Synchronization of Multimedia Events based on Semantic Extraction –A Survey

S.Vigneswaran
Department of Computer Applications, Anna University Regional Centre Coimbatore.

A.Leelamani, Ph.D
Assistant Professor of Mathematics, Anna University Regional Centre Coimbatore.

K.Divya
Senior Research Fellow, Tamilnadu Agricultural University,Coimbatore.

ABSTRACT
This paper explores the various synchronization schemes of multimedia events and how they are synchronized. Mostly the synchronization is based on the ECA rules [Event, Condition and Action]. When multimedia streams are presented to the user it should be synchronized, even though it is in distributed system. A distributed multimedia system involved for ECA that should be synchronized while presented to the user. Based on certain constraints like time, space and user requirements, multimedia events are also synchronized at the time of deliver on the display. User can apply their own synchronization technique that is developed for their self usage else the events itself will be synchronized automatically.

General Terms
Event, Intra stream, Inter stream, Ontology, Frames

Keywords
Multimedia synchronization, Multimedia Streams, Semantic web

1. INTRODUCTION
An interactive multimedia presentation in a distributed multimedia system requires synchronization of media streams, preprocessing media for content-based retrieval and low bandwidth transmission over network and user interface for interacting multimedia presentation. In multimedia synchronization, we handle the synchronization by synchronization rules which include events, conditions and actions. The synchronization works are divided based on how the data are generated, on demand and real-time. In this research, we are mainly interested in the real time category. Inside this branch, there are two categories: the intra stream, which is carried out in one stream, and the inter stream, which is carried out among various streams in order to maintain the connection of the applications.

A timeline for events and actions have to be generated to handle user interactions like skip and change direction. In video processing, video streams are compressed using Discrete Cosine Transform (DCT), which express a sequence of finite streams in terms of sum of cosine functions which gives the exact event(s) oscillating at different time intervals. The background generation can be performed by grouping of similar blocks at different intervals of the video stream. Information storage means for storing data in the form of information blocks and control information data for managing the information blocks, including a multilevel composition data showing how the information blocks are linked to each other. The group of related streams called clusters, which contains more elements than others, is the candidate cluster which contains the blocks for the background. The extraction of moving objects can be performed by using the background model, in which objects appears with some indirect properties and will be passive in nature. Common patterns in objects and background produce shadow this is a major problem. To reduce this problem the edges of background and frames, which is a collection of related pixels, are compared. If object and background edges are similar the object edges may be removed.

The morphological operations such as reduplication or conversion are applied to detect video object after the detection of edges but it doesn't consider the contents of frame .When the edges are eliminated or cannot be detected the highest gradient in the neighbor pixels are used to connect the edges. If a closed boundary is created the inside region is filled with the object. In High-level user interface, most of the user interfaces uses interfaces like VCR [Videocassette Recorder] –type that is used to store and retrieve data for multimedia presentation. The multimedia presentation is presented like a book which is more convenient to the users.

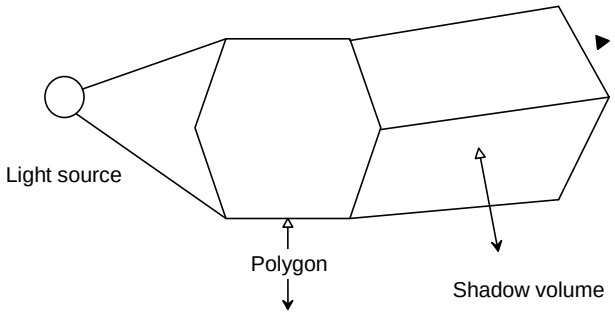

Figure 1 : Example - Shadow problem

The above figure is the example for shadow problem which shows that the object in the polygon surface which is reflected by the light source shows the shadow of the object in the other end. Shadow volume appears as a duplicate object which may leads to the following constraints

- Increases spatial gap between pixels.
- Miss-match surface retrieval.
- Occupies more space.
- Inconsistency.

