Media Streaming Synchronisation and Video Interaction: a Survey

R. Y. D. Xu[1], R. Wang[2] N. Parameswaran[2] and J. S. Jin[1,2]

[1]School of Information Technologies, The University of Sydney, NSW 2006 Australia

[2]School of Computer Science & Engineering, The University of NSW, Sydney 2052, Australia

{richardx,jesse}@it.usyd.edu.au,{s3053700,paramesh}@cse.unsw.edu.au

Abstract

Digital video has become an important element in a multimedia package. Most of the traditional video players and the video formats they support only provide ways of linear interaction, ie, a user can play the video in forward and backward direction or jump to a certain frame number indicated by a sequence or time on a random access device. This kind of linear interactions is also restricted to local content, ie, within a single video stream. Many attempts have been made to provide more flexible way to interactive with video. Such interaction, similar to hypertext systems available on the World Wide Web, provides multi-dimensional access cross various media in a distributed environment.

This paper discusses the fundamentals of video interaction, reviews several approaches of non-linear interaction of video streams and multimedia synchronisation, and presents our view of the direction of video interaction.

Keywords: interactive video, streaming media, video cataloguing, multimedia synchronisation

1 Introduction

Video contains rich information and it is difficult to interactive with a video stream. This is due to the organization of digital video. Video does not have the same structure and organization as text. Video bits are meaningless by themselves. The interactive information is difficult to be built and retrieved from video bits data. In the past, there are a few attempts in exploring and standardizing video interactions. MPEG-4 and MPEG-7 standards have been introduced two years ago for this purpose.

MPEG-4 [2001] enables interactivity and a combination of natural and synthetic materials, coded in the form of objects and allows IPMP for video information protection. MPEG-7 [2001] specifies a set of audiovisual description tools (the metadata elements and their structure and relationships, that are defined by the standard in the form of Descriptors and Description Schemes) to create descriptions (i.e., a set of instantiated Description Schemes and their corresponding Descriptors at the users will), which will form the basis for applications enabling the needed effective and efficient access (search, filtering and browsing) to multimedia content. MPEG-21 [2002] standard will be released in future with more focus on interoperability for multimedia food chain from supplier to end-users.

Although MPEG is a large organization, and its standards are easier to be adopted by multimedia suppliers and consumers, however, it has its own drawbacks. The major one of course is unable to support video format changes. For example, the video would lose all its interaction information if MPEG-7 format were exchanged to another video format. This could be a huge problem for hardware manufacturers to adopt new MPEG standard, considering most of the videos in the past have been built using mainly MPEG-1 and MPEG-2 standards. Secondly, the new changes in MPEG-4 and 7 standards are quite complex for hardware and software implementations. Thirdly, secure information using MPEG-7 cannot be transparently built into the video and it may attract deliberate attacks. Lastly, information embedded in MPEG-7 is still considered as metadata description. They are not robust. Think of a scenario where a user has captured an image or a short sequence of video using an MPEG-7 capable video player, then information associated with those captured video sequence or images can no longer be associated with them.

Much research is needed to achieve flexible access video streams and provide multiple linkages from video to other media. This paper provides a general survey of the field and related issues. It is organised as follows. Section 2 discusses the fundamentals of video interaction, its definition, aims and functionality. Sections 3, 4 and 5 review several approaches of non-linear interaction of video streams and multimedia synchronisation. Section 6 reveals the related field. In the conclusion, we present our view of the direction of video interaction.

2 Video Interaction

An interactive video is a digital video with a hyperlink type of interaction capabilities for browsing. Like hyperlinks in a hypertext document, interactive videos contain hyperlinks in a video stream to provide users with interactive maenuvier over variety of document. The hyperlinks are attached to selectable objects, sprites, regions, frames and shots within a video sequence, and activated by interactive selection. Video hyperlinks provide the user with dynamic access and better comprehension of the video. For example, it can activate a formula, a web page, an image, an audio file or even another video, to provide extra information, or a questionnaire and switch to a relevant site according to the answer.

An interactive video contains four components:

- a digital video stream,
- selectable video components such as region, objects, sprites, etc
- hyperlinks or tags,
- the definition of activation.

This research is inspired by HTML. There is long history to use interactive video in psychology study [Dowrick 1991]. However, the interaction in this circumstance is merely controlled by an operator while our research is trying to provide a user-centralized interaction.

