Volume 3, Issue 11, November 2013 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

A Survey of Audio Visual Retrieval by Content based Analysis

V. Shankar Ganesh M.Tech (MMT)*
Department of Computer Science and Engineering,
K.S.R. College of Engineering, India

G. Nagarajan, M.E
Department of Computer Science and Engineering,
K.S.R. College of Engineering, India

Abstract- This study survey is about the current trends/ methods in video retrieval using the content based analysis. The major topics covered in the study includes automatic and manual annotation video description in TRECVID campaign, video IR systems, Thesaurus enrichment approach, Audio visual segmentation, dynamic programming problem, retrieving techniques, shot retrieval, program retrieval, similarity of approaches and futurework. This work is done in an aim to guide upcoming researches in the techniques and approaches available for video retrieval.

Keywords: CBVR, Shot segmentation, TRECVID, GTAA, WordNet, Dynamic programming, Histogram, Evaluation methodology, Query definitions, Retrieval tasks, Keyframe.

I. INTRODUCTION

Content-based video retrieval (CBVR) [23] is the technique for audio-visual retrieval problem; it is useful for searching videos in large databases. Content based analysis gives video indexing (VI) for effective audio visual archiving and retrieving. It is a source for media professionals in different field to archive video from database for reusage. Content based analysis provides a solution for the inevitably tedious and incomplete video fragments archive. Fine grained manual and automatic annotation source are helpful for the users to retrieve exact data. Common initial steps for most content based video analysis techniques are to segment a video into elementary shots using video extractor. Based on the description and similarity between the shots result is obtained by query given by the user. Some query input methods are text, image and even video as query for audio visual archive. In images query, the histogram approaches and video query is by shot-by-shot detection. Textual query is the most common query used in the search system to find the relevant data based on the description given in query interface. In some approaches, automatic or manual queries are composed using query composer. Finally, all the data are stored in storage server as shown in Figure.1. The file system stores the video data, the indexes and metadata are stored in the storage server.

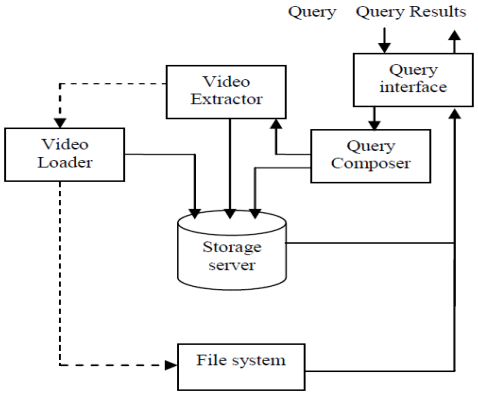

Figure.1. the system architecture

TRECVID [27] is one of the valuable instigators in the advancement of techniques for content based video retrieval, set the base for finding the advancement techniques to archive video data in real world queries. TRECVID started in the year 2001 for evaluation new methods in archiving video; in this study we are going to evaluate some of the optimal approaches in content based video retrieval (CBVR) systems. Due to advancement in technology, the video data are easily captured and stored, we can compress it, transmit and even render it on different platforms, being navigate it (i.e.,) browsing and searching it based in inherent or values of, us is what we are going to guide for new researchers.

II. APPROACHES IN CONTENT BASED VIDEO RETRIEVAL

For video retrieval the largest collaborative benchmarking activity for content based activities is the series of TRECVID [27] worships, in this each participating group have to submit the results of running several statements need for searching data. Thus each person participating for search tool testing has to find the relevant data within the limit; the efficient search engines are tested based on this trademark. Some of the approaches in content based retrieval for video data are:

A. Video IR systems

In 2007, TRECVID [15] illustrates the capabilities of content based video retrieval in three systems such as, K-space interactive video retrieval by Dublin City University, MediaMill fork browser [9] by University of Amsterdam and VisionGo by National University of Singapore and institute of computing technology. In K-space video retrieval [21], the video is processed every second and extracting low-level feature descriptors for each K-frame and then combined in user interface of the system.