2. CORRELATION AND SYNCHRONIZATION

The media correlation and media synchronization in a composite multimedia document is used to assist the multimedia for authoring, presenting and accessing. The three levels of media correlation in temporal, spatial and content domains are Syntactic level and Semantic level correlation .Syntactic level correlation means that the media streams are only based on objects and their locations for specific time stamps of the media stream. Semantic level correlation is the integration between objects and object locations for specific segments within the media stream. The Marking-based Synchronization Multimedia Tutoring [MSMT] system integrates voice and event streams to provide audiovisual lectures for English Composition studies. The MSMT System architecture consists of Marking-based Synchronization Multimedia Recording [MSMR], MSMT, Marking-based annotations Indexer, Marking-based Synchronization Multimedia Voice [MSMV], explains how to synchronize the voice and how to give a better presentation in audiovisual scenes. In Syntactic level correlation explains about the temporal, spatial and content correlation. In Semantic level correlation explains about Speech-Event Binding, Tele-pointer Movements Interpolation, Adaptable Handwriting Presentation and Erasing Handling. The two levels of media correlation are explored to simplify the authoring process, enhance the presentation and facilitate cross-media access. Three types of relations including temporal, spatial and content relationships between different media streams are investigated.

3. APPORACHES

There are various approaches to ensure synchronization of multimedia streams even after extraction from the semantic web but only few algorithms are widely used for synchronization, the following discussion tells how the procedure is carried out to ensure the synchronization.

3.1 Search Assumptions

The traditional approach to search is based on two assumptions: First, we should know exactly, what we are looking for; and second, we can organize the search space where we are looking to find it. These assumptions are not true in all searching situations. For instance, when humans try to recall information, in that situation our minds are not in one steady state or mood.

Elina Megalou and Thanasi's Hadzilacos gives the information about semantic abstractions in the Multimedia domain gives information about searching that gives exactly the matching content which is traditionally a valid point of machine level searching however, how human memory works and is rarely satisfactory for advanced retrieval tasks in any domain, multimedia in particular, where the presentational aspects can be equally important to the semantic content. A combined conception of the abstract and presentational characteristics of multimedia applications, leads to their conceptual structure and to the presentational structure. Conceptual structure includes with classic semantics of the real-world modeled by entities, relationships, and attributes with their presentational structure including media type, logical structure, temporal(time) synchronization, spatial (on the screen) synchronization and interactive behavior. Multimedia applications are construed as consisting of Presentational Units elementary and composite .The primal concept introduced is that of Semantic Multimedia Abstractions (SMA): qualitative abstract descriptions of multimedia applications interim of their conceptual and presentational properties at an adjustable level of abstraction.

3.2 Scene Segmentation

Scene segmentation is also known as story unit segmentation. In general, a scene is a group of contiguous shots that are coherent with a certain subject or theme. Scenes have higher level semantics than shots. Scenes are identified or segmented out by grouping successive shots with similar content into a meaningful semantic unit. The grouping may be based on information from texts, images, or the audio track in the video. According to shot representation, scene segmentation approaches can be classified into three categories: key framebased, audio and visual information integration-based, and background-based.

3.3 Key Frame-Based Approach:

This approach represents each video shot by a set of key frames from which features are extracted. Temporally close shots with similar features are grouped into a scene. Similar shots are linked, and scenes are segmented by connecting the overlapping links. A motion-based key frame selection strategy is, thus, used to compactly represent shot contents. Scene changes are detected by measuring the similarity of the key frames in the neighboring shots. The limitation of the key frame-based approach is that key frames cannot effectively represent the dynamic contents of shots, as shots within a scene are generally correlated by dynamic contents within the scene rather than by key frame-based similarities between shots.

3.4 Audio and Vision Integration-Based Approach:

This approach selects a shot boundary where the visual and audio contents change simultaneously as a scene boundary. A time-constrained nearest neighbor algorithm is used to determine the correspondences between these two sets of scenes. The limitation of the audio and visual integration based approach is that it is difficult to determine the relation between audio segments and visual shots.