3 VAML Project at UNSW

A project was conducted in the University of New South Wales to develop a video annotation mark-up language dating back to 1996 (http://www.cse.unsw.edu.au/~vip/#vp). It aimed to provide an annotation scheme for distance education. In current distant learning program the students receive video types. They play video at home. This is passive learning and we are not able to customize the video to reflect the individual needs. If we annotate the video to include the intra-relation of video frames and inter-relation between video and other media (audio, text, 3D animation, other video sequences), the student will be able to select the subject and control the pace himself or herself. Annotating a video sequence involves many issues, e.g., synchronization between audio and video, hyper-link to the external text source, buffers for multiple streams of video, segmentation of video objects, representation of video sequences, and the organization of source data. The project has developed a demonstration prototype but the annotation using tracking has proved very time-consuming. Many problems remain unsolved. Some issues have been discussed in Liu and Jin [2001].

4 MOVieGoer System

MOVieGoer interactive video system was developed to provide extra information over the video stream [Finke & Gerfelder 2000]. It is based on a Server-Client architecture. When a user interacts with the video content, i.e. when he clicks on an object within a video scene, interaction data is generated and transmitted to the server's side, consisting of the following parameters:

- spatial and timing information,
- the user profile ID, and
- the title of the movie.

The process handler on the server's side determines data and definition of the selectable objects at the specific interaction. If the chosen object is determined and the result is positive, the process handler provides additional information, filtered by the user and device profile, and sends the result to the client application where it is presented.

It is implemented as an open system, i.e. each of the three main data elements - video content, selectable objects, hyperlinks - is handled as separate data types. The advantage of this concept is that the actual video data is not modified, neither by information about selectable objects, nor by hyperlinks. Therefore, existing and upcoming video formats can easily be integrated and used. It also has some drawbacks which will be discussed in Section 5.

Within the MOVieGoer project, an authoring tool is developed. Supplying the MOVieGoer system with an own authoring tool gives the content provider the advantage of a completely independent system. The authoring tool which is used for producing interactive digital videos for the MOVieGoer environment is called MOVieEditor. The input for this tool is a digital video that can be of different formats. The MOVieEditor generates two data files for the definition of selectable video objects and the additional information linked to them. Since a video object is linked with additional information, the system must know the location of the object at any time. Therefore, object tracking is needed. As was discussed before, tracking an object in a video stream remains one of the major difficulties in authoring interactive videos. Tracking an object manually in each frame of a sequence is too time-consuming and therefore ineffective. The MOVieEditor system provides a key frame method, which is a half-automatic tracking method.

5 Streaming Synchronisation

There has also some multimedia integration languages and its commercial software been introduced recently for interactive videos. Some well known examples include BIBS, SAMI and SMIL.

BIBS - Berkeley Internet Broadcasting System presents a way to receive and browse a host video from a client side [BIBS 2001]. It synchronizes audio and video with other media (e.g., slides, which can be viewed from a HTTP browser), interaction with remote participants (e.g. asking questions during a live lecture) and tools to automatically generate keyword indexes using speech-to-text conversion of a lecture audio track. It uses Java-script on the client side to interact with host video using standard HTTP browser.

A simpler video interactive method that inspired by BIBS was proposed by Jin and Wang [2001]. It used a controlled module on the client's side, which is able to configure its browser to retrieve certain clips in the video and its captions at the server. Also browsing is done using metadata information on the client side on an interactive tree using JavaScript. Each individual clips were segmented and encoded on the host side.

More schematic approach of interactive video authoring tool can be found in both Microsoft Synchronized Accessible Media Interchange (SAMI) and The Synchronized Multimedia Integration Language (SMIL).

SAMI [2001] is a tool that allow closed caption to a video, SAMI was designed and developed to caption the digital media available in PC systems. SAMI captions coexist with digital media as separate text files. The captions can be modified, maintained and localized for different languages. SMIL [2001] enables authoring of interactive audiovisual presentations. SMIL is used for multimedia presentations, which integrate streaming audio and video with images, text or any other media type [SMIL 2001]. SMIL is a HTML-like language. In 2001, SMIL 2.0 Define an XML-based language that allows authors to write interactive multimedia presentations. Both SAMI and SMIL use a HTML-alike text file. It contains what the video interaction information only in video's temporal domain.

