Chapter IV

Statistical Audio-Visual Data Fusion for Video Scene Segmentation

Vyacheslav Parshin, Ecole Centrale de Lyon, France
Liming Chen, Ecole Centrale de Lyon, France

Abstract

Automatic video segmentation into semantic units is important in order to organize an effective content-based access to long video. In this work, we focus on the problem of video segmentation into narrative units called scenes—aggregates of shots unified by a common dramatic event or locale. In this work, we derive a statistical video scene segmentation approach that detects scenes boundaries in one pass, fusing multi-modal audiovisual features in a symmetrical and scalable manner. The approach deals properly with the variability of real-valued features and models their conditional dependence on the context. It also integrates prior information concerning the duration of scenes. Two kinds of features extracted in visual and audio domain are proposed. The results of experimental evaluations carried out on ground truth video are reported. They show that our approach effectively fuses multiple modalities with higher performance compared with an alternative rule-based fusion technique.

Copyright © 2007, Idea Group Inc. Copying or distributing in print or electronic forms without written permission of Idea Group Inc. is prohibited.
Introduction

A constantly growing amount of available digitized video stored at centralized libraries or even on personal computers gives rise to the need for an effective means of navigation that allows a user to locate a video segment of interest. Searching of such a segment sequentially using simple fast-forward or fast-reverse operations provided by most of the existing players is tedious and time-consuming. A content-based access could greatly simplify this task, giving to a user the possibility to browse a video organized as a sequence of semantic units. Such an organization also could facilitate the task of automatic video retrieval, restricting the search by the scope of meaningful semantic segments. Another potential area of application is an automatic generation of video summaries or skims that preserve the semantic organization of the original video.

As the basic building blocks of professional video are shots—sequences of contiguous frames recorded from a single camera—it is natural to divide a video into these units. Unfortunately, the semantic meaning they provide is at too low of a level. Common video of about one or two hours (e.g., a full-length film) usually contains hundreds or thousands of shots—too many to allow for efficient browsing. Moreover, individual shots rarely have complete narrative meaning. Users are more likely to recall whole dramatic events or episodes, which usually consist of several contiguous shots. In this work, we consider the task of automatic segmentation of narrative films, such as most movies, into something more meaningful than shots—high-level narrative units called scenes, or aggregates of shots unified by a common dramatic event or locale. We need shot segmentation at the first preliminary processing step since scenes are generated as groups of shots. Segmentation into scenes can be considered the next level of content generation, yielding a hierarchical semantic structure of video in which shots are preserved to form the lower level. In this work, we are not concerned with the problem of shot segmentation or adopting one of already existing techniques (Boreczky & Rowe, 1996; Lienhart, 1999) but rather focus on the task of video segmentation into scenes.

Sharing a common event or locale, shots of a scene usually are characterized by a similar environment that is perceivable in both the visual and audio domains. So, both the image sequence and the audio track of a given video can be used to distinguish scenes. Since the same scene of a film usually is shot in the same settings by the same cameras that are switched repeatedly, it can be detected from the image track as a group of visually similar shots. The visual similarity is established using low-level visual features such as color histograms or motion vectors (Kender & Yeo, 1998; Rasheed & Shah, 2003; Tavanapong & Zhou, 2004). On the other hand, a scene transition in movie video usually entails abrupt changes of some audio features caused by a switch to other sound sources and sometimes by film editing effects (Cao, Tavanapong, Kim, & Oh, 2003; Chen, Shyu, Liao, & Zhang, 2002; Sundaram & Chang, 2000). Hence, sound analysis provides useful information for scene segmentation as well. Moreover, additional or alternative features can be applied. For example, editing rhythm, which usually is preserved during a montage of a scene, can be used to distinguish scenes as groups of shots of predictable duration (Aigrain, Joly, & Longueville, 1997); classification of shots into exterior or interior ones would allow for their grouping into the appropriate scenes (Mahdi, Ardebilian, & Chen, 1998), and so forth.
On the Current State of Linked Open Data: Issues, Challenges, and Future Directions
www.igi-global.com/article/on-the-current-state-of-linked-open-data/210655?camid=4v1a