A MULTI-LANGUAGE SYSTEM FOR KNOWLEDGE EXTRACTION IN E-LEARNING VIDEOS

By

APARESH SOOD * ANKUSH MITTAL ** DIVYA SARTHI ***

* - *** Department of Electronics & Computer Engineering, Indian Institute of Technology Roorkee, Roorkee.

ABSTRACT

The existing multimedia software in E-Learning does not provide par excellence multimedia data service to the common user, hence E-Learning services are still short of intelligence and sophisticated end user tools for visualization and retrieval. An efficient approach to achieve the tasks such as, regional language narration, regional language captioning system and keywords based efficacious seek to save the bandwidth, is introduced in this paper. The goal is to reduce the language barrier and seeking time for E-Learning videos. The paper presents an innovative solution of E-Learning multimedia delivery package for client-server model.

The functioning of the software tool can be described in different modules. Initially the system integrates keywords enriched subtitle stream with the video file at the server side. At the client side, the algorithm parses the video file into different streams. The synchronized text stream is then transformed into the uncompressed video stream and later overlaid onto the primary video to reproduce regional language captioned video. The system also encompasses a text annotation pad with intention to facilitate the future comprehension. Moreover, the user can upload his notes on server to make it available publicly. Regional language narrator is one of the advanced features supported by the system. Component Object Model (COM) has been used extensively for extending the system to include multiple numbers of language translators and narrators at both the server as well as the client sides.

Keywords: Filter, Filter Graph, Annotation, Captioning and Speech Synthesizer, COM (Component Object Model).

INTRODUCTION

In developing countries, a majority of population dwells in non-urban setup where the educational infrastructure and resources are usually meager and scanty. Trained teaching faculty at primary, secondary and technical educational levels are lacking. The students who come out from such background are less likely to excel than those who are exposed to the best education. However, internet based distance learning (E-Learning) is changing the global scenario occurring in education today. E-Learning techniques [1, 2, 3, 4] like virtual classrooms, provide the exposure to quality education and are very beneficial to students in remote educational institutions. Many institutes such as MIT (USA) and IIT Delhi (India) have opened their web servers for free lecture-on-demand on several courses. Communication and advanced computer technology enable common user to receive instruction despite geographical and time disparity that would otherwise traditional classroom instruction impossible. Nevertheless, the bandwidth restriction on communication channel makes E-Learning a challenging problem.

Research has shown that audio/video mediated communication has helped students achieve higher grades than conventional classroom lecture [12]. Several works also show that recently e-Education has got huge potentials to govern the education system on the global basis in the near future. Within a short span of 9 months, a total of 870 courses (that accounts for 51% of Nanyang Technological University, Singapore (NTU) courses) went on-line by end March 2001, resulting in a large quantum increase from zero courses in early July 2000. Statistics show that edveNTUre (NTU E-Learning client-server system) receives average 30,000 to 80,000 hits per day. Above statistics supports the facts that on-line learning services expose students to new learning approaches where they acquire skills for life-long learning, a critical asset in today's

knowledge economy.

E-Learning is now considered to be one of the more significant and growing research and application areas of multimedia computing. Many areas like data compression, data streaming, data retrieval, etc. are demanding a lot of research efforts in the field of E-Learning. Some of the presently commercial available Web Learning Environment products include WebCT [9], Blackboard [10] and Cisco IP/TV [11]. However, all the products share one common shortcoming among them, that is, they all mainly employ outdated low level representation techniques, i.e., the knowledge access is restricted to a "one-step" level. Even publicly available conversant web search engine as Google and Yahoo! employ general manual filtering process to get a short relevant matching list.

Apart from the technology limitations the major hurdle in accomplishment of worldwide establishment of E-Learning education system is the semantic gap between inhabitants at different geographic location. The objective is to furnish display of subtitles in the regional language along with the video/audio at real time. However, the exact conversion of the speech is not requisite to the extent that the delivered idea remains the same.