In MediaMill fork browsers, interface is used for combining query by keyword, query by example, semantic concepts (using 572) concepts, time and by program in shape of fork. In VisionGo, the combination of text from ASR (Automatic Speech Recognition) [17] , High-level features [7] and motion. In this to maximize user's interaction, UI is used for fast keystroke actions. The shot image relevant to query are automatically refreshed to display next row of keyframes in the rank list of video.

In above three systems, the keyframes are used as visual result for the search result of users either by text or using the image input. The video data are filtered either by static or dynamic threads of manual or automatic annotation. In all above, VisionGo gives the interactive query by user and results of the search.

B. Thesaurus Enrichment

In thesaurus enrichment concept [15], to investigate the audio-visual documents which are indexed with an in-house thesaurus to retrieve by improving using external thesaurus. Thus indexing of the audio-visual content automatically using this method. GTAA is in-house thesaurus with limited structure for indexing and searching the collection of video.

WordNet [29] is external thesaurus by combining both; the metadata of video data is resulted approximately. GTAA is common thesaurus for audio-visual archive contains items such as location, person name, marker and subject of video. WordNet [5] is nouns, adjectives, verbs and adverbs. Thus new relation is categorized based on trivial and non-trivial combination of words.

In this, of each run, measuring the precision (Prec), recall (Rec) and the mean of precision and recall, F_1-measure:

$$\text{Prec} = \frac{|Retrieval \ \& Relevant|}{|Retrieved|} \quad \text{-----} \quad (1)$$

$$\text{Rec} = \frac{|Retrieval \ \& Relevant|}{|Relevant|} \quad (2)$$

$$F_1 = 2.\frac{Prec.Rec}{Prec+Rec} \quad \text{-----} \quad (3)$$

where |Retrieved| is the number of programs a run retrieved, and |Relevant| is the number of programs that is relevant for a query. -----

C. Audio-visual segmentation

In the audio-visual segmentation approach [28], videos are segmented into content classes that do not directly correspond to topics. The overall correct classification rate is 97.3% for station dependent recognition and 93.7% for station independent recognition. The following types are the six content classes,

1. Begin,
2. End,
3. Newscaster,
4. Report,
5. Interview
6. Weather Forecast.

In the content classes, there are four different edit effects can be used such as four classes are defined for the edit effects:

1. Cut (a hard cut)
2. Dissolve
3. Wipe
4. Window Change.

Each of these classes is modeled by using Hidden Markov Model (HMM) [26]. In each image represents a feature vector consisting of 12 video features. The features are center, velocity and variance of motion, a modified intensity of motion, a modified difference histogram and a feature which improves the detection of dissolve edit effects and then classify the scenes. The extensions have to be made which we implemented in two different ways:

- After scene classification [13], there is some rules for extracting the topic boundaries which is based on the typical structure of a topics.
- The features of video and audio segmentation are combined into the HMM structure. An modified video model is used which represents topic structures.

In the audio-visual segmentation, the audio is extracted from the video and are segmented with boundaries the segmentation algorithm uses BIC, in this window of n audio features $X_1,\ldots\ldots X_n$ and arbitrarily places boundary in position P, generated two segments $X_1,\ldots\ldots X_i$ and $X_{i+1},\ldots\ldots X_n$ by two different modes θ_{21} and θ_{22}.

$$\Delta_{BICi} \overset{!}{<} 0 \text{ with} \quad \text{-----} \quad (4)$$

$$\Delta_{BICl} = \frac{n}{2}\log|\sum_w| + \frac{i}{2}\log|\sum_f| + \frac{n-i}{2}\log|\sum_s| + \frac{1}{2}\lambda\left(d + \frac{d(d+1)}{2}\right)\log n. \quad \text{-----} \quad (5)$$

$\sum_w$ denotes the covariance matrix of all window feature vectors $X_1,\ldots\ldots X_n$. $\sum_f$ and $\sum_s$ are the covariance matrices of the features of the first and second segment respectively. 4 is the feature vector dimension.

