Chapter 6

Video Summarization by Redundancy Removing and Content Ranking

Tao Wang
Intel Labs China, China

Yue Gao
Intel Labs China, China & Tsinghua University, China

Patricia P. Wang
Intel Labs China, China

Wei Hu
Intel Labs China, China

Jianguo Li
Intel Labs China, China

Yangzhou Du
Intel Labs China, China

Yimin Zhang
Intel Labs China, China

ABSTRACT

Video summary is very important for users to grasp a whole video’s content quickly for efficient browsing and editing. In this chapter, we propose a novel video summarization approach based on redundancy removing and content ranking. Firstly, by video parsing and cast indexing, the approach constructs a story board to let user know about the main scenes and the main actors in the video. Then it removes redundant frames to generate a “story-constraint summary” by key frame clustering and repetitive segment detection. To shorten the video summary length to a target length, “time-constraint summary” is constructed by important factor based content ranking. Extensive experiments are carried out on TV series, movies, and cartoons. Good results demonstrate the effectiveness of the proposed method.

DOI: 10.4018/978-1-60960-024-2.ch006
INTRODUCTION

Rapid advances in the technology of media capture, storage and network have contributed to an amazing growth of digital video content. Existing so many and long videos, it is very time-consuming for us to know the video content before we browse them and decide which part to watch. To deal with the problem, video summarization becomes very important to help users to grasp the video content efficiently.

Video summarization is generally a condensed sequence of still or moving images, which provide the essential content of a video in a general, logical, and connected way. According to the summary mode, summarization approaches can be categorized into “story board” and “video skimming”. Story board is a collection of still images, such as a key frame list of important shots or scenes (Uchihashi, 1999). Story board can be constructed fast with small storage but their descriptive ability is limited since they lose lots of dynamic audiovisual content in the original video. Compared with story board, video skimming is made up of video clips which show important scenes, actors, objects and events for efficient browsing, e.g. highlights of a Hollywood movie. Based on the information theory, Gong and Liu (2001) proposed a video skimming approach with minimal visual content redundancies. They first cluster key frames and then concatenate short video segments of representative key frames to construct a video skimming. Li and Schuster (2005) formulated the optimal video summarization problem as finding a predetermined number of frames that minimize the temporal constraint. Otsuka and Nakane (2005) proposed a highlights summarization approach. They use audio feature to detect sports highlights as the video skimming.

Video summarization is so active research field that the National Institute of Standards and Technology (NIST) hold the evaluation of rushes video summarization in both 2007 and 2008 (Paul Over, 2007 and 2008). Rushes are the raw material (extra video, B-rolls footage) used to produce a video. The rushes summarization task is to automatically create a MPEG-1 summary clip less than or equal to a maximum duration (e.g. 2%) that shows the main objects and events of the rushes videos. Since there are many redundant, repeated, unstructured and bad video clips, e.g. color bar, near uniform-color, abrupt and clapboard frames etc, rushes video summarization is more challengeable than summarization of teleplays and movies. There are about 30 teams to join in the evaluation including Carnegie Mellon University, Dublin City University, City University of Hong Kong, National Institute of Informatics, FX Palo Alto Laboratory Inc., AT&T Labs, and Intel etc. In rushes summarization of TRECVID, the popular approaches are video sampling, key frame clustering and iteratively selecting video clips. For video sampling, the CMU team’s baseline1 approach selects 1-second video segments from every 25 seconds of an original video (Alexander, 2007). The 1-second segments are then concatenated to generate the summary video. Sampling approach is very simple but hard to beat other methods due to losing analysis of video structure and content. For key frame based clustering, many teams employ the approach to find representative video clips and remove redundant clips, e.g. k-means clustering of CMU-baseline2 (Alexander, 2007) and UCAL (Anindya, 2007), and hierarchical clustering of JOANNEUM(Werner, 2007) and THU-ICRC(Wang, 2007) etc. After key frame clustering, the representative key frames (which are near to cluster center, last appeared or longest in duration) are selected and concatenated into the summary video by their temporal order. Key frame clustering approach is effective, but how to automatically select the most representative clips and decide their concatenated order are the main problems. Different from clustering approaches, the greedy method of iteratively selecting video clips divides a video into 1-second video clips and generate the video summary by iteratively selecting the clips with the highest representative score (Wang, 2007). This greedy approach can quickly select representative video clips and