IJCSE International Journal of Computer Sciences and Engineering Open Access
Survey Paper Vol.-7, Issue-4, April 2019 E-ISSN: 2347-2693

Content Based Video Retrieval System Using Video Indexing

Jaimon Jacob[1*], Sudeep Ilayidom[2], V.P. Devassia[3]

[1]Govt. Model Engineering College, Ernakulam
[2]Division of Computer Engg, School of Engineering, Cochin University of Science and Technology
[3]Former Principal, Govt. Model Engineering College, Ernakulam

Corresponding Author: jaimon@mec.ac.in, Tel.: +91-0484-2301430

DOI: https://doi.org/10.26438/ijcse/v7i4.478482 | Available online at: www.ijcseonline.org

Accepted: 08/Apr/2019, Published: 30/Apr/2019

Abstract- Searching for a Video in World Wide Web has augmented expeditiously as there's been an explosion of growth in video on social media channels and networks in recent years. At present video search engines use the title, description, and thumbnail of the video for identifying the right one. In this paper, a novel video searching methodology is proposed using the Video indexing method. Video indexing is a technique of preparing an index, based on the content of video for the easy access of frames of interest. Videos are stored along with an index which is created out of video indexing technique. The video searching methodology check the content of index attached with each video to ensure that video is matching with the searching keyword and its relevance ensured, based on the word count of searching keyword in video index. Video captions are generated by the deep learning network model by combining global local (glocal) attention and context cascading mechanisms using VIST-Visual Story Telling dataset. Video Index generator uses Wormhole algorithm, that ensure minimum worst-case time for searching a key with a length of L Also, Video searching methodology extracts the video clip where the frames of interest lies from the original huge sized source video. Hence, searcher can get and download a video clip instead of downloading entire video from the video storage. This reduces the bandwidth requirement and time taken to download the videos.

Keywords- Video Indexing, Video Searching methodology, VIST- Visual Story Telling dataset

I. INTRODUCTION

Video is one of the main active medium to convey messages effectively. Consequent to the development in computer networking technology, video is now accessible to everyone through various social media networking platforms. Video is more effective than text or audio messages because it grabs the attention of people very easily, engage the targeted viewers and it is easy to memorize. In addition, video can embrace all the other forms of information including text, audio, music, photographs, links etc.

Prediction based on current CISCO Visual Networking Index (VNI), shows that global IP traffic would increase by three times in 2022 compared to that in 2017[1]. According to the forecast, the IP video which include Internet video, IP VoD, Video file exchanged through file sharing, video-streamed gaming and conferencing, will continue to be in the range of 80 to 90 percent of total traffic. Globally, IP video traffic will account for 82 percent of traffic by 2022. This remind the significance of managing the Video traffic by reducing the downloads only to the intended specific part of video.

Video indexing formulate a technique of indexing a video as in the text books. When a download request with a search keyword is raised, the relevant part of video is found by checking the word density using the video index and transfer that video clip only. This avoid the transfer of entire video and thus reduces the video traffic drastically.

This paper is organised as Introduction in section I, describes an overview of work proposed in this paper. In section II, various works already done related with this area described. It also ensured that no similar works has been done as presented in this paper. The proposed work described in detail with the support of block diagram in Section III. Results of the implementation of this proposed work described in section IV. Section V describes the conclusion and future scope of this proposed system.

II. RELATED WORK

Works on video searching based on video content is not seen reported in the literature so far. Most of the proposed algorithms are based either on textual content or audio content.

An Efficient Video Similarity Search Algorithm is introduced in [2] for the convenience of content-based video retrieval in large storage devices. Spatial-temporal distribution of video frame sequences is used as the base for similarity measurement and searching. The video similarity is restrained based on the computation of the number of similar video components.

A fast search algorithm for large video database using Histogram of Oriented Gradients [HOG] based features is proposed in [3]. HOG based features is used as a feature vector of a frame image in this paper. For achieving a fast and robust video search, it is proposed to combine this with active search algorithm.