The motivation of work is to devise low cost and user-friendly techniques which are capable of rendering multilingual captions (subtitles) along with the video. The adoption of presented technique would make E-Learning application more appealing for the masses. However, the technique is not only restricted to E-Learning application but it can also be extend to transmit other video data like agricultural science, news, sports, oration, medical, etc. Out proposed scheme is very efficient both in terms of storage requirement at server as well as processing required at the client. Efficiency in storage results in low download time of video and low processing requirement at client side enables real time usage of the software even in resource scanty devices such as PDA's and other handheld devices.

The rest of the paper is organized as follows. Section II highlights some of the well recognized past work. Section III discusses integration of regional language text stream with primary video data to reproduce the captioned video. Section IV gives a brief overview of Microsoft® DirectShow® to illuminate the concepts of Filters, Filter graph and Multimedia Data Transformations. Section V explains the functioning of our client-server based software product which incorporates regional language narrator. Section VI covers in-depth elucidation of the highlevel file format of video which is introduced in the software solution. Section VII discusses the interactive graphical user interface of client side multimedia player with some snapshots of the running video. Some of the possible applications of proposed product have been covered in section VIII. Conclusion and future work follow in section IX.

1. Previous Works

Many researchers have addressed the issues related to automatic captioning and annotation. Wakamatsu et al. [13] had designed Video Caption Markup Language (VCML) (based on XML 1.0) and developed VCML player which can play video data with captioning according to VCML document. Gao et al. [14] introduced the concepts of keywords-based news story indexing and retrieval. Smith et al. [16] have presented the idea of "annotation as argumentation" to help the learners to articulate more than contents summary. Shih et al. [17] proposed multistory annotation system specially designed for distance learning application. In an approach, Wilcox et al. [18] proposed method for indexing and retrieval of multimedia data based on annotation and segmentation. However, the aforementioned works fail to recognize the importance of regional language captioning and the retrieval techniques used by them are solely dependent on English language texts only.

2. Text Assimilation

Text assimilation is a vital processing step at the server end. Text assimilated with video data stream is used to create the subtitle promptly at the time of playback. Such text-assimilated video file makes the file management facile and reduces the incurred transmission overhead as compared to the other possibility of synchronous transmission of separate text and video files. The format of the assimilated text is so as to allow identification of the keywords within the text stream. These keywords can then assist in extracting useful information during video

playback.

Many video standards allow integration of various data streams (text, etc.) in addition to the video/audio stream. Commonly used standards include MPEG (Motion Picture Expert Group) and AVI (Audio-Video Interleaved). An overview of these standards is provided in subsequent subsections.

A. MPEG Video and Text Embedding

Define in ISO/IEC 11172, MPEG [5, 6, 7] is an international standard for coding of moving pictures and associated audio for digital storage media. The section 1 of part 1 of the two part standards specifies the system-coding layer. It defines a multiplexed structure for combining elementary streams, including coded audio, video and other data streams, and specifies means of representing the timing information needed to replay synchronized sequences in real-time.

MPEG system specifies the syntax and semantics of information that is necessary to reproduce data streams in a system. An ISO 11172 stream consists of one or more elementary streams multiplexed together and organized into two layers: the pack layer and the packet layer.

The pack layer is for system operations and the packet layer is for stream specific operations. Figure1 shows a logical MPEG pack. An ISO 11172 stream consists of one or more packs. The pack header stores system clock reference and bit rate information and mux_rate specifies the rate at which the decoder receives the ISO 11172 multiplexed stream during the pack. The system header indicates decoding requirements for each of the elementary streams. It stores data rate, the number of streams, and the buffer size limits for the individual elementary streams. The data from elementary streams are stored in packets. A packet consists of a packet header, which identifies the stream, followed by packet data.

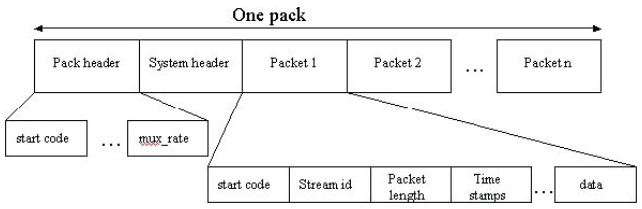

Figure 1. A MPEG pack containing various packets

In order to add a text stream for subtitling in a MPEG file, text packet stream has to be created. Some of the issues involved in text stream creation are updating the mux_rate, system header for all packs including the packet stream, buffer size and time-stamps for the new packet stream. Fig. 2 shows the addition of text stream in a pack.

