objective function under selection criteria, the optimal data points are selected as key frames. For instance, Shih [23] combines the visual saliency-based attention features with the contextual game status information of sports videos and selects the key frames with the maximal visual attention scores. Liu et al. [24] model the summarization procedure as a maximum a posterior problem by integrating both shot boundary detection and key frame selection. In [2], the key frames are selected via a novel dictionary selection model, and similar ideas are also applied in [25–28].

For medical endoscopy, one important field is the inspection of the gastrointestinal tract (GT tract) [29]. Many computer-aided endoscopy diagnosis systems have been proposed to assist clinicians in improving the accuracy of medical diagnosis using the images or videos recorded in the inspection of a GT tract. According to the specific lesions, these systems can be classified to handle bleeding [30,31], tumors [32,33], Helicobacter pylori [34], cancer [35,36], Crohn's disease [37] and polyps [38]. Moreover, some other applications include pose detection for endoscopy [39], video segmentation [40] and three-dimensional reconstruction of the digestive wall [41]. For video summarization, as discussed above, most previous works mainly focus on the summarization of structured videos, which have well-defined temporal structures and characteristics for selecting key frames. In comparison, the endoscope video of our problem is more subjective and difficult for summarization. To our best knowledge, the state-of-the-art works for endoscope video summarization merely adopt wireless capsule