$k = \frac{l+m}{2}$ with $\Delta_{BICl} = 0$, $\Delta_{BICm} = 0$, l<k<m (6)

Where k is the middle point of the boundary and it is considered instead of i.

Here, video segmentation is done by scene and video features. The audio segmentation [8] and video segmentation are combined using HMM [23] structure. The system results the audio-visual topic input using the BIC and HMMs by duration of the topic and length of proceeding scenes.

D. Content based Analysis

In content based analysis [2], there are three dominant content based video retrieval methods such as, transcript based search, low-level feature based search and detector based search. The first method, transcript based search uses automatic speech recognition and given textual query [22] to retrieve high quality recordings are retrieved. The second method, low-level feature based search visual information of image keyframes are matched to query image [7], which is done by similarity metrics between global image histograms. The third method, detector-based search utilizes shot based detection scores by human defined concepts to retrieve video and it is shown in Figure.2.

The evaluation methodology includes manual catalog annotation for aggregating different fields by free text, tags and technical metadata. In multimedia content analysis dominant search methods above mentioned are used. The query sets such as archive queries are concatenation of text queries in various sessions by users the query is defined. In lab queries, the query [7] is defined based on if the program video contains shots relevant to query, then entire video is considered relevant to query, video collections varies from year to year for finding relevant shots. In future queries, the analyzing search sessions and use them to formulate multimedia queries by using transaction logs. In simulation query by using simulation framework of [3] using logs archive users to generate simulated query.