B. AVI Video and Text Embedding

The Microsoft AVI (Audio Video Interleaved) format is a RIFF (Resource Interchange File Format) file specification. A chunk is a logical unit of multimedia data and is the basic building block of a RIFF file. Each chunk incorporates FOURCC (four-character code) type chunk identifier, data size and data. The AVI RIFF form is identified by the FOURCC 'AVI ' as the chunk identifier. All AVI files include two mandatory LIST chunks, namely 'hdrl' and 'movi'. The LIST 'hdrl' chunk defines the format of the data. The LIST 'movi' chunk contains the data for the AVI sequence. In addition, "RIFF" and "LIST" chunks can contain sub-chunks.

The AVI file begins with the main header with identifier 'avih'. This header contains global information for the entire AVI file, such as the number of streams within the file, the width and height of the AVI sequence, suggested buffer size etc. One or more 'strl' chunks follow the main header. Each 'strl' chunk must contain a stream header (FOURCC 'strh') and stream format chunk (FOURCC 'strf'). Chunk 'strl' might also contain a stream-header data chunk (FOURCC 'strd') and a stream name chunk (FOURCC 'strn'). The stream header specifies the playback rate for the stream, the type of data the stream contains, etc. In order to add text stream for subtitling in an AVI file, the system header needs to be updated and 'txts' stream header (text stream) and 'txts' stream data need to be added. To provide high-quality video and audio playback or capture, Microsoft's DirectShow is used in design and is discussed in the following section.

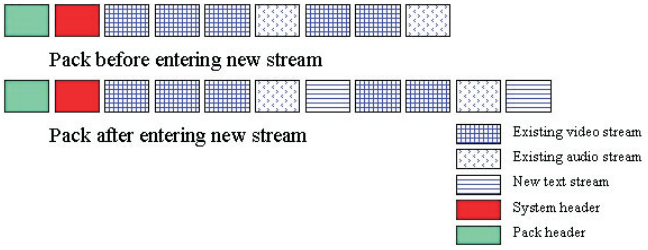

Figure 2. The addition of text stream

3. Microsoft® Directshow®

The Microsoft's DirectShow API (Application Programming Interface) is a media-streaming architecture for the Microsoft Windows platform. The building block of DirectShow is a software component called a Filter. The data travels from source to sink through filters chained together to form a filter graph. All data in DirectShow is streamed between filters. A filter can be written to intercept data, manipulate it in some way, and pass it downstream. Filters connect with one another in the standard way defined by COM.

Most common types of filters present in this architecture are source filter (introduces data into the filter graph), parse filter (separates all the constituent data streams from the incoming data), transform filter (takes an input stream, processes the data, and creates the output stream) and render filter (receives data and presents a stream to the user). Any filter chain has source filter(s) as the first filter and the render filter(s) at the end. The connection points on a filter are known as input and output pins, and a filter can have several of them. All input/output of stream data is done through these pins. Pins have media types associated with them, which include major type, subtype and format type. For details please refer to Microsoft® DirectShow® documentation [xxxx].

4. Methodology

A video along with 'multi-lingual subtitles' can be rendered simultaneously and synchronously by creating a new DirectShow transform filter. Such a transform filter must be capable of processing text stream to generate uncompressed video stream. The logical view of whole process can be understood from Figure 3. The scheme proposes design of TB filter (Transform Filter) which works as a bridge between parse filter and video renderer. The input pin of TB filter takes the text stream from the parse filter and its output pin is connected with the input pin of video renderer, to provide the uncompressed video stream to the renderer. The connection establishment demands negotiation between peer filters, i.e., between parse filter and TB filter, and between TB filter and video renderer. In the negotiation process the downstream filter checks for compatibility with the upstream filter, i.e. it checks for at least one common supported stream type. After the compatibility check, TB filter negotiates with the allocator, which manages the buffer space between TB filter and downstream filter. Once the buffer space is decided the setup is equipped to work.

