Volume 4, No. 12, December 2013
Journal of Global Research in Computer Science
ISSN-2229-371X
RESEARCH PAPER
Available Online at www.jgrcs.info

A FRAMEWORK FOR GROUP BASED IMAGE RETRIEVAL AND VIDEO ANNOTATION

Dr. V. Radha [*1] and K.Tamil Selvi[2]
[*1]Professor, [2]Research Scholar,
Department of Computer Science
Avinashilingam Institute for Home Science and Higher Education for Women. Coimbatore, India.
radhasrimail@gmail.com*, ktamilselvimsc23@gmail.com

Abstract: In this research two automatic video annotation techniques are considered. The first technique uses ontology to reduce the semantic gap during video retrieval and other performs a group based image retrieval using video files. The proposed algorithm uses GIR algorithm to create similar image group. From this refined set of images, SIFT features are extracted and the steps used by ASVA algorithm is performed to annotate the video in a semantic fashion. The Automatic Semantic based Video Annotation algorithm performs annotation in three steps. The first step calculates the video similarity using SIFT features, sentence and synonym analysis is performed in the second to find similar meaning annotations and finally the conjunction of the sentences are analyzed to increase the certainty of each annotation using Concept Net.

Keywords: Ontology, Video annotation, Group based image retrieval,wavelet feature, Earth Mover's distance.

INTRODUCTION

Due to the technological advancements both in hardware and software, the amount of image and video content used is increasing in alarming fashion. Various applications like video-on-demand, computer aided educational Compact discs apart from films, advertisement is using videos abundantly. In today's digital environment, videos are used in many applications, ranging from simple PowerPoint presentation to complex video-on-demand, entertainment to computer aided tutorials as teaching aid. Presently, World Wide Web (WWW) is rich with huge video databases. For example, as of August 2008, YouTube has more than 144 million videos [12] which has grown in multifold today. While the number of video databases present in the WWW is increasing, the ability of efficiently utilizing these video files is limited because of the existence of semantic gapwhich is the difference between the low-level visual features and the human's perception. One of effectively dealing the problem of semantic gap is to use tags to the digital data. These tags are termed as annotation, andwhen used to represent objects related to video files, are referred to as video annotation.

Video annotation is considered as a crucial task to improve the search process and provide fast access to video files in huge databases. Currently, search engines like Google, Yahoo and MSN use text based search which is not very effective with video files. Retrieving videos by key words requires semantic knowledge of the videos.The main motivation of automating this process is because the manual labeling of video data is not only labor intensive and time consuming but also is subject to human errors. With the growing use of videos, the importance of video annotation several researchers are contributing to the problem of automatic video annotation (Mu, 2010) in various aspects. Existing proposed solutions can be categorized into two groups, namely, image-feature based or ontology (semantic) based algorithms. Inspite of various proposed scheme, there is a lack of satisfactory techniques for extracting information from raw video and researchers seek alternative schemes for achieving the same goal of supporting indexing and query of content. This paper proposes an alternative enhanced method that combines both image feature based and semantic based algorithm to annotate videos.

[2] proposed a semantic based approach for video annotation with the aim to bridge the semantic gap. On receiving a new video input to be annotated, this framework uses a pre-annotated video dataset to identify similar videos. The matched annotations are then semantically analyzed and the best description for this new video is obtained using commonsense knowledge base. Commonsense is the term referred to identify information and facts that are expected to be known by ordinary people. Using these results, the new video is annotated. This system, referred to ASVA (Automatic Semantic Video Annotation) in this paper.

The ASVA algorithm compares all the dominant moving objects in the input video frame with that of objects in all frames of each video in the pre-annotated dataset using low level features called SIFT (Scale Invariant Feature Transformation). In the proposed method, a group based image retrieval method [21] is used first to identify videos with similar moving objects after which the semantic and concept base steps can be used to annotate the video. The technique proposed by Murabayashi *et al*. (2008)is referred as GIR system in this paper is used for this purpose. The proposed model is referred as GIRVA. For this purpose, wavelet features and k-means clustering are used. This change in ASVA can improve the precision and recall while reducing the large feature space produced while using the SIFT features. Reduction in feature space results in fewer computations and hence increases the speed of the algorithm.