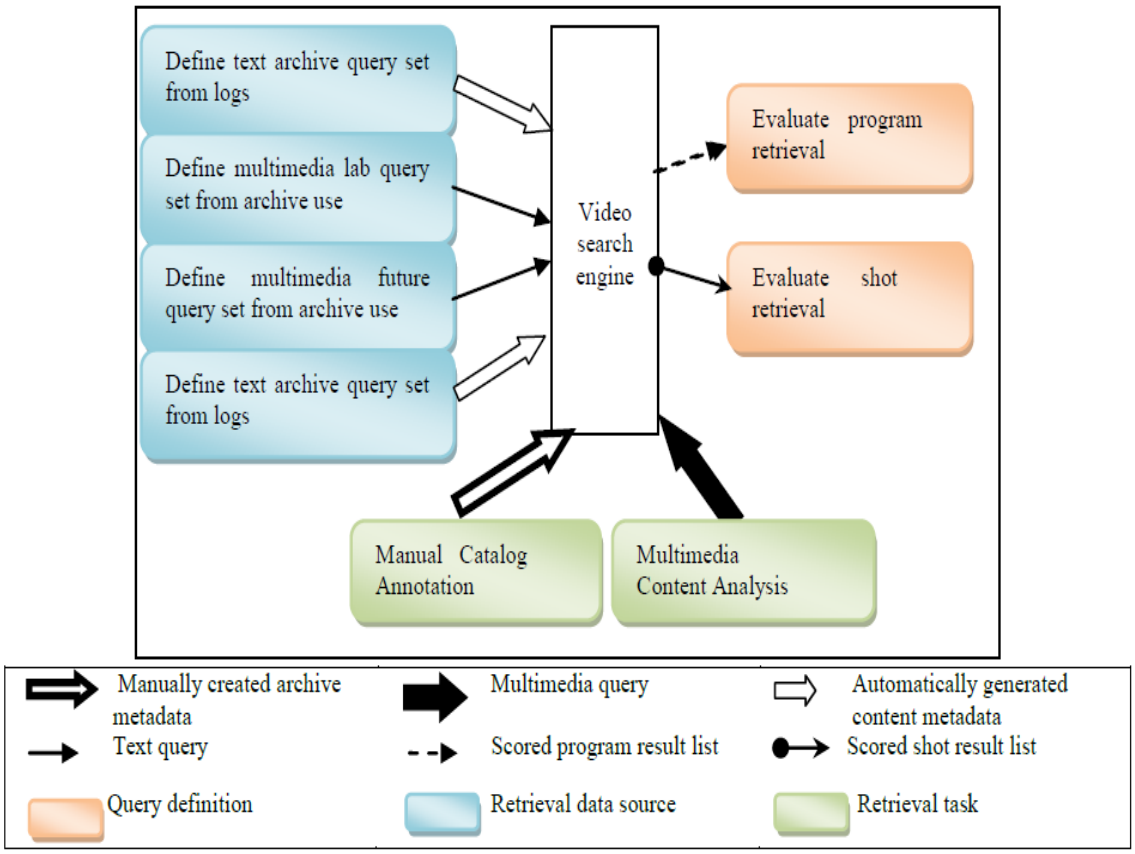

Figure 2: Evaluation methodology used to evaluate the potential impact of content-based video retrieval in the audiovisual archive.

In concept-based video retrieval [6], two retrieval tasks shot retrieval and program retrieval combining both annotation [17] phrase and multimedia contents to retrieve high quality data as result.

E. Multimodal Video Indexing and Retrieval

The framework for multimodal video indexing and retrieval has two techniques such as SODA and DI for different videos and indexing as shown in the Figure.3. The indexed videos can be retrieve by using SODA (Shrinkage optimized directed information assessment). The DI (Directed Information) [19] is used to capturing the video contents with respect to scenes [13], audio and so on. The captured information can be displayed and it is applied directly to the empirical probability distributions of both audio-visual features over successive frames. The focused features are RASTA-PLP features used to audio feature representation and SIFT features used to visual feature representation.

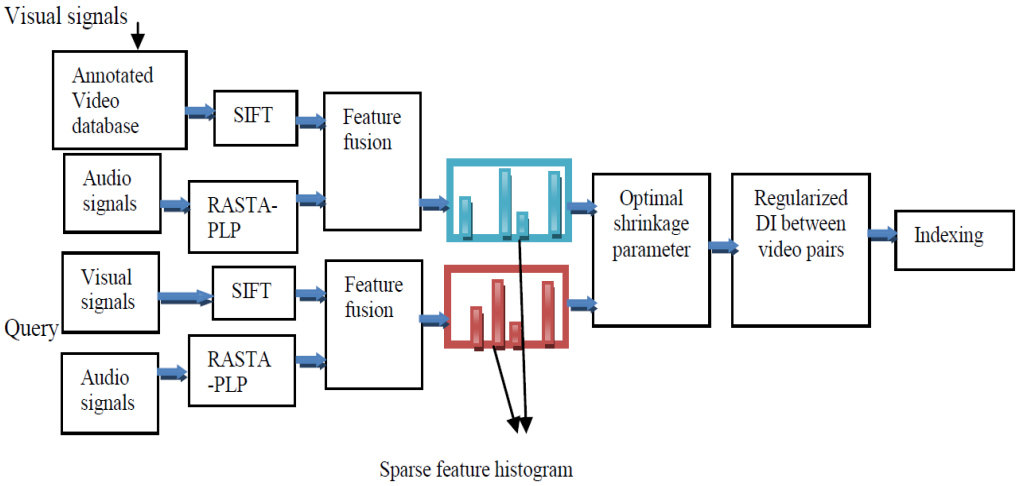

Figure.3. Block diagram of shrinkage optimized directed information (SODA) for fusion of audio and visual features for video indexing.

The authority of the SODA approach in video indexing, retrieval and activity recognition as compared to the state-of-the-art methods such as Hidden Markov Models (HMM) [4], Support Vector Machine (SVM) [4], Cross- Media Indexing Space (CMIS) and other non-causal divergence measures such as mutual information (MI). SODA used to the estimation of the joint probability distribution of audio and visual features and fuses the audio-visual signals. The Relative Spectra Transform- Perceptual Linear Prediction (RASTA-PLP) for speech feature representation due to their superiority in smoothing over short-term noise variations and SIFT features for visual feature representation, due to their invariance to image scale, rotation and other effects, and the bag of visual words (BOW) model for representing image content in each frame. The mutual information (MI) between V_x and V_y is

$$MI(V_x; V_y) = E\left[\ln \frac{f(X^{(M_x)}, Y^{(M_y)})}{f(X^{(M_x)}) f(Y^{(M_y)})}\right] \quad \text{-----} \quad (7)$$

