



in a practical video content management system, as generated key frames and skims provide users an efficient way to browse or search video content. With the proliferation of digital video, this process will become an indispensable component to any practical content management system. A video summary can be displayed without the worry of timing issues. Moreover, extracted key frames could be used for content indexing and retrieval. However, from the viewpoint of user, a video skim may provide a more attractive choice since it contains audio and motion information that makes the abstraction more natural, interesting and informative.

Film is an art form that offers a practical, environmental, pictorial, dramatic, narrative, and musical medium to convey a story [1]. Although it can be viewed as a type of generic video, complex film editing techniques, such as the selecting, ordering, and timing of shots; the rate of cutting; and the editing of soundtracks, are required to produce a successful movie. Consequently, all of these special features need to be taken into account for better content analysis, understanding, and management. There has been recent work on movie content abstraction, which produces a static storyboard, a summary sequence, or a highlight [2]. A summary sequence provides users a small taste of the entire video, while a highlight contains only the content that may appear interesting to viewers such as the movie trailer. Despite the large amount of research on generic video abstraction, it remains a challenge to generate meaningful movie abstracts due to the special filming and editing characteristics.

## SURVEY ON VIDEO ABSTRACTION

### VIDEO SUMMARIZATION

Based on the way a key frame is extracted, existing work in this area can be categorized into three classes: sampling based, shot based, and segment based. Most of the earlier summarization work belongs to the sampling-based class, where key frames were either randomly chosen or uniformly sampled from the original video. The video magnifier [3] and the MiniVideo [4] systems are two examples. This approach is the simplest way to extract key frames, yet such an arrangement may fail to capture the real video content, especially when it is highly dynamic.

More sophisticated work has been done to extract key frames by adapting to dynamic video content. Since a shot is defined as a video segment taken from a continuous period, a natural and straightforward way is to extract one or more key frames from each shot using low-level features such as color and motion. A typical approach was proposed in [5], where key frames were extracted in a sequential fashion via thresholding. More sophisticated schemes based on color clustering, global motion, or gesture analysis could be found in [6]–[8]. Realizing that regular key frames cannot represent the underlying video dynamics effectively, researchers have looked for an alternative way to represent the shot content using a synthesized panoramic image called the *mosaic*. Along this direction, various types of mosaics such as static background mosaics and synopsis mosaics have

been proposed in [9] and [10]. An interchangeable use of regular key frames and mosaic images has also been studied in [11]. Some other work applied mathematical tools to the summarization process. For instance, a video content could be represented by a feature curve in a high-dimensional feature space with key frames corresponding to the curvature points [12]. One drawback of the shot-based key frame extraction approach is that it does not scale up well for long video.

More recently, efforts have been made in extracting key frames at a higher unit level, referred to as the segment level. Various clustering-based extraction schemes have been proposed. In these schemes, segments are first generated from frame clustering and then the frames that are closest to the centroid of each qualified segment are chosen as key frames [13], [14]. Yeung and Yeo [15] reported their work on video summarization at the scene level. Based on a detected shot structure, they classified all shots into a group of clusters using a time-constrained clustering algorithm, and then extracted meaningful story units (or scenes) such as dialogue and action. Next, representative images (R-images) were selected for each story unit to represent its component shot clusters. All extracted R-images of a story unit were resized and organized into a single regular-sized image following a predefined visual layout called the *poster*. Other schemes based on sophisticated temporal frame sampling [16], hierarchical frame clustering [17], fuzzy classification [18], singular value decomposition, and principle component analysis techniques have been tried with some encouraging results.

### VIDEO SKIMMING

A three-layer system diagram for video skimming is shown in Figure 1. In this system, low-level features are extracted and preprocessing tasks (such as commercial break and/or shot detection) are performed at the first layer. At the second layer, mid- to high-level semantic features are derived, which can be accomplished using techniques such as face detection, audio classification, video text recognition, and scene or event detection. The third layer assembles clips that possess user-desired length and content into the final abstract.

Previous work on video skimming can be classified into two categories: summary oriented and highlight oriented. A summary-oriented skim keeps the essential part of the original video and provides users a summarized version [19]. In contrast, the highlight-oriented skim only comprises a few interesting parts of the original video. Movie trailers and highlights of sports are examples of this skim type [20].

Defining which video segments to be highlighted is a subjective and difficult process. It is also challenging to map human perception into an automated abstraction process. Hence, most current video skimming work is summary-oriented. One straightforward approach is to compress the original video by speeding up the playback. As pointed out by Omoigui et al. [21], a video program could be watched in a fast playback mode without distinct pitch distortion using a time compression technology. Similar work was also reported by Amir et al. [22], where an