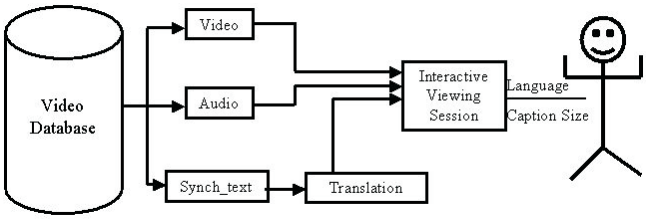

Figure 3. Logical View of Regional Subtitling system

The TB filter exposes some COM interfaces for providing access to its features like the selection of languages, fonts, background colors, enabling/disabling of the subtitles, etc. These features will affect the mapping procedure from text in Unicode format to the video subtitle in uncompressed video format.

The TB filter has been suited to extract keywords from the text stream. These keywords can be highlighted for better user understanding. Moreover, the keywords can be passed on to the media player for providing keyword-specific video seek as well. The logical position of TB filter in the filter graph can be seen through graph edit tool provided with DirectX. Figure 4 shows the filter graph for a video with two streams, namely a video stream and a text stream for subtitles. A parse filter separates all the streams from the file and provides the input for the TB filter. The output of the TB filter will be passed to the rendering filter for display. A rendering filter which is capable of accepting multiple streams and provides mixing and blending features can be used to superimpose subtitles on the primary video. Figure 5 shows a high level view of the work done by the TB Filter.

To facilitate the comprehension the narration feature in regional language is provided. The logical implementation of narrator is shown in Figure 6.

The language specific text parser filter rejects the unwanted

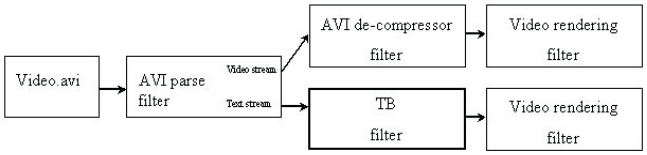

Figure 4. Filter Graph

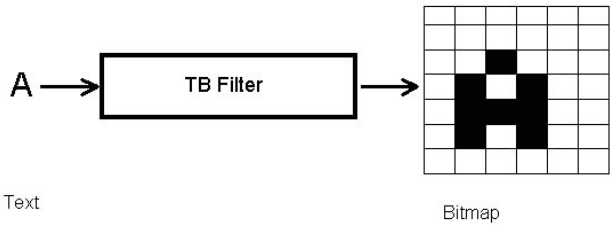

Figure 5. Logical representation of TB filter's Work

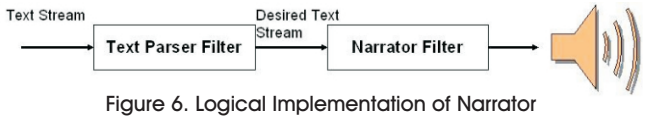

Figure 6. Logical Implementation of Narrator

text streams and passes only the desired language text stream to the narrator filter. Narrator filter is a language specific text to speech converter. For each desired language a separate audio narrator is required.

5. Video File Format

The need of the hour is the video file format that can support both the keyword indexing mechanism and Text Annotation. In our approach we have devised a video file format which has provision to support both of our needs. The format is shown in Figure 7.

The system header contains the system specific details and flags such as total number of streams, location of various other sections in the video file, maximum buffer size required, etc. The stream header is particular to the stream such as video stream header which includes width and height of video, compression algorithm used, frame rate, etc. Followed by the stream headers is the start of the actual multimedia data that is packets of audio/video and text streams, which may be interleaved. The text stream, described in the previous section, contains subtitles of video in one or more languages. The table of contents (TOC) created at the time of initial video files processing at the server end is a read-only part of the file. It contains the information about the organization of contents of the lecture video as per the time scale. This information is made available at the user end for the full length of the video. This information is helpful to the user to access the desired part of the video on time seek basis. The keyword index is a created database which outputs the time of the appearance of the queried keyword in the captioned video. The system also highlights the TOC entries which are relevant to the queried keyword. It helps the user to seek the video based on keyword query. The flowchart of keyword query mechanism is illustrated in Figure 8.

