Chapter 2.14
Context-Based Interpretation and Indexing of Video Data

Ankush Mittal
IIT Roorkee, India

Cheong Loong Fah
The National University of Singapore, Singapore

Ashraf Kassim
The National University of Singapore, Singapore

Krishnan V. Pagalthivarthi
IIT Delhi, India

ABSTRACT

Most of the video retrieval systems work with a single shot without considering the temporal context in which the shot appears. However, the meaning of a shot depends on the context in which it is situated and a change in the order of the shots within a scene changes the meaning of the shot. Recently, it has been shown that to find higher-level interpretations of a collection of shots (i.e., a sequence), intershot analysis is at least as important as intrashot analysis. Several such interpretations would be impossible without a context. Contextual characterization of video data involves extracting patterns in the temporal behavior of features of video and mapping these patterns to a high-level interpretation. A Dynamic Bayesian Network (DBN) framework is designed with the temporal context of a segment of a video considered at different granularity depending on the desired application. The novel applications of the system include classifying a group of shots called sequence and parsing a video program into individual segments by building a model of the video program.

INTRODUCTION

Many pattern recognition problems cannot be handled satisfactorily in the absence of contex-

tual information, as the observed values underconstrain the recognition problem leading to ambiguous interpretations. Context is hereby loosely defined as the local domain from which observations are taken, and it often includes spatially or temporally related measurements (Yu & Fu, 1983; Olson & Chun, 2001), though our focus would be on the temporal aspect, that is, measurements and formation of relationships over larger timelines. Note that our definition does not address a contextual meaning arising from culturally determined connotations, such as a rose as a symbol of love.

A landmark in the understanding of film perception was the Kuleshov experiments (Kuleshov, 1974). He showed that the juxtaposition of two unrelated images would force the viewer to find a connection between the two, and the meaning of a shot depends on the context in which it is situated. Experiments concerning contextual details performed by Frith and Robson (1975) showed a film sequence has a structure that can be described through selection rules.

In video data, each shot contains only a small amount of semantic information. A shot is similar to a sentence in a piece of text; it consists of some semantic meaning which may not be comprehensible in the absence of sufficient context. Actions have to be developed sequentially; simultaneous or parallel processes are shown one after the other in a concatenation of shots. Specific domains contain rich temporal transitional structures that help in the classification process. In sports, the events that unfold are governed by the rules of the sport and therefore contain a recurring temporal structure. The rules of production of videos for such applications have also been standardized. For example, in baseball videos, there are only a few recurrent views, such as pitching, close up, home plate, crowd and so forth (Chang & Sundaram, 2000). Similarly, for medical videos, there is a fixed clinical procedure for capturing different video views and thus the temporal structures are exhibited.

The sequential order of events creates a temporal context or structure. Temporal context helps create expectancies about what may come next, and when it will happen. In other words, temporal context may direct attention to important events as they unfold over time.

With the assumption that there is inherent structure in most video classes, especially in a temporal domain, we can design a suitable framework for automatic recognition of video classes. Typically in a Content Based Retrieval (CBR) system, there are several elements which determine the nature of the content and its meaning. The problem can thus be stated as extracting patterns in *the temporal behavior of each variable and also in the dynamics of relationship between the variables, and mapping these patterns to a high-level interpretation.* We tackle the problem in a Dynamic Bayesian Framework that can learn the temporal structure through the fusion of all the features (for tutorial, please refer to Ghahramani (1997)).

The chapter is organized as follows. A brief review of related work is presented first. Next we describe the descriptors that we used in this work to characterize the video. The algorithms for contextual information extraction are then presented along with a strategy for building larger video models. Then we present the overview of the DBN framework and structure of DBN. A discussion on what needs to be learned, and the problems in using a conventional DBN learning approach are also presented in this section. Experiments and results are then presented, followed by discussion and conclusions.

RELATED WORK

Extracting information from the spatial context has found its use in many applications, primarily in remote sensing (Jeon & Landgrebe, 1990; Kittler & Foglein, 1984), character recognition (Kittler & Foglein, n.d.), and detection of faults and cracks