Formerly, metadata is used for tagging the videos and clustering or captions are used for repossessing the videos[4]. Many algorithms proposed earlier used two stage pipeline to identify the semantic content (subject, Verb, Object) and generate a meaningful sentence using a predefined prototype[5]. Here, trained classifiers are used to recognise objects, actions and sights.

The method proposed in [6,7] is generating a fixed dimension vector representation of images by extracting features using a CNN. This fixed dimension vector representation is decoded into a sequence of words describing the content of image. LSTM models are used as sequence decoders, which is appropriate for learning long range dependencies, and overcome the inferior performance of RNN, if used as sequence decoders.

In [8], LSTMs is used to create video explanations by combining the images of individual frames. This technique excerpts the CNN features from frames in the video and then mean-pools the outcomes to get a single feature vector expressing the entire video. They also use an LSTM as a sequence decoder to create the explanation from this vector. A major inadequacy of this methodology is that this illustration completely disregards the gathering of the video frames and fails to accommodate any time-based information.

In [9] video classification is done using action, scene and object concepts as semantic characteristics. Also it is suggested to model events using a variety of harmonizing semantic characteristic features developed in a semantic concept space. It has systematically exploited the merits of this concept-based event representation (CBER) in applications of video event cataloguing and understanding.
Semantic signature representation reported in [10] is implemented for web video query for complex events using handful of video query examples, in which off-the-shelf concept detectors are used to capture the variance in semantic presence of events.

In order to overcome the complex dynamics in the Real-world videos, [11] propose a novel end-to-end sequence-to-sequence model which create captions for videos. Recurrent neural networks, specifically LSTMs, have been used for state of-the-art performance in image caption generation. In this work, LSTM model is trained on video-sentence pairs and to associate a sequence of video frames to a sequence of words in order to generate a description of the event in the video clip. This model is capable to learn the time-based structure of the frames in sequence as well as the sequence model of the resulting sentences.

A detailed survey on various indexing techniques is described in [12] which include video indexing, content based video indexing, video retrieval, and multimodal video indexing, key frame extraction and various indexing techniques etc.

In [13], an algorithm for content-based video indexing and retrieval is proposed using key-frames texture, edge, and motion features. The algorithm abstracts key frames from a video using k-means clustering based method, followed by abstraction of texture, edge, and motion features to characterize a video with the feature vector.

Automatic video description generation has recently been getting awareness after swift development in image caption generation. Automatically generating description for a video is more inspiring than an image due to its temporal dynamics of frames. Most of the algorithms relied on Recurrent Neural Network (RNN). Recently attentional mechanisms have also been applied to make the model learn to focus on some frames of the video while generating each word in a describing sentence. In [14] a novel algorithm on a sequence-to-sequence method with temporal attention mechanism is presented.

III. PROPOSED WORK

All video searching algorithms reported in contemporary research are accessing the whole video in a stretch while the stake holder is interested in a small portion of the video. However, the proposed methodology enables content wise retrieval of the video frames of interest only, considerably reducing the bandwidth.

In this methodology, an innovative idea is proposed to speed up the video data retrieval with the help of a Video Index. Video Index is similar to page index that appear in text books, where each significant keyword is arranged in sorted order along with the dense and frame.

Figure 1 shows the work flow of proposed method using video indexing. It comprises distinct modules like Video index generation, Audio to Text Converter, Text Extractor, and Video caption Generator. The video given as input is broken into different frames. Each frame is treated as an image

When a keyword is given as input of Video search engine, it checks the video index table and identify the most relevant video by looking the dense of keyword given as input. The relevant video clip can be recognized by checking the lower and upper range of frame numbers corresponding to the key word in video index.