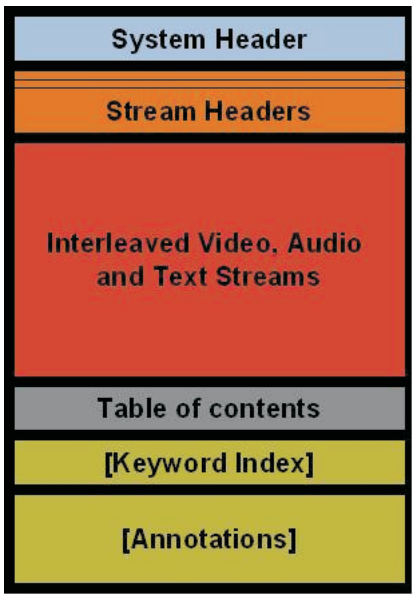

Figure 7. Video File Format

Annotation is an optional part in which the notes can be created by the user to better comprehend the video. The storage format of annotation is shown in Figure 9.

The language part specifies the language used. It is to assist the translation of annotation/comment in different language. The presentation time specifies the beginning and the end of the annotation in the video. Compression

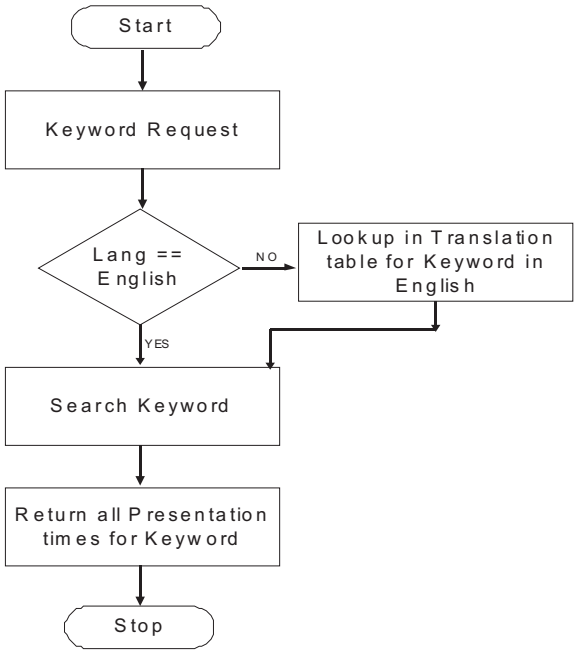

Figure 8. Flow chart of the Keyword Query Mechanism

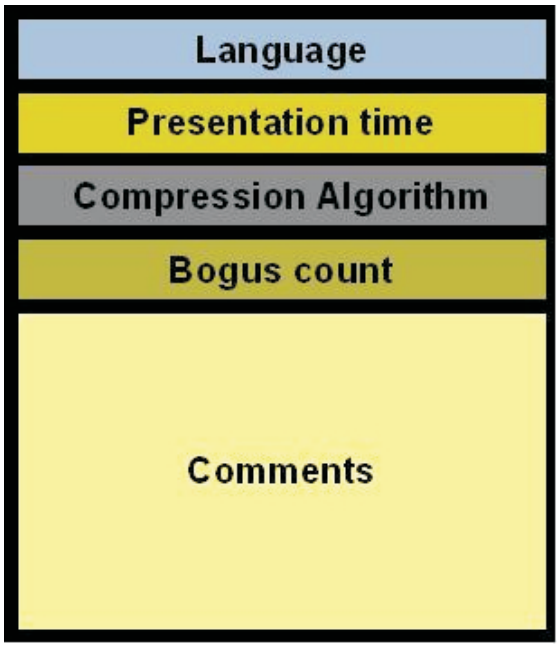

Figure 9. Storage format of Annotation

algorithm part signifies the presence/absence of the compression and type of compression for the annotation text [15]. Bogus count keeps the track of the number of hits on the annotation as bogus. This helps to remove the counterfeit comments from the video. The comments section of the format contains the actual comments on the video. It may contain hyperlinks also. Figure 10 shows an example.