3.5 Background-Based Approach:

This approach segments scenes under the assumption that shots belonging to the same scene often have similar backgrounds. Then, the color and texture distributions of all the background images in a shot are estimated to determine the shot similarity and the rules of filmmaking are used to guide the shot grouping process. The limitation of the background based approach is the assumption that shots in the same scene have similar backgrounds: sometimes the backgrounds in shots in a scene are different. According to the processing method, current scene segmentation approaches can be divided into four categories: merging based, splitting-based, statistical model-based, and shot boundary classification-based.

3.6 Merging-based approach:

This approach gradually merges similar shots to form a scene in a bottom-up style. In the first pass, over segmentation of scenes is carried out using backward shot coherence. In the second pass, the over segmented scenes are identified using motion analysis and then merged. The algorithm takes each shot as a hidden state and loops upon the boundaries between consecutive shots by a left-right HMM.

3.4.1 Splitting-based approach:

This approach splits the whole video into separate coherent scenes using a top-down style.

3.7 Statistical model-based approach:

This approach constructs statistical models of shots to segment scenes. The scene boundaries are updated by diffusing, merging, and splitting the scene boundaries estimated in the previous step. Each scene is modeled with a Gaussian density. A boundary voting procedure decides the optimal scene boundaries.

3.8 Shot boundary classification-based approach:

In this approach, features of shot boundaries are extracted and then used to classify shot boundaries into scene boundaries and non scene boundaries.

In their method, scene segmentation is based on a classification with the two classes of "scene change" and "non scene change." A SVM is used to classify the shot boundaries. Hand-labeled video scene boundaries from a variety of broadcast genres are used to generate positive and negative training samples for the SVM.

The common point in the merging-based, splitting-based, and statistical model-based approaches is that the similarities between different shots are used to combine similar shots into scenes. This is simple and intuitive. However, in these approaches, shots are usually represented by a set of selected key frames, which often fail to represent the dynamic contents of the shots. As a result, two shots are regarded as similar, if their key frames are in the same environment rather than if they are visually similar. The shot boundary classification based approach takes advantage of the local information about shot boundaries. This ensures that algorithms with low computational complexities are easy to obtain. However, lack of global information about shots inevitably reduces the accuracy of scene segmentation. It is noted that most current approaches for scene segmentation exploit the characteristics of specific video domains such as movies, TVs, and news broadcasts for example, using the production rules by which movies or TV shows are composed. The accuracy of scene segmentation is improved, but it is necessary to construct *a priori* model for each application.

3.6 Semantic Web

Semantic Web and semantic Web Services explains about how father and son or indivisible Twins? And this is researched by Martin Heep in Digital Enterprise Research Institute [DERI], University of Innsbruck. Semantic Web is used to gaining strength in industry and Education. The recent International Semantic Web Conference [ISWC] attracted more than 500 researchers; major vendors such as IBM, Oracle, and Software AG have been released. We can realize the Semantic Web by gradually augmenting existing data [mainly HTML and XHTML] via ontological annotations derived from today's human-readable Web content. The symbolic level of abstraction covers the raw multimedia information represented in well-known format for video, image, audio, text, metadata etc...

