Automation in Construction 81 (2017) 355–368

Contents lists available at ScienceDirect

Automation in Construction

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

Automatic key frame extraction in continuous videos from construction monitoring by using color, texture, and gradient features

Ling Chen, Yuhong Wang*

Department of Civil and Environmental Engineering, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong

ARTICLE INFO

Keywords:
Construction video archiving
Scene transition detection
Image feature selection
Key frame extraction

ABSTRACT

On-site video recording systems are increasingly used for monitoring construction activities. The recorded videos contain rich and useful jobsite information that can be used for a variety of purposes. The large amount of video data generated by continuous monitoring, however, creates tremendous challenges on data storage and retrieval. Due to the relatively slow pace of construction activities, a significant portion of the recorded data is redundant. Therefore, archiving raw construction videos into a concise and structured set of key frames would facilitate data storage, retrieval and analysis. Three key issues in automatic key frame extraction from construction videos are studied, including the selection of proper video features, scene segmentation, and key frame extraction. New image features and methods are developed to address the three issues. A validation experiment indicates that the developed features and methods can effectively and efficiently extract representative key frames from the complex and dynamic construction videos. The developed techniques can be used to develop a construction video summary system that serves the purpose of effectively archiving construction jobsite videos.

1. Introduction

Video recording systems are widely used on construction sites nowadays. On some projects, it is even compulsory for contractors to install video cameras of specified resolution and at designated locations to perform continuous recording [1]. The generated videos may be viewed real-time and/or become part of project records for various purposes. Compared with still photos and time-lapse images, videos provide more and richer information. They form a useful repository of project information from the perspective of project documentation management.

Although potentially useful, two obstacles are encountered in using raw construction videos. One issue is that too much data are generated through continuous recording. A typical video camera on a construction site may generate 2–3 gigabytes of data per hour. A project that lasts for several years will create huge amount of data. Another challenge is to retrieve useful information from the lengthy and unstructured video records. It is inconvenient and time-consuming to browse over the entire video footages to find the desired information. As a result, some contractors and owners think such video records are of limited use. They often discard the continuous videos shortly after construction due to limited electronic storage space [2]. This compromises the intention of using the video records as permanent project documents, training materials, or materials for knowledge sharing.

A close scrutiny of the video records, however, often reveals that a significant portion of the data is of little use. For example, unless for security reasons, videos recorded during non-work hours contain almost no new project information. Even during work hours, there is too much redundant information when the project progresses slowly. Therefore, there is great potential to "**lean**" the construction videos by removing those redundant frames. Time and structure information may be also attached to the "leaned" video records to make them better indexed and more easily retrieved. The processed records will then contain useful yet condensed information that can be economically stored, well organized, easily retrieved, and effectively used.

This paper presents the research of developing a construction video archiving system that aims to remove a large number of redundant image frames. The research focuses on solving three key issues in the development of the system, including: (1) identification of features that are appropriate for abstracting construction videos, (2) identification of effective method for segmenting construction videos into scenes based on the features, (3) identification of effective method for extracting key frames from the segmented scenes. More emphasis is placed on the comparison of existing image features and the development of new ones for analyzing construction videos. New methods for video segmentation and key frame extraction are also developed. The obtained key frames and scenes from trial videos are chronically organized into an easily retrievable Extensible Markup Language (XML) document to facilitate

* Corresponding author.
E-mail address: ceyhwang@polyu.edu.hk (Y. Wang).

http://dx.doi.org/10.1016/j.autcon.2017.04.004
Received 16 May 2016; Received in revised form 10 March 2017; Accepted 5 April 2017
Available online 25 April 2017
0926-5805/

web-based query and browsing (not presented in this paper for the interest of brevity). Although video summary systems have been developed in other fields (e.g., sports videos, medical videos) [3–13], the approaches used in the existing systems cannot effectively address the unique characteristics of construction videos.

The rest of the paper is organized as follows. The second section introduces relevant existing studies, the characteristics of construction videos, and the general directions of feature selection. The third section discusses three types of image features, with emphasis on how to define the features for the purpose of processing construction videos. The fourth section introduces the use of the computed image features for scene segmentation and key frame extraction. The fifth section summarizes the experimental results of examining the accuracy and efficiency of scene segmentation and key frame extraction methods. The last section summarizes the findings and introduces future research needs.

2. Related work and the architecture of the developed system

2.1. Related work

In response to the explosive growth of multimedia data, video summary systems have been researched and developed for various purposes. Example applications include television news [3–7], film videos [8,9,14], sports videos [10,11,15], ecological videos [12], medical videos [13] and geographical survey videos [16]. For instance, video summary systems have been developed to remove irrelevant and low-quality images from endoscopic videos [13] and laparoscopic videos [17] for more effective use of medical records. The typical framework of a video summary system is illustrated in Fig. 1. In the framework, the most important components are video feature selection and computation, video segmentation, and key frame extraction.