6. Media Player

The Media Player provides a user interface and means for setting the subtitle stream in addition to the video and subtitle's playback along with optional narration at the client side. The player is able to pass the user's language preference to the filter through one of the exposed interfaces of the filter which is discussed in the previous section.

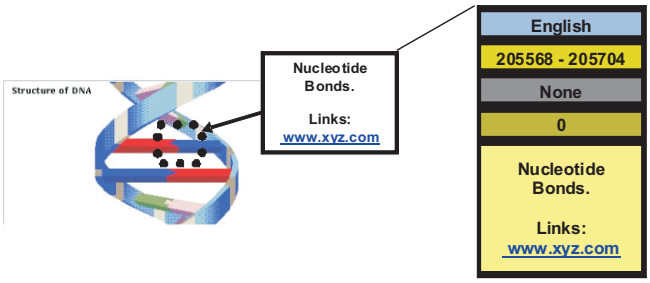

Figure 10. Example of Annotation with storage format

The player creates the Filter Graph Manager and builds an empty Filter Graph. The video to be opened is then parsed for the streams. The codec for all the streams are searched through the registry of the system and is added to the filter graph. For the text stream, the Media Player selects the TB Filter. A video renderer and subtitle's location is chosen depending on the graphics capability of the user's system and his preference. Figure 11 shows a video wherein a teacher is giving lecture. Figure 12 shows the same video from the player when the subtitles are enabled and the language chosen is Hindi.

7. Applications

The proposed technique will be beneficial in many areas. The software package can be used advantageously in various fields including sports, farming videos, news, etc. In sports it can be used to translate the commentary in the

Figure 11. Video without subtitle

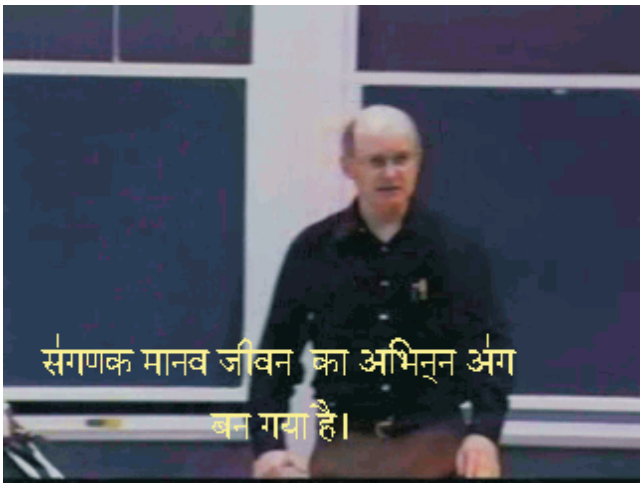

Figure 12. Video with subtitle in Hindi language

user desired language. The farming videos with different regional languages can generate the desired education for diversified set of people.

The proposed technique will be very beneficiary for news videos as well. The technique is also a boon to physically challenged people with low auditory response. Moreover, the narrator feature will be able to aid blind people as well.

8. Conclusion and Future Work

Language barrier is of prime concern in knowledge assessment process. We have presented a way after exploiting existing technology along with some innovative ideas to design regional language narration and subtitling system. The technique is very efficient in terms of storage requirement and can be employed for any internationally acclaimed language supported under UNICODE [8]. Once subtitle stream is added to a video file in one language, subtitles in other languages can be inserted automatically using language translators and automatic language summarizers. The system does not put any limit on the number of languages that can be supported simultaneously for such a system. Also, the facility is provided to the user to change the subtitle language during runtime. Video searching, with the help of keywords present in subtitle stream, is also a potential facility of our design. The advantages of the proposed method over conventional methods are summarized in Table 1.

The text assimilation process can provide security by encrypting the text. System can easily deliver facilities like different font colors to increase the contrast between background and subtitled text as well as the choice of subtitle location in the video.