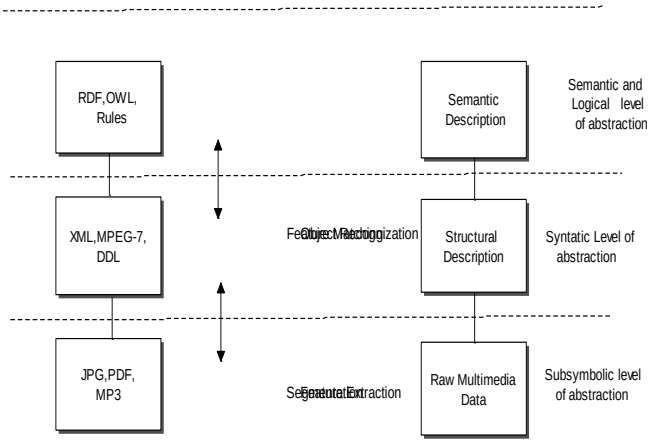

Figure 2 : Abstraction Levels of Multimedia Annotations

The above figure shows how semantic web provides the data for the presentation and how the information's are hided to the users in different levels. Only 7 percent of vendor operated sites offered room-availability information, which is the most important fact when searching for a suitable offer. Only 21 percent of the accommodations give availability data. The remaining 79 percent require a potential guest to either call or communicate via email to check availability. At least half the sites covered only 7 of 16 typically relevant categories in sufficient detail for decision-making. All of which are limitations that Semantic Web technologies promise to overcome. Entities are more willing to expose functionality than data in business settings**.** Work already exists on annotating dynamic Web content, but the fact that results to queries for availability and price aren't a functional value to this input isn't the same as whether a Web site is based on static HTML/XHTML pages or dynamic Web pages (PHP, active server pages, and so on) that are generated on the fly via a background database.

This is exactly what Semantic Web services frameworks, such as the Web Service Modeling Ontology (WSMO), OWL-S, or the Semantic Web Services Language (SWSL), offer. The SPARQL protocol, which will provide a standard query interface to Resource Description Framework (RDF) databases, can also be regarded as a simplistic framework for exposing functionality, albeit limited to database queries. Exposing functionality in the form of Web services is generally more attractive for market participants than publishing all relevant facts directly on the Web. I believe that the Semantic Web must include annotation of functionality through Semantic Web services technologies such as WSMO, SWSL, or OWL-S. I'm also convinced that it's possible to describe SPARQL endpoints using something like WSMO and thus embed this promising approach into a more generic Semantic Web services framework. As far as Semantic Web services are concerned, we should think about whether fully automated discovery , composition , and orchestration is a realistic scope, or whether more lightweight approaches are appropriate. Quite appealing is that both can likely fit well into a single comprehensive representational framework for exposing and finding functionality on the Web, such as WSMO.

4. CONCLUSION

A major feature that differentiates multimedia applications from other traditional applications is the integration of various media streams that have to be presented in a synchronized fashion. One of the main problems that have to be addressed in a multimedia system is the way media streams are synchronized when they are presented to the users. In order to guarantee high-quality multimedia presentations, real-time support from the network and operating system is important, for this we concentrate on synchronization algorithms that consider best-effort environments, for two reasons: first, algorithms designed for non-real-time environments can also work in real-time environments. First, we address the effect of workload variation on the synchronization specification of multimedia streams and on the display time of video frames. Second, we introduce a model for expressing the synchronization condition between continuous streams with different frame durations and apply it to fine-grain lip synchronization. Our lip synchronization algorithm based on the variation of the display time for a video frame. The video stream is the slave stream and it is synchronized with the audio stream. Audio stream plays continuously. Due to the transmission delays, the causal relation between two streams may be affected.

5. FUTURE DEVELOPMENTS

Although a large amount of work has been done in visual content-based indexing and retrieval, many issues are still open and deserve further research, especially in the following areas.

1) Motion Feature Analysis. The effective use of motion information is essential for content-based video retrieval. To distinguish between background motion and foreground motion, detect moving objects and events, combine static features and motion features, and construct motion-based indices are all important research areas.

2) Hierarchical Analysis of Video Contents. One video may contain different meanings at different semantic levels. Hierarchical organization of video concepts is required for semantic based video indexing and retrieval. Hierarchical analysis requires the decomposition of high-level semantic concepts into a series of low-level basic semantic concepts and their constraints. Low-level basic semantic concepts can be directly associated with low-level features, and high-level semantic concepts can be deduced from low-level basic semantic concepts by statistical analysis. In addition, building hierarchical semantic relations between scenes, shots, and key frames, on the basis of video structural analysis; establishing links between classifications with the three different levels: genres, event and object; and hierarchically organizing and visualizing retrieval results are all interesting research issues.