The rest of the paper is organized as below. A brief literature study is provided in Section 2and a general video annotation system described in section 3 followed by a

detailed explanation of the steps in the proposed annotation algorithm in Section 4. The experimental results obtained while testing with TRECVID 2005 BBC Rushes dataset is presented and discussed in Section 5. The work is concluded with future research directions in Section 6.

LITERATURE STUDY

Video annotation has wide usage in many domains including education and media research. Several real time video annotation tools have been proposed. Examples include Marquee , Audio Notebook [27], Logjam [6], R frames [3], MSR Video Skimmer [17], Hierarchical Video Magnifier [20], Jabber [16], VoiceGraph [23, 26] Media Streams [7] and DIVA [19]. All these tools require the help of users to annotate and have been proved to work efficiently. However, as indicated before, manual annotation has serious drawbacks like labor intensive, time consuming and high human errors. Proposal that perform automatic annotation have also been probed. OVID (Object-oriented Video Information Database [24] automatically identify meaningful scenes in a video using inheritance based on interval inclusion relationship and a generalized hierarchical model along with adhoc query facility called VideoSQL to annotate and retrieve video. Similarly, [30] presented a video conceptual model designed to cater for all aspects of digital video management.VideoSTAR (Video Storage And Retrieval, [11] is a database system developed at the Norwegian Institute of Technology. It proposes a comprehensive conceptual model designed to handle media files, virtual video documents, video structure and content-based annotations; and parts of it have been implemented.

AVIS (Advanced Video Information System) is a video database approach focusing on query processing proposed by [1]. This is a formal model for a video database, as well as index structures and algorithms for queries and updates. Veggie [11] is an application for describing video with Dublin Core-based metadata, developed at the state library and university of Queensland, Australia. Its purpose is mainly to enable quick, easy, cost-effective generation of standardized metadata that can be used to create online detailed visual summaries of videos.In 1997, another video annotation engine called Vane was proposed by [4] was developed at Boston University. It is a tool for semi-automatic production of metadata and is designed to be as open as possible for multiple domain-specific applications. Later in 2003, a commercial product called Qualitative Media Analyzer (QMA) was developed for annotating videos of interviews and media documents by creating scores and independent variables. Observer [22] a competitor for QMA, is another commercial product developed by Noldus Information Technology.Bil Video [8] is a video database system developed at Bilkent University in Ankara, Turkey. Its main contribution is the advanced, rule-based spatio-temporal modelling and querying functionality along with more conventional temporal semantic annotations. The VideoText model [15] is a video data model based on the concepts of logical video segment and free text video annotations with arbitrary mapping between them.

All of the above mentioned models, on the semantic expressivescale, end up on the weak side of structured data values. Recent solutions focus on proposing solutions to strengthen this weakness by including ontology, machine learning and genetic algorithms. Examples include [14], [31], and [5] In continuation with these researches, this paper uses semantics with image features to annotate video files.

A GENERAL VIDEO ANNOTATION SYATEM

A general video annotation system is presented in Figure 1 and consists ofthree major steps, namely, segmentation, semantics annotator and descriptors. The first step, video segmentation, cuts a video sequence into smaller units. These smaller video units are semantically analyzed to regulate the video content descriptions and to assign relevance scores that reflect the segments importance with respect to these descriptions. Finally, these descriptions are combined with input video to output the resultant annotated video. Thus, the main aim of any video annotation algorithm is to categorize the semantic content of each video unit, assign the corresponding relevance score and output the description file.