Advantages of proposed method over conventional methods

Feature supported	*Proposed method*	*Conventional method*
Keyword specific video seek	Yes	No
Language specific subtitle's summarization	Yes	Possible
Modifiable subtitle's orientation during playback	Yes	No
Closed captioning support	Yes	Not always
File size compared to original video file (In discussed file format)	Almost equal	More
Supports automated multi-lingual audio generation	Yes	Not possible
Provision for annotation	Yes	No

Table1. Advantages of proposed method over conventional methods

References

[1]. **H. Ahmad, Z. M. Udin, and R. Z. Yusoff,** *"Integrated process design for E-Learning: a case study"*, The Sixth International Conference on Computer Supported Cooperative Work in Design, pp. 488 – 491, 2001.

[2]. **P. R. Polsani,** *"E-Learning and the Status of Knowledge in the Information Age",* International Conference on Computers in Education, pp.952 – 956, 2002.

[3]. **K. Seki, W. Tsukahara, and T. Okamoto,** *"System Development and Practice of E-Learning in Graduate School",* Fifth IEEE International Conference on Advanced Learning Technologies, ICALT, pp. 740 – 744, 2005.

[4]. **S. M. Jeong and K.S. Song,** *"The Community-Based Intelligent e- Learning System",* Fifth IEEE International Conference on Advanced Learning Technologies, ICALT, pp. 769 – 771, 2005.

[5]. **G. J. Lu, H. K. Pung, and T. S. Chua,** *"Mechanisms of MPEG Stream Synchronization",* ACM SIGCOMM. Vol 24, Issue 1, pp. 57-67, Jan 1994.

[6]. **M. Azimi, P. Nasiopoulos, and R. K. Ward,** *"Implementation of MPEG System Target Decoder",* Canadian Conference of Electrical and Computer Engineering, Vol 1, pp. 943-948, 2001.

[7]. **P. N. Tudor,** "Tutorial MPEG-2 video compression", *Journal of Electronics and Communication Engineering,* pp. Dec 95, Available online at http://www.bbc.co.uk/rd/pubs/papers/paper_14/paper_14.html (last accessed 15th March 2006)

[8]. http://www.unicode.org/ , last accessed 15th March 2006

[9]. **WebCT,** http://www.webct.com/, last accessed 15th March 2006

[10]. **Blackboard,** http://www.blackboard.com/, last accessed 15th March 2006

[11]. **Cisco IP/TV application,** http://www.cisco.com/ , last accessed 15th March 2006

[12]. **C. S. Lee and T. H. Tan,** *"Humanizing E-Learning,"* International Conference on Cyberworlds, pp. 418 – 422,

2003

[13]. K. Watanabe, N. Fukada and M. Sugiyama, *"Design of Video Caption Markup Language VCML and development of VCML player,"* IEEE conference on Multimedia and Expo (ICME), Volume 1, pp.163-166, July 2000.

[14]. G. Xinbo, X. Hong and J. Hongbing, *"A study of intelligent video indexing system,"* Proceedings of the 4th World Congress on Intelligent Control and Automation. Vol. 3, pp. 2122-2126, June 2002.

[15]. S. Das, S. Manna and U. Garain, *"Compression of Indian Language Electronic Text: A case study on Bengali text corpus,"* International conference on Emerging Applications of Information Technology, pp. 11 – 14, 2006

[16]. B. K. Smith, E. Blankinship and T. Lackner, *"Annotation and education,"* IEEE Multimedia, Volume 7, issue 2, pp. 84 – 89, April-June 2000.

[17]. T. K. Shih, Y.-C. Liao, H.-B. Chang and M.-Y. Kuan, and G. Yee, *"Multistory annotation system: A novel application of distance learning,"* 18th International Conference on Advanced Information Networking and Applications (AINA). volume 2, pp. 116 – 119, 2004.

[18]. L. Wilcox and J. Boreczky, *"Annotation and segmentation for multimedia indexing and retrieval,"* Proceedings of the Thirty-First Hawaii International Conference on System Sciences. pp. 259 – 266, volume 2, Jan. 1998.