The advantages of these two methods are text files can be edited and maintained easily without the presence of the host video. And interaction information can be easily retrieved and edited. However, it has some major drawbacks. Firstly, it lacks of the spatial information. It is difficult to have a separate file describing the spatial information of the video. Secondly, there is a security problem. A text file can be easily corrupted and deliberate attacked may occur in some applications. This may cause the whole interactions to be erased or replaced easily. Thirdly, when the host video is edited by other video player/editor without updating of SMIL file. E.g., video maybe resized, and certain frames have been cut from the original video. This will cause the SMIL or SAMI to have incorrect references.

6 Semantic Cataloguing Video Content

One topic closely associated with interactive video is semantic cataloguing video content. Although interactive video provides a multi-dimensional access to video streams, without a proper semantic catalogue, the multiple accesses are meaningless. Few attempts of cataloguing video streams have been investigated in developing content-based video information retrieval methods, and query by video's features [Hampapur & Jain 1998;].

Tandiaus et al. [2001] presents video using a fixed three-level hierarchical structure, users can access and retrieve specific video information through browsing of pre-generated metadata on the existing ViMeta-VU system. Users use XML to represent such data. Video frame image texture analysis was performed using Gabor filters. Query methods include query by feature and selection the structure browsing. This paper shows how a video's metadata can be built from both spatial and temporal domain, which provides a semantic structure of video. However, the interactions are still concentrated on video's temporal domain. The video information is still described in its metadata as a separate data to the actual video bits.

7 Conclusion

The major aim of digital TV is not for the quality of picture but the interaction between audience and the TV. Interactive video is a necessary technique for future digital TV. We have reviewed the development of interactive video and related fields. In conclusion, we suggest that:

- it is necessary to develop a standard interaction protocol for video streams;
- the interactive tag should be embedded into the video streams;
- the design of interaction should be consistent with the semantic structure of video streams.

To meet this need, much research has to be done in the areas of video segmentation, object tracking, sprite modelling, tag embedding, semantic cataloguing, and multimedia synchronisation.

References

BIBS (2002) The Berkeley Internet Broadcasting System http://bmrc.berkeley.edu/bibs/

Dowrick, P. W. (1991). *Practical Guide to Using Video in The Behavioural Sciences*. Chichester: John Wiley & Sons.

Finke, M & Gerfelder, N (2000). Video interaction and information personalization for new interactive broadcast services, CG topics 3/2000, pp.27-30. (also www.inigraphics.net/publications/topics/2000/issue3/3_00a09.pdf)

Hampapur, A. & Jain, R. (1998). Video Data Management Systems: Metadata and Architecture, in *Multimedia Data Management*, A. Sheth, W. Klas (eds.), McGraw-Hill, 1998.

Jin, J. S. & Wang, R. (2001) The Development of an Online Video Browsing System, *Conferences in Research and Practice in Information Technology (Visualisation 2001)*, pp.3-9.

Liu, C. R. & Jin, J. S. (2001). Modelling and design of VAML, *Conferences in Research and Practice in Information Technology (Visualization 2001),* vol 11, pp.151-152.

MPEG-4 (2001). Overview of the MPEG-4 Standard, http://mpeg.telecomitalialab.com/standards/mpeg-4/

MPEG-7 (2001). Overview of the MPEG-4 Standard http://mpeg.telecomitalialab.com/standards/mpeg-7/mpeg-7.htm

MPEG21 (2002). From MPEG-1 to MPEG-21: Creating an Interoperable Multimedia Infrastructure, http://mpeg.telecomitalialab.com/documents/from_mpeg-1_to_mpeg-21.htm

SAMI (2001) Microsoft® Synchronized Accessible Media Interchange (SAMI) V1.0 October 2001 http://msdn.microsoft.com/library/default.asp?url=/library/en-us/dnacc/html/atg_samiarticle.asp

SMIL (2001). Synchronized Multimedia Integration Language (SMIL 2.0) Specification, http://www.w3.org/TR/smil20/

Tandianus, J. E.; Chandra, A. & Jin, J. S. (2001). Video Cataloguing and Browsing, *Conferences in Research and Practice in Information Technology (Visualisation 2001)*, pp.39-45.