METHODOLOGY

In this paper, the proposed video annotation method is performed through a series of steps, namely, video segmentation, similarity calculation, analysis and annotation. Different algorithms exist for each of these steps. These algorithms are explained in this section and illustrated in Figure 2. The GIRVA (Group based Image Retrieval and Video Annotation) system addresses two main issues. The first is the selection of visual features that can efficiently give knowledge about the video content and second is the technique that can be used to represent these features in annotation format. These two details are addressed in three steps.

Video Similarity Calculation:

The similarity between video content is performed in four steps. The first step performs motion segmentation using Daubechies 4 (D4) wavelet transformation [29] is used as they are more robust to degraded videos than SIFT. The video frames are initially are divided into 4 x 4 blocks. Then D4 transformation is applied which results in four subbands, LL, HL, LH and HH. To use the magnitude of spatial frequency resolution, the absolute value of the coefficients is used and a feature vector as given below is created.

$$v_{ij} = (|D_{LH_{ij}}|, |D_{HL_{ij}}|, |D_{HH_{ij}}|) \quad (1)$$

where $D_{LH_{ij}}, D_{HL_{ij}}, D_{HH_{ij}}$ are the subband components, i and j are the coordinates of the coefficients in each wavelet transformed data. Thus v_{ij} is a three di*mensiona*l vector, which is clustered using k-means algorithm. The center values and ratio of the cluster size is calculated and used as feature vector (F_v) of block m (Equation 2).

$$F_v = \{(f_1, wf_1), \ldots,, (f_n, wf_n), \ldots, (f_n, wf_n)\} \quad (2)$$

where f_n is the centre of cluster n (Equation 3) and w_n is the ratio of the cluster size (Equation 4).

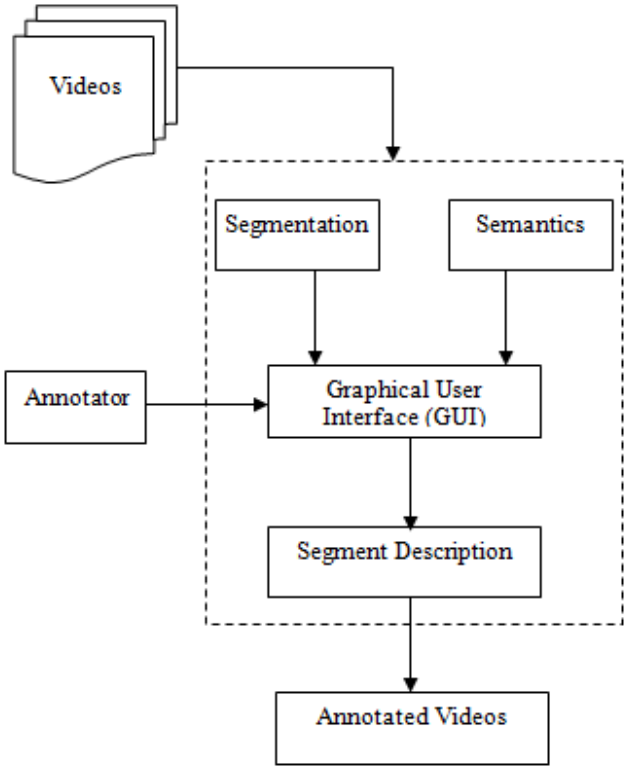

Figure: 1 Video Annotation System

$$\beta_n = \left(\frac{\sum_{v_{ij} \in cluster\ n} |d_{LH_{ij}}|}{\text{No. of } v_{ij} \in cluster\ n}, \frac{\sum_{v_{ij} \in cluster\ n} |d_{HL_{ij}}|}{\text{No. of } v_{ij} \in cluster\ n}, \frac{\sum_{v_{ij} \in cluster\ n} |d_{HH_{ij}}|}{\text{No. of } v_{ij} \in cluster\ n} \right) \quad (3)$$

$$w_n = \frac{\text{No. of } v_{ij} \in \text{cluster } n}{\text{No. of all } v_{ij}} \quad (4)$$

In the next step, the distance between the feature spaces are calculated using EMD method [25] between the same blocks. The output after applying EMD is a weighted list of text entries that accompanied to the top similar matched videos.