A video flow consists of images, captions and audios [18]. Feature extraction is to convert raw information in recorded videos into abstract information. Although caption [18] and audio [15] features are often selected in some news and film video summary systems, the most widely used feature is image feature. Unlike image pixels, image features are not affected by image transformation such as transition, zooming, and stretching and hence are more reliable. Image feature extraction is an essential step in image/video content analysis [19]. Various image features can be generally divided into two classes: high-level and low-level features. The typical high-level features are shapes or objects (e.g., human beings) [14]. Low-level features include global [20–22], local [13,23], and motion [24] features. For a global feature, the entire image is the subject of interest, whereas only a set of blobs in the image is interested for a local feature. For a motion feature, time is used as a variable in conjunction with the image information. The selected features need to match the characteristics and purpose of a particular video summary system.

The calculated image features can be used for image segmentation, which requests to detect the boundaries of shots or scenes in the continuous video footage. Shots may be generated by camera cut, edited changeover [25] and camera work state change (lens transition, rotation and zooming). Although a construction video created by continuous monitoring may not experience camera state changes, the videos may be divided into scenes that designate the major changes of construction tasks, e.g., switching from steel rebar installation to concrete placement. Such segmentation may provide better organization of construction videos based on work content. Shot or scene boundary detection is usually achieved by measuring the difference between several consecutive image frames [26]. According to the number of compared frames, the commonly used boundary detection methods can be divided into 2-frame method [27] and 3-frame method [28,29]. The performance of scene segmentation is also closely related to the chosen features and frame discrimination computation method. Another aim of this study is to choose a proper frame discrimination method and compare the performance of different features in scene segmentation.

Key frames are extracted from each scene after scene boundary detection. Key frames are the set of the representative frames from a scene. Existing key frame extraction methods can be categorized into four classes: shot boundary method, temporal sampling, clustering method, and visual difference curve based method. Shot boundary method uses shot boundaries (SB) as key frames [30]. This method, however, often misses the salient (distinguished) frames in the video stream. Temporal sampling methods include uniform temporal sampling (UTS) and adaptive temporal sampling (ATS). In UTS, key frames are extracted at a fixed time interval. UTS has two drawbacks: (1) the

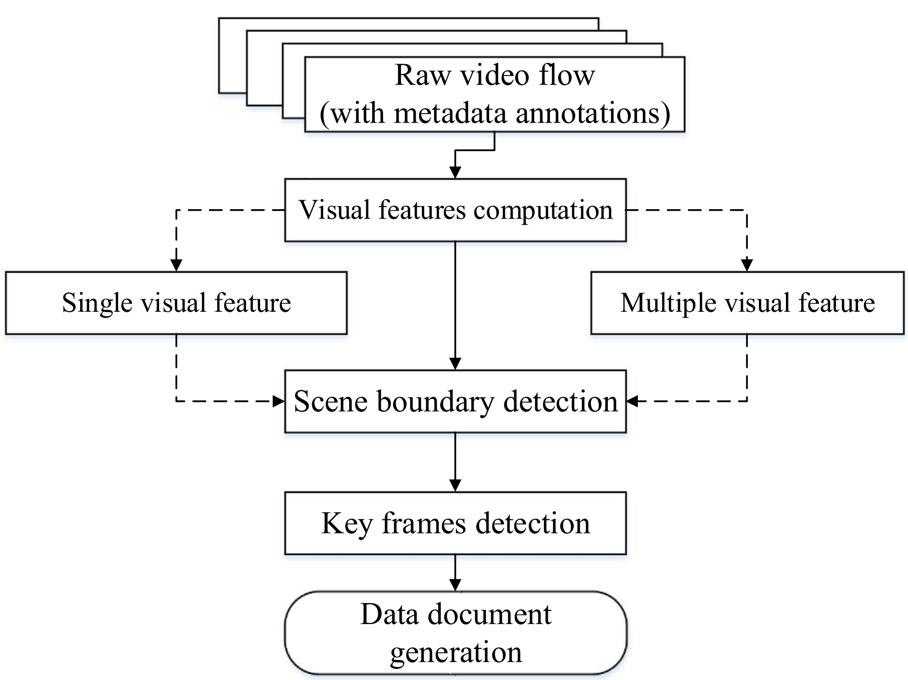

Fig. 1. The general framework of a video summary system.