3) Hierarchical Video Indices. Corresponding to hierarchical video analysis, hierarchical video indices can be utilized in video indexing. The lowest layer in the hierarchy is the index store model corresponding to the high-dimensional feature index structure. The highest layer is the semantic index model describing the semantic concepts and their correlations in the videos to be retrieved. The middle layer is the index context model that links the semantic concept model and the store model. Dynamic, online, and adaptive updating of the hierarchical index model, handling of temporal sequence features of videos during index construction and updating, dynamic measure of video similarity based on statistic feature selection, and fast video search using hierarchical indices are all interesting research questions.

4) Fusion of Multi models. The semantic content of a video is usually an integrated expression of multiple models. Fusion of information from multiple models can be useful in content based video retrieval . Description of temporal relations between different kinds of information from multiple models, dynamic weighting of features of different models, fusion of information from multiple models that express the same theme, and fusion of multiple model information in multiple levels are all difficult issues in the fusion analysis of integrated models.

5) Semantic-Based Video Indexing and Retrieval. Current approaches for semantic-based video indexing and retrieval usually utilize a set of texts to describe the visual contents of videos. Although many automatic semantic concept detectors have been developed, there are many unanswered questions: How to select the features that are the most representative of semantic concepts? How should large-scale concept ontology for videos be constructed? How to choose useful generic concept detectors with high retrieval utility? How many useful concepts are needed? How can high-level concepts be automatically incorporated into video retrieval? How can ontology be constructed for translating the query into terms that a concept detector set can handle? How can inconsistent annotations resulting from different people's interpretations of the same visual data be reconciled? How can elaborate ontology be established between the detector lexica? How can multimodality fusion be used to detect concepts more accurately? How can different machine learning approaches be fused to obtain more accurate concepts?

6) Extensible Video Indexing. Most current video indexing approaches depend heavily on prior domain knowledge. This limits their extensibility to new domains. The elimination of the dependence on domain knowledge is a future research problem.

6. REFERENCES

[1] Ramazan Sava Aygijn. An Integrated Framework for Interactive Multimedia Presentations in Distributed Multimedia Systems. In ACM, pages 1-581, '01 Proceedings of the ninth ACM international conference on Multimedia October 5 2001.

[2] Kuo-Yu Liu and Herng-Yow Chen. Exploring Media Correlation and Synchronization for Navigated Hypermedia Documents. annual ***ACM*** symposium on User interface software ACM 1-59593-044-2/05/0011 November 2005.

[3] Martin Hepp. Semantic Web and Semantic Web Services,IEEE Internet Computing, April 2006.

[4] Elina Megalou and Thanasis Hadzilacos. Semantic Abstractions In the Multimedia Domain IEEE Transactions on Knowledge and Data Engineering, vol. 15, no. 1, January/February 2003.

[5] Emilia Stoica, Hussein Abdel-Wahab and Kurt Maly. Synchronization of Multimedia Streams in Distributed Environments.,IEEE, 0-8186-781949, 1997.

[6] Zhou Y., Murata T., 2001, Modeling and Analysis of Distributed Multimedia Synchronization by Extended Fuzzy-Timing Petri Nets, Transactions of the Society for Design and Process Science, Volume 5, Number 4, ISSN:1092-0617, pp. 23-37,December 2001.

[7] Y. Gao, W.-B. Wang, and J.-H. Yong, "A video summarization tool using two-level redundancy detection

for personal video recorders," *IEEE Trans. Consum. Electron.*, vol. 54, no. 2, pp. 521–526, May 2008.

[8] J. Ayars. *Synchronized Multimedia Integration Language*. W3CRecommendation,2001. http://www.w3.org/TR/2001/REC-smil20- 20010807.

