Chapter IX
Content-Based Video Semantic Analysis

Shuqiang Jiang
Chinese Academy of Sciences, China

Yonghong Tian
Peking University, China

Qingming Huang
Graduate University of Chinese Academy of Sciences, China

Tiejun Huang
Peking University, China

Wen Gao
Peking University, China

ABSTRACT

With the explosive growth in the amount of video data and rapid advance in computing power, extensive research efforts have been devoted to content-based video analysis. In this chapter, the authors will give a broad discussion on this research area by covering different topics such as video structure analysis, object detection and tracking, event detection, visual attention analysis, and so forth. In the meantime, different video representation and indexing models are also presented.

INTRODUCTION

In recent years, digital equipments such as digital camera and digital video camera have become very popular, which makes the digital video data growing in an exponential speed. Information in the Internet has also changed from one-fold text information to multimedia information such as graphs, image, music.
and particularly video. Broadly speaking, the Internet can be regarded as a database which contains a big size of video data. Digital video is playing more and more important role in human's life in terms of working, education, entertainment, and so on.

The increasing amount of digital video brings up a lot of problems which are never met before. For example, it is time-consuming and labor-intensive to manually manage the large scale, rapidly increasing digital video collection. As a result, it is more difficult for users to find the desired video contents from the large scale video database. Therefore it is crucial in nowadays to automatically understand digital video data and make them convenient for people to access. During the recent two decades, video content analysis and understanding has been an active research area, and a whole range of models and methods have been proposed. One of the most famous achievements is the Content Based Video Retrieval (CBVR). CBVR firstly extracts the low-level visual and/or acoustical features of videos, and then finds videos that have low-level features similar to those of the query example. As this technology needs little human intervention and does not take human's perception into consideration, it is not satisfactory in most real applications. In fact, human often judge the similarity of videos according to the objects they describe, the events they contain, and the feeling they express, rather than the video features such as color, shot, motion intensity. This difference between the low-level features and high level understanding is referred to as semantic gap, which in fact has become an important barrier of natural communication between computer and human. In order to bridge the semantic gap, lots of methods have been proposed, including relevance feedback, object recognition, event detection, and so on.

In this chapter, we will outline the structure for content-based video semantic analysis and discuss some related technologies from low-level content analysis such as feature extraction, structure analysis, object detection and tracking, to high-level semantic analysis such as event detection, attention analysis and video mining.

LOW LEVEL VIDEO FEATURE EXTRACTION AND REPRESENTATION

Generally speaking, low-level feature representation includes visual feature extraction, description, dimension reduction and indexing. After the video is segmented and key frames are chosen, low-level image features can be extracted from these key frames. Low-level visual features such as color, texture, edge and shapes can be extracted from the key frame set in video and represented as feature descriptors. After post-processing on the feature descriptors such as dimension reduction, they can be stored in the database using indexing models for future queries. There are two categories of visual features: global features that are extracted from a whole image, and local or regional features that describe the chosen patches of a given image.

Global Features

In the sense of human perception, an image can be characterized from various aspects, consequently leading to different image feature extraction methods. Typically, three categories of these methods are commonly used: color-based, texture-based and shape-based (especially edge-based) methods.

**Color**: Color is the most widely used feature for its close relationship with the objects and the scene in the image and its little dependence with the image scale, view direction and view point. Due to these