Sentence Analysis:

This step aims is to find similar meaning annotations irrespective of different names of the same or similar objects, method used to describe an event or action, different spelling versions. The step divides a sentence into Object, Event and Location triplet using Stanford NLP Log-linear Part-Of-Speech Tagger (POS Tagger). A POS Tagger is a software tool that reads text and assigns parts of speech to each word, such as noun, verb and adjective. These tags indicate which part is the object, which is the subject in linguistic terminology and which is the event, that is, the verb along with its related prepositions and the location, if exists. Three separated lists are generated from this analysis. The Object and Location lists are considered as list of nouns when entered to WordNet [9] and the Events list is considered as a list of verbs. The "isA" relationship in WordNet, which gives the synonyms, is selected because it gives equal meaning words with little amount of abstraction. Each list, separately, is extended then intersected using this relationship. The process is simply done by obtaining each item's synonyms, which match its part of speech (i.e. nouns for items in nouns' and locations' lists, and verbs for events' list). This assigns a suitable weight $S_w$for each synonym, calculated based on the initial word weight $W_w$and an un-trust decreasing constant C_d and can be formulated using Equation (5).

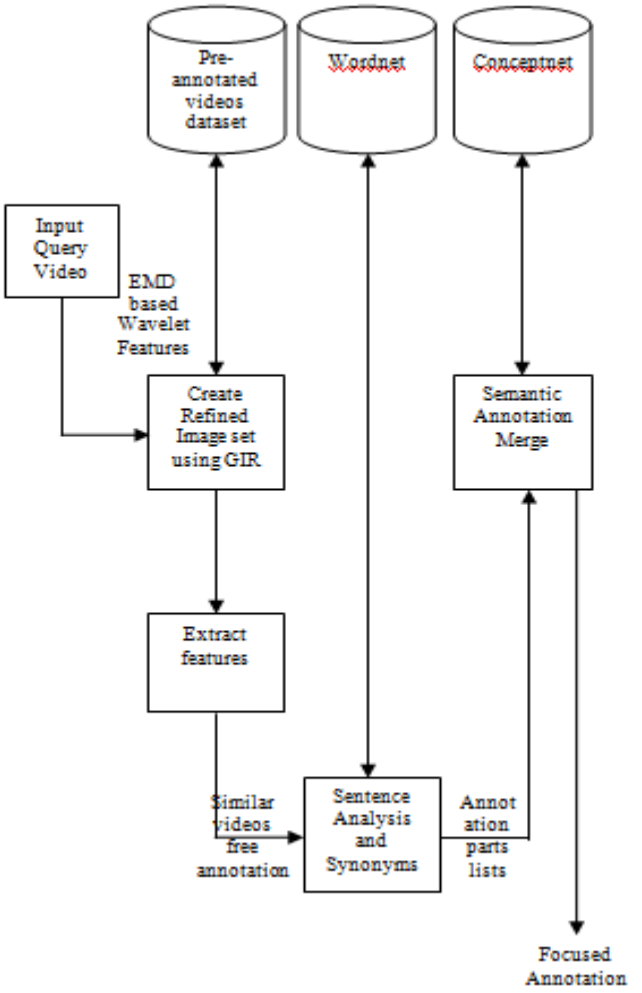

Figure 2: Proposed GIRVA System

$$S_w = W_w \times C_d \quad (5)$$

The decreasing constant $C_d$holds a value between 0 and 1, giving less weight for synonyms than the original word. A high value of C_d(more than 0.8), indicates similar strength of the synonym and the original word. This leads to increased false alarms and therefore, an average value of 0.5 was chosen. Matched words are grouped to increase their trust and the resulted lists are normalized. The output of this step is three sorted lists, each of which contains weighted entries for one part of the scene elements (object, event and location).