Where,

$$f(X^{(M_x)}, Y^{(M_y)}) = f(X^{(M,a)}, X^{(M,v)}, Y^{(M,a)}, Y^{(M,v)}) \quad \text{-----} \quad (8)$$

Equation (8) is the joint distribution for fusion of the audio and video features for both the sequences V_x and V_y, and

$$f(X^{(M_x)}) = f(X^{(M,a)}, (X^{(M,v)}) \quad \text{-----} \quad (9)$$

and

$$f(X^{(M_x)}) = f(Y^{(M,a)}, Y^{(M,v)}) \quad \text{-----} \quad (10)$$

Equation (9) and (10) are joint distributions of audio-visual features for each sequence. Finally, in this approach multimodal indexing is done as shown in Figure.3.

III. CONCLUSION

In this paper, we probably focused on only a small part of all existing video retrieval systems. From this survey it is difficult to evaluate content based video retrieval systems based on effectiveness, efficiency and flexibility. However in the above researches, the goal was not only to guide for building an application but also to investigate the possibilities of retrieving video data in large databases. The study convinces us that we make a good basis for future improvements in the field of content based video analysis. In futurework, content based video searching will make good results by using effective techniques, from that content based video archive can be retrieved with better improvement.

REFERENCES

[1] "Alert homepage," http://alert.uni-duisburg.de/start.html.
[2] Alan F. Smeaton, Peter Wilkins, Marcel Worring, Ork de Rooij, Tat-Seng Chua, Huanbo Luan, "Content-Based Video Retrieval: Three Example Systems from TRECVid" May 2008
[3] Bouke Huurnink, Cees G. M. Snoek,Content-Based Analysis Improves Audiovisual Archive Retrieval" August 2012.
[4] C. Chang and C. Lin. Libsvm: A library for support vector machines. 2001.
[5] C. Fellbaum, editor. *WordNet: an electronic lexical database*, Cambridge, MA, USA, 1998. MIT press.
[6] C. G. M. Snoek and M. Worring, "Concept-based video retrieval," *Found. Trends Inf. Retriev.*, vol. 4, no. 2, pp. 215–322, 2009.
[7] C. Hauff, V. Murdock, and R. Baeza-Yates, "Improved query difficulty prediction for the web," in *Proc. CIKM*, 2008, pp. 439–448, ACM.
[8] C. Petersohn, "Fraunhofer HHI at TRECVID 2004: Shot boundary detection system," in *Proc. TRECVID*, Gaithersburg, MD, 2004.
[9] C.G.M. Snoek, J.C. van Gemert, Th. Gevers, B. Huurnink, D.C. Koelma, M. Van Liempt, O. De Rooij, K.E.A. van de Sande, F.J. Seinstra, A.W.M. Smeulders, A.H.C. Thean, C.J. Veenman, and M. Worring. The MediaMill

TRECVID 2006 Semantic Video Search Engine. In Proceedings of TRECVID 2006, Gaithersburg, MD, November 2006.

[10] E. Voorhees. Query expansion using lexical-semantic relations. In W. B. Croft and C. J. Rijsbergen, van, editors, *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 61–69, New York, NY, USA, 1994. Springer-Verlag.

[11] G. Schwarz, "Estimating the dimension of a model," *The Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[12] International Organization for Standardization. *ISO 2788:1986. Guidelines for the establishment and development of monolingual thesauri*. ISO, Geneva, 1986.

[13] J. Huang, Z. Liu, Y. Wang, Y. Chen, and E. K.Wong. Integration of multimodal features for video scene classification based on hmm. In IEEE Signal Processing Society 1998 Workshop on Multimedia Signal Processing, 1999.

[14] J. R. Smith and S.-F. Chang, "Visually searching the web for content," *IEEE MultiMedia*, vol. 4, no. 3, pp. 12–20, 1997.

[15] Laura Hollink, V´eronique Malais´e, and Guus Schreiber, " Enriching a Thesaurus to Improve Retrieval of Audiovisual Documents"

[16] M. Assem, van, V. Malais´e, A. Miles, and A. Th. Schreiber. A method to convert thesauri to skos. In *In Proceedings of the Third European Semantic Web Conference*, pages 95–109, Budvar, Montenegro, 2006.

[17] M. Huijbregts, R. Ordelman, and F. de Jong, "Annotation of heterogeneous multimedia content using automatic speech recognition," in *Proc. SAMT*, Berlin, Germany, 2007, LNCS, Springer Verlag.

[18] Mauro Cettolo and Marcello Federico, "Model selection criteria for acoustic segmentation," in *Proc. of the ISCA ITRW ASR2000 Automatic Speech Recognition*, Paris, France, 2000, pp. 221–227.

[19] P. Amblard and O. Michel. On directed information theory and granger causality graphs. In Journal of Computational Neuroscience, volume 30, 2011.

[20] P. Joly, J. Benois-Pineau, E. Kijak, and G. Que´not, The argos campaign: Evaluation of video analysis tools. In Proceedings of the International Workshop on Content-Based Multimedia Indexing, 2007. CBMI'07, 2007, pp. 130–137.

[21] P. Wilkins, T. Adamek, P. Ferguson, M. Hughes, G.J.F. Jones, G. Keenan, K. McGuinness, N.E. O'Connor, D. Sadlier, and A.F. Smeaton, K-Space at TRECVid 2007, In Proceedings of TRECVID 2007, Gaithersburg, MD, November 2007.

[22] R. Yan and A. Hauptmann, "A review of text and image retrieval approaches for broadcast news video," *Inf. Retriev.*, vol. 10, no. 4, pp. 445–484, 2007.

[23] S.-Y. Neo, J. Zhao,M.-Y. Kan, and T.-S. Chua, "Video retrieval using high level features: Exploiting query matching and confidence-based weighting," in *Proc. CIVR*, Heidelberg, Germany, 2006, pp. 143–152, Springer-Verlag.

[24] Stefan Eickeler and Stefan M¨uller, "Content-based video indexing of tv broadcast news using hidden markov models," in *Proc. IEEE ICASSP*, 1999, pp. 2997–3000.

[25] T. Adamek and N. O'Connor, Using Dempster-Shafer theory to fuse multiple information sources in region-based segmentation, In ICIP 2007 – Proceedings of the 14th IEEE International Conference on Image Processing, 2007.

[26] T. Hospedales, S. Gong, and T. Xiang. A Markov clustering topic model for mining behaviour in video. In ICCV, 2009.

[27] T.S. Chua, S. Neo, Y. Zheng, H. Goh, X. Zhang, S. Tang, Y. Zhang, J. Li, J. Cao, H. Luan, Q. He, and X. Zhang, TRECVID 2007 Search Tasks by NUS-ICT. In Proceedings of TRECVID 2007, Gaithersburg, MD., November 2007.

[28] *U. Iurgel, R. Meermeier, S. Eickeler, G. Rigoll, "*New Approaches to Audio-Visual Segmentation of TV News for Automatic Topic Retrieval"

[29] V. Malais´e, A. Isaac, L. Gazendam, and H. Brugmann. Anchoring dutch cultural heritage thesauri to wordnet: two case studies. In *ACL 2007 Workshop on Language Technology for Cultural Heritage Data*, 2007.

[30] Xu Chen, Alfred Hero and Silvio Savarese, "Multimodal Video Indexing and Retrieval Using Directed Information"

[31] Y. Ke, R. Sukthankar, and M. Hebert. Event Detection in Crowded Videos. In International Conference on Computer Vision (ICCV). IEEE, 2007.