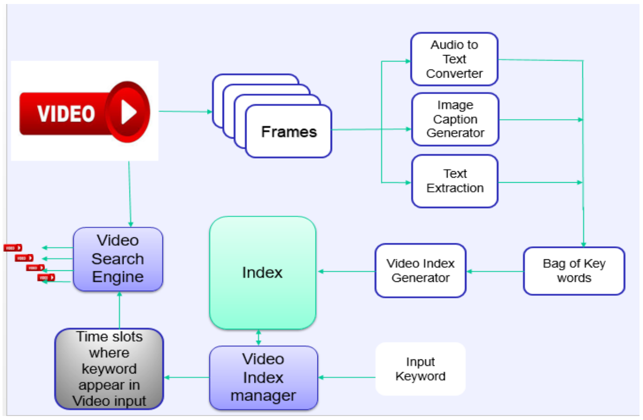

Fig 1 Video search using video index- Work flow

Audio to Text converter

The audio component present in an input frame is extracted and fed to an RNN to generate corresponding text. RNN is a neural network structure having ability to memorize that resolves the future estimates After feeding chunks of audio slices of 20ms to the neural structure, it will generate letter which equals the vocalized sound.

Text Extractor

The module text extractor extract text from the frames if it present. Fundamentally it works in three stages: In first stage, edge generation using Line edge detection mask is applied. In second stage, text localization using projection profiles based, is done. In the third stage, text segmentation and text recognition is carried out on the Localized frames. An efficient algorithm for text localization and extraction for detection of both graphics and scene text in video images proposed in [15] is used for the task.

In the Text Segmentation stage, each localized video image frame is converted to intensity-based edge map, using a Sobel edge operator. A morphological process is now applied to edge based intensity image. These steps are used for diluting the Strong and weak sub bonds.

Video Caption Generator

In this stage, Captions are generated for the video using NLP. The major difficulty is how to create image-specific descriptions within the context of overall frames. Here, we use the VIST-Visual Story Telling dataset for the generation of multi stage cued story, composed of Five sentences [16,17]. Visual story is generated by the deep learning network model by combining global local (*glocal*) attention and context cascading mechanisms. i.e., global-overall encoding level and local-image feature level. In the image sequence encoders, the bi-directional LSTMs is used encoding the global context of the storyline using features of five images, Global attention on the context and, Local attention to image features. Both of these are united and sent to RNN-based sentence generators. Large number of parameters are used for the standard attention configuration. It is realized by hard connections from the outputs of encoders or image features onto the sentence generators. The coherency of the generated stories can be improved, by conveying the latest hidden vector in the sentence generator to the subsequent sentence generator as an initial hidden vector.

Initially the video is separated as a set of image sequence. The features are extracted from each image using the ResNet-152. Then extracted features are fed sequentially into the bi-LSTM(Bidirectional LSTMs) so that the context of the images can be uniformly presented in the perfect story. The glocal vectors are created using bi-LSTM outputs and image-specific features go through the fully connected layers. Then, it is concatenated to the word tokens in order to be used as inputs to the decoder. One glocal vector is used until the decoder meets a token '<END>' which represents the end of the sentence; five glocal vectors created for each image in the same method. The cascading mechanism carries the hidden state (context) of the previous sentence to the following sentence. The hidden state of the LSTM is set to zeros only at the opening of the first sentence of the story for keeping story context.

Video Index Generator

The bag of words generated out of the modules Video Caption generator, Text Extractor, and Speech Recognition System is given as the input of a Video Index Generator. Index generator use a new algorithm named Wormhole[18], that takes *O(logL)* worst-case time for searching a key with a length of L. Finally, the Video index generator generate an index which contain the keyword, dense, lower-upper frame range of the significant video clip.

IV. RESULTS AND DISCUSSION

The proposed work implemented in Python IDE and tensor flow deep learning framework. Video captions are generated by the deep learning network model by combining global local (glocal) attention and context cascading mechanisms using VIST-Visual Story Telling dataset. Video Index generator used Wormhole algorithm to ensure minimum time complexity. 96% accuracy achieved in a cricket match video.