Semantic annotation:

The final step of ASVA algorithm checks the possible conjunctions of the sentences' parts under real constraints so as to assign more certainty to higher potential actions in daily life. For this step, ConceptNet [18]is used after some adaptation. ConceptNet consisted of a huge number of concept nodes; each concept is a semi-sentence or a phrase. Again, WordNet "isA" relationship is utilized and a full intersection operation is applied between objects' list and events' list using this relationship. Cross weights are calculated using Equation 6. Then the same operation between objects' list and locations' list is repeated using "locationAt" relation.

$T_w = N_w x\ V_w x\ R_s$ (6)

Where $T_w$is the sentence weight, $N_w$and $V_w$are the noun and verb phrases weights respectively and Ris the relation score. Each ConceptNet node is analyzed to obtain the core phrase that matches its type. Finally, the rest of the node is deleted if it does not hold a full meaning or another node suits this meaning is created. This task is performed using the steps given below:

a. Each node's words are tagged using Stanford previously mentioned tagger [10] then non-useful parts of sentence in visual field are deleted. These parts vary from some prepositions and stop words to some common used adjectives and adverbs, which are included in a manual written table. For example, "fast" is visually a useful adjective because it holds a meaning related to motion, but "better" is not.
b. A split operation is applied to divide some complex nodes into parts causing new relationships to be established

To achieve more effective comparing between analyzed nodes and resulted candidate annotations' parts, this comparing operation is performed on the stemming level. This is done by stemming all the words of each entry, i.e. obtain the root of the word, then sorting the resulted stemmed words alphabetically. This causes the nodes that contain the same words but in different format to be comparable.

EXPERIMENTAL RESULTS

To evaluate the systemsvideos from TRECVID 2005 BBC Rushes [28] is used. TRECVID 2005 is a group of standard databases for information retrieval. This dataset contains 335 single-shot video clips containing various types of moving vehicles like cars, tanks, airplanes and boats. These challenging uncontrolled videos contain considerable range of variations like size, appearance and shapes, viewpoint and motion of object. Also all possibilities of unknown camera quality and motion, like moving and zooming, are exists. The framework currently operates on individual video shots, but it can easily be extended by plugging a shot boundary detection layer. Some example frames from the dataset is shown in Figure 3.

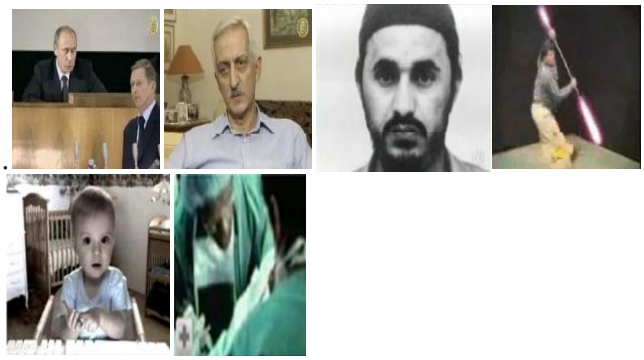

Figure 3 : Sample Video Frames

To ascertain the performance of the models, several experiments were conducted. All the experiments were conducted using a Pentium IV machine with 4GB RAM. Performance evaluation was done vigorously for both the existing and proposed systems. Four performance metrics, namely, average Precision, average Recall, average F-measure and speed were selected during evaluation.The precision, recall and F Measure were calculated using Equations 7, 8 and 9.