[9] T. Berners-Lee, J. Hendler, and O. Lassila. The semantic web. *The Scientific American Journal*, 2001.

[10] N. Kodali, C. Farkas, and D. Wijesekera. Enforcing integrity in multimedia surveillance. In *IFIP 11.5Working Conference on Integrity and Internal Controlin Information Systems*, 2003

[11] E. Damiani, S. D. C. di Vimercati, S. Paraboschi, and P. Samarati. Securing XML documents. *Lecture Notes in Computer Science*, 1777:121–122, 2000.

[12] N. Kodali, C. Farkas, and D. Wijesekera. Enforcing semantic-aware security in multimedia surveillance. In *Journal of Data Semantics*, 2003.

[13] E. Damiani, S. D. C. di Vimercati, S. Paraboschi, and P. Samarati. A fine grained access control system for xml documents. *ACM Transactions on Information and System Security*, 5, 2002.

[14] E. Damiani and S. D. C. di Vimercati. Securing xml based multimedia content. In *18th IFIP International Information Security Conference*, 2003.

[15] E. Bertino, M. Braun, S. Castano, E. Ferrari, and M. Mesiti. Author-x: A java-based system for XMLdata protection. In *IFIP Workshop on DatabaseSecurity*, pages 15–26, 2000.

[16] J. Calic, D. Gibson, and N. Campbell, "Efficient layout of comic-like video summaries," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 7, pp. 931–936, Jul. 2007.

[17] P. Over, G. Awad, J. Fiscus, and A. F. Smeaton. (2010). "TRECVID 2009–Goals, tasks, data, evaluation mechanisms and metrics," [Online]. Available: http://www-nlpir.nist.gov/projects/ tvpubs/tv.pubs.org.html.

[18] Weiming hu, *senior member, ieee*, nianhua xie, li li, xianglin zeng, and stephen maybank" a survey on visual content-based video indexing and retrieval", ieee transactions on systems, man, and cybernetics—part c: applications and reviews, vol. 41, no. 6, november 2011.

[19] M. G. Christel, A. G. Hauptmann, W.-H. Lin, M.-Y. Chen, B. Maher,and R. V. Baron, "Exploring the utility of fast-forward surrogates for BBC rushes," in *Proc. 2nd ACM TREC Video Retrieval Eval. VideoSummarization Workshop*, 2008, pp. 35–39.

[20] W. Ren, S. Singh, M. Singh, and Y. S. Zhu, "State-of-the-art on spatiotemporal information-based video retrieval," *Pattern Recognit.*, vol. 42, no. 2, pp. 267–282, Feb. 2009.

[21] M. Wang, X. S. Hua, J. Tang, and R. Hong, "Beyond distance measurement: Constructing neighborhood similarity for video annotation," *IEEE Trans. Multimedia*, vol. 11, no. 3, pp. 465–476, Apr. 2009.

[22] M. Bertini, A. Del Bimbo, and C. Torniai, "Automatic video annotation using ontologies extended with visual information," in *Proc. ACM Int. Conf. Multimedia*, Singapore, Nov. 2005, pp. 395–398.

[23] A. G. Hauptmann, M. Christel, and R. Yan, "Video retrieval based on semantic concepts," *Proc. IEEE*, vol. 96, no. 4, pp. 602–622, Apr. 2008.

[24] J. Fan, H. Luo, and A. K. Elmagarmid, "Concept-oriented indexing of video databases: Towards semantic sensitive retrieval and browsing," *IEEE Trans. Image Process.*, vol. 13, no. 7, pp. 974–992,Jul. 2004.

[25] F. Pereira, A. Vetro, and T. Sikora, "Multimedia retrieval and delivery: Essential metadata challenges and standards," *Proc. IEEE*, vol. 96, no. 4, pp. 721–744, Apr. 2008.

[26] Y. Aytar, M. Shah, and J. B. Luo, "Utilizing semantic word similarity measures for video retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2008, pp. 1–8.