Flexible Querying Using Structural and Event Based Multimodal Video Data Model

Hakan Öztarak[1] and Adnan Yazıcı[2]

[1] Aselsan Inc, P.O. Box 101, Yenimahalle, 06172, Ankara, Turkey
hoztarak@aselsan.com.tr
[2] Department of Computer Engineering, METU, 06531, Ankara, Turkey
yazici@ceng.metu.edu.tr

Abstract. Investments on multimedia technology enable us to store many more reflections of the real world in digital world as videos so that we carry a lot of information to the digital world directly. In order to store and efficiently query this information, a video database system (VDBS) is necessary. We propose a structural, event based and multimodal (SEBM) video data model which supports three different modalities that are visual, auditory and textual modalities for VDBSs and we can dissolve these three modalities within a single SEBM model. We answer the content-based, spatio-temporal and fuzzy queries of the user by using SEBM video data model more easily, since SEBM stores the video data as the way that user interprets the real world data. We follow divide and conquer technique when answering very complicated queries. We give the algorithms for querying on SEBM and try them on an implemented SEBM prototype system.

1 Introduction

Since multimodality of the video data comes from the nature of the video, it is one of the important research topics for the database community. Videos consist of visual, auditory and textual channels, which bring the concept of multimodality [1]. Modelling, storing and querying the multimodal data of a video is a problem, because users want to query these channels from stored data in VDBS efficiently and effectively. In [5], a **s**tructural and **e**vent **b**ased, **m**ultimodal (SEBM) video data model for VDBSs is proposed with querying facilities. SEBM video data model supports these three different modalities and we propose that we can dissolve them within a single SEBM video data model, which makes us find the answers of multimodal queries easily.

Definition of multimodality is given by Snoek et. al. as the capacity of an author of the video document to express a predefined semantic idea, by combining a layout with a specific content, using at least two information channels, [1]. Moreover they give the explanations of the modalities that we use in SEBM as:

- *Visual modality*: contains everything, either naturally or artificially created, that can be seen in the video document;

H. Legind Larsen et al. (Eds.): FQAS 2006, LNAI 4027, pp. 75 – 86, 2006.

- *Auditory modality*: contains the speech, music, and environmental sounds that can be heard in the video document;
- *Textual modality*: contains textual resources that can be used to describe the content of the video document.

Nowadays researches are concentrating on efficient and effective ways of querying the multimodal data, which is integrated with temporal and spatial relationships. Modelling is as important as querying, because it is an intermediate step between data extraction and consumption. In general, researchers propose their querying algorithms with their data models. Snoek et. al. give the definition of multimodality and focus on similarities and differences between modalities in [1]. They work on multimodal queries in [18]. They propose a framework for multimodal video data storage, but only the semantic queries and some simple temporal queries are supported. They define collaborations between streams when extracting the semantic from the video. Oomoto et. al. don't work on multimodality but investigate the video object concept which is a base for spatio-temporal works [7]. Day et. al. extend the spatio-temporal semantic of video objects [17]. Ekin et. al. introduce object characteristics, and actors in visual events [4]. Köprülü et. al. propose a model that defines spatial and temporal relationships of the objects in visual domain which includes fuzziness, [3]. Durak in [2], extends the model proposed in [3]. She introduces a multimodal extension of the model and gives two different structures for visual, auditory and textual modalities. BilVideo is a good example for a VDBS, which considers spatio-temporal querying concepts, [8].

Main contributions of our work can be summarized as follows: In this study, we work on querying features of SEBM, which is based on human interpretation of video data. This interpretation is like telling what is happening in videos. If one can express information in digital world as human does in real world, then we think that all of the queries coming from a user can be handled more accurately and effectively. So we can bypass the problem of handling the models in different data structures and handle them separately as done in [2]. In SEBM, actor entities that are only defined for visual domains in [4] are modelled for multimodal domains. These entities give us the ability to express and query the structure of events in multimodal domains. Moreover object characteristics that involve a particular feature of an object or relation of an object with other objects are also introduced in SEBM for multimodal domains different than [7] which considers only visual domain. We propose some algorithms to query these stored relationships of video objects and events and follow divide and conquer approach in query processing to answer complex, nested, conjunctive, spatial, temporal, content-based and possibly fuzzy video queries. This approach gives us the ability to deal with much more complex and compound multimodal queries different than ones in [2] and [3]. We support these algorithms with an implemented querying prototype system that uses SEBM while modelling the data.

The rest of the paper is organized as follows: Section 2 presents how SEBM models the video data with exploring video segmentation, video entities and video actions. In Section 3, query processing on SEBM is investigated and content based, spatio-temporal, hierarchical and fuzzy queries are explored. Throughout the parts 2 and 3 the usage of SEBM is also explored. The last section provides conclusion with some future extensions of our model.