$$\text{Precision} = \frac{\text{Similar\_videos} \cap \text{Retrieved\_videos}}{\text{Retrieved\_videos}} \quad (7)$$

$$\text{Recall} = \frac{\text{Similar\_videos} \cap \text{Retrieved\_videos}}{\text{Similar\_videos}} \quad (8)$$

$$F = 2 \text{ x (Precision x Recall) / (Precision + Recall)} \quad (9)$$

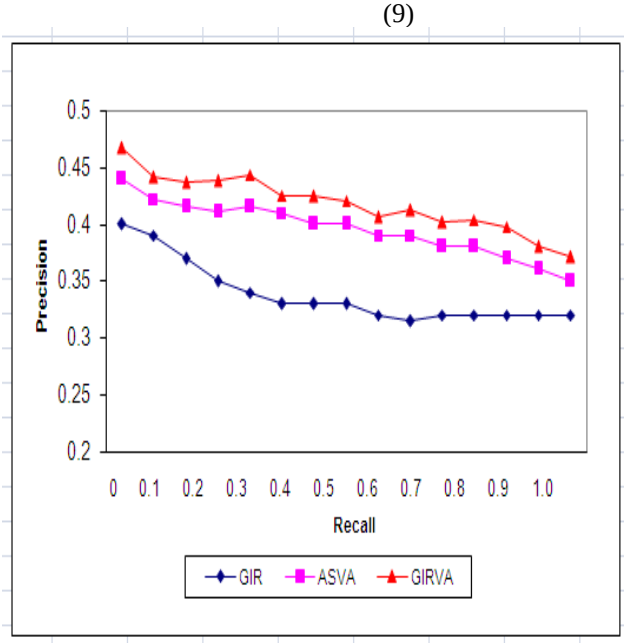

Figure 4 : Recall Vs. Precision

Figure 4 shows ROC diagram using the average of Recall vs. Precision while using ASVA and proposed GIRVA to annotate all database videos. Each time, one video is taken as a test and all other files in the database are considered as the pre-annotated database. Similarity has been evaluated by comparing correct similar retrieved files to all similar files and enhancement has been evaluated by comparing correctly retrieved annotations to all possible correct annotations for the input video. From the results, it is evident that the proposed method has enhanced the annotation process in terms of precision and recall.

To analyze the effectiveness of the pre-annotated dataset, F-measure parameter was used and the result of F-measure when used with different number of top ranked files is presented in Figure 5 from the results it can be seen that the proposed GIRVA algorithm performs better than ASVA algorithm. Similarity results show that the performance degrades when the number of files is more than 20. This improves to 30 for ASVA and 33 for GIRVA. After this number all the three algorithms performs in a similar manner. As the aim of the framework is not to retrieve the whole corrected list of annotations, but to find few representative annotations for the input video, the proposed algorithm achieves improved.

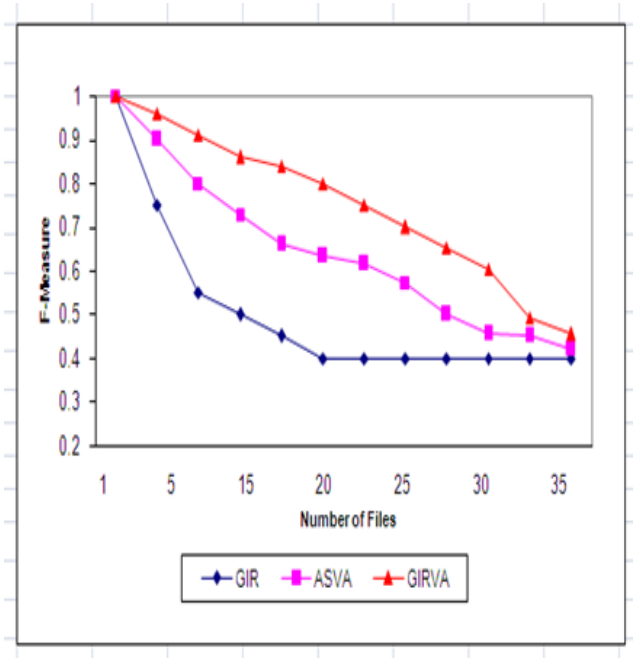

Figure 5Effectiveness of the Pre-annotated Dataset

While considering the speed of the three algorithms. When presented with a video file on average the GIR algorithm took less than 5.91 seconds, the ASVA algorithm took less than 6.75 seconds and the GIRVA algorithm annotated the video file within 8.21 seconds. The extra time taken by the proposed algorithm is due to the combined computations required to perform some steps of GIR and ASVA. But as the difference is very small when compared with GIR and ASVA algorithms (less than 2.3 seconds and 1.4 seconds respectively), the GIRVA algorithm can be considered as an enhanced version.