V. CONCLUSION AND FUTURE SCOPE

In this paper, an Efficient Video Searching methodology using Video Indexing is proposed using the Video, Audio, and Textual information. RNN based speech recognition model is used for audio to text conversion, OCR technique is used for Text extraction from preprocessed frames. Video captions generated for preparing the video index from video content uses the VIST-Visual Story Telling dataset for the generation of multi stage cued story. Visual story is generated by the deep learning network model by combining global local attention and context cascading mechanisms. i.e. global-overall encoding level and local-image feature level.
Effective implementation of this methodology in Video Search Engine, will initiate incredible changes in data traffic by minimizing the size of video transport. Also, from the user point of view, the intended part of video only need be accessed.
As a future scope of work, the same algorithm can be implemented and tested in a news video which contains various contextual contents in various topics.

REFERENCES

[1] Cisco, *"Visual Networking Index: Forecast and Trends, 2017–2022"*, CISCO, **February 27, 2019**.

[2] Z. Cao, and M. Zhu, *"An Efficient Video Similarity Search Algorithm"*, IEEE Transactions on Consumer Electronics, Vol. **56**, No**. 2, May 2010**.

[3] Q. Chen , K. Kotani , F. Lee and T. Ohmi, *"A fast search algorithm for large video database using HOG based features"*, David C., Wyld et al. (Eds) : ITCS, JSE, SIP, ARIA, NLP-**2016**, pp. **35–41, 2016**.

[4] H.Aradhye, G. Toderici, and J. Yagnik. *"Video2text: Learning to annotate video content"*,.ICDM Workshop on Internet Multimedia Mining, Google, Inc,USA,**2009**.

[5] N. Krishnamoorthy, G. Malkarnenkar, R. J. Mooney, K. Saenko, and S. Guadarrama,. *"Generating Natural-language video descriptions using text-mined knowledge"*. In Proceedings of the Workshop on Vision and Natural Language Processin*g*, pp **10-19,July 2013**.

[6] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, *"Long-term recurrent convolutional networks for visual recognition and description", arXiv:**1411.4389v4** [cs.CV]*, **May 2016**

[7] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, *"Show and tell: A Neural Image Caption Generator", arXiv:1411.4555v2 [cs.CV]*, **April 2015**.

[8] S. Venugopalan, H. Xu, J. Donahue, M. Rohrbach, R. Mooney, and K. Saenko. *"Translating videos to natural language using deep recurrent neural networks",* arXiv:1412.4729v3 [cs.CV], **April, 2015**.

[9] J. Liu , Q. Yu , O. Javed , S.Ali, A. Tamrakar, A. Divakaran, H.Cheng, H. Sawhney *"Video event recognition using concept attributes",* IEEE Workshop on Applications of Computer Vision (WACV), **March 2013**

[10] M. Mazloom, A. Habibian and C.G. M. Snoek ISLA,. *"Querying for Video Events by Semantic Signatures from Few Examples".* Proceedings of the 21st ACM International Conference on multimedia, **pp 609-612 October, 2013.**

[11] S. Venugopalan, M. Rohrbach, J.Donahue Raymond Mooney, T. D. Kate Saenko, *"Sequence to Sequence – Video to Text"*, Proceedings of the 2015, IEEE International Conference on Computer Vision (ICCV),Pages **4534-4542 December, 2015**.

[12] N.Gayathri, K.Mahesh, "*A Systematic study on Video Indexing"*, *International Journal of Pure and Applied Mathematics* Volume **118** No. **8 2018, 425-428**

[13] M.Ravinder, T.Venugopal, Sultanpur, Medak, "*Content-Based Video Indexing and Retrieval using Key frames Texture, Edge and Motion Features"*, International Journal of Current Engineering and Technology, Vol.**6,** No.**2,April,2016**.

[14] N. Laokulrat, S. Phan, N. Nishida, R. Shu , Y.Ehara , N. Okazaki, Y. Miyao and H. Nakayama, *"Generating Video Description using Sequence-to-sequence Model with Temporal Attention"*, Proceedings of International Conference on Computational Linguistics: *Technical