From the results, it could be seen that irrespective of the number of files the performance of GIRVA algorithm is superior to the existing annotation models namely, GIR and ASVA in terms of precision, recall and speed. The experiments further demonstrate that the proposed annotation framework can produce a small list of candidate annotation when compared with the existing frameworks.

CONCLUSION

This paper proposed a wavelet-based semantic system for video annotation. The usage of wavelets reduced the false alarms raised by illumination.The experimental results proved that annotation is effective. The results further show that the performance of the proposed algorithm has improved the annotation process in terms of all parameters except for speed. The speed of the proposed algorithm increased on average by 1.8 seconds due to the extract computations required while combining the algorithms. However, the high accuracy, low error rates show that the proposed algorithm is efficient and can be used to annotate large video databases. In future, plans to combinewavelets with SIFT is envisaged.

REFERENCES

[1] Adali, S., Candan, K.S. , Chen, S.S., Erol, K. and Subrahmanian, V.S. (1996) The Advanced Video Information System: Data Structures and Query Processing, Multimedia Systems, Vol.4, No.4, Pp.172–186.

[2] Altadmri, A. and Ahmed, A. (2009) Automatic Semantic Video Annotation in Wide Domain Videos Based on Similarity and Commonsense Knowledge bases, IEEE International Conference on Signal and Image Processing Applications, UK, Pp. 74 – 79.

[3] Arman, F., Depommier, R., Hsu, A. and Chiu, M. (1994) Content-Based Browsing of Video Sequences, ACM Multimedia, Pp.97-103

[4] Carrer, M., Ligresti, L., Ahanger, G. and Little, T.D.C. (1997) An Annotation Engine for Supporting Video Database Population, Journal of Multimedia Tools and Applications, ACM Digital Library, Vol. 5, No. 3, pp. 233-258.

[5] Chiu, C.Y., Lin, P.C. Li, S.Y., Tsai, T.H. Tsai, Y.L. (2012) Tagging Webcast Text in Baseball Videos by Video Segmentation and Text Alignment, IEEE Transactions on Circuits and Systems for Video Technology, Vol. 22 , Issue: 7, Pp. 999 – 1013.

[6] Cohen, J., Withgott, M. and Piernot, P. (1999) Logjam: a tangible multi-person interface for video logging, CHI 99 conference on human factors in computing systems, ACM Press, Pittsburgh, Pennsylvania.

[7] Davis, M. (2003) Media Streams: An Iconic Visual Language for Video Annotation," Proceedings of the 1993 IEEE Workshop on Visual Languages, Bergen, Norway, Pp.196-202.

[8] Dönderler, M.E., Saykol, E., Ulusoy, Ö. And Güdükbay, U. (2003) BilVideo: A Video Database Management System, IEEE MultiMedia, Vol. 10, No.1, Pp.66-70.

[9] Fellbaum, C. (1998) WordNet: an electronic lexical database. Cambridge, Mass: MIT Press.

[10] Group, S.N. (2009) Thestanfordnlp log-linear part of speech tagger, http://nlp.stanford.edu/software/tagger.shtml, Access Date 21-07-2012.

[11] Hjelsvold, R., Langørgen, S., Midtstraum, R. and Sandstå, O. (1995) Integrated Video Archive Tools, ACM Multimedia, Pp.283-293.

[12] http://www.youtube.com

[13] Hunter, J. and Newmarch, J. (1999) An Indexing, Browsing, Search and Retrieval System for Audiovisual Libraries, S. Abiteboul and A. Vercoustre (Ed.), Research and Advanced Technology for Digital Libraries (ECDL), Pp.76-91.

[14] Jeong, J.W., Hong, H.K. and Lee, D.H. (2011) Ontology-based automatic video annotation technique in smart TV environment, IEEE Transactions on Consumer Electronics, Vol. 57 , Issue: 4, Pp. 1830 – 1836

[15] Jiang, H., Montesi, D. and Elmagarmid, A.K. (1997) VideoText database systems, Proceedings of the 4th IEEE International Conference on Multimedia Computing and Systems, Pp. 334—351.

[16] Kominek, J. and Kazman, R. (2007) Accessing Multimedia through Concept Clustering, CHI, Pp.19-26.

[17] Li, F. C., Gupta, A., Sanocki, E., He, L. and Rui, Y. (2000) Browsing digital video, CHI, Pp.169-176.

[18] Liu, H. and Singh, P. (2004) Conceptnet a practical commonsense reasoning tool-kit, BT Technology Journal, Vol. 22, No. 4, Pp. 211–226.

[19] Mackay, W. E. and Beaudouin-Lafon, M. (2008) DIVA: Exploratory Data Analysis with Multimedia Streams, CHI, Pp.416-423.

[20] Mills, M., Cohen, J. and Wong, Y. Y. (2002) A Magnifier Tool for Video Data, CHI, Pp.93-98.

[21] Murabayashi, N., Kurahashi, S. and Yoshida, K. (2008) Group-based Image Retrieval Method for Video Annotation, International Symposium on Applications and the Internet, Tokyo, Pp. 126 – 132.

[22] Noldus Information Technology (2003) The Observer 5.0, www.noldus.com/ products/observer/index.html, Last Access Date : 22-07-2012.

[23] Oard, D. W. (2007) Speech-based Information Retrieval for Digital Libraries, Notes from AAAI Spring Symposium on Cross-Language Text and Speech Retrieval, Stanford University, California.

[24] Oomoto, E. and Tanaka, K. (1993) Ovid: Design and implementation of a video-object database system, TKDE, Vol.5, No.4, Pp.629–643.

[25] Rubner, Y., Tomasi, C. and Guibas, L.J. (2000) The earth mover's distance as a metric for image retrieval, International Journal of Computer Vision, Vol. 40, No. 2, Pp. 99–121.Saleh, A., Rahman, M., Cha, J. and Saddik, A.E. (2009) Authoring Edutainment Content through Video Annotations and 3D Model Augmentation, International Conference on Virtual Environments, Human-Computer Interfaces and Measurements Systems, China, IEEE Explore, Pp. 1-5.

[26] Slaughter, L. A., Oard, D. W., Warnick, V. L., Harding, J. L. and Wilkerson, G. J. (2008) A Graphical Interface for Speech-Based Retrieval, Proceedings of the 3rd ACM International Conferenceon Digital Libraries, Pp.305-306.

[27] Stifelman, L., Arons, B. and Schmandt, C., (2011) The audio notebook: paper and pen interaction with structured speech, CHI, 2011, Pp.182-189.

[28] Trecvideo Retrieval Track (2005) http://wwwnlpir.nist.gov/projects/trecvid, Last Access Date 21-07-2012.

[29] Wang, J.Z., Wiederhold, G., Firschein, O. and Wei, S.X. (1997) Content-based image indexing and searching using daubechies wavelet, International Journal on Digital Libraries, Vol. 1, No. 4, Pp. 311-328.

[30] Weiss, R., Duda, A. and Gifford, D.K. (1995) Composition and Search with a Video Algebra. IEEE MultiMedia, Vol. 2, No.1, Pp.12–25.

[31] Zhang, T. (2012) A Generic Framework for Video Annotation via Semi-Supervised Learning, IEEE Transactions on Multimedia, Vol, 14 , Issue: 4, Pp. 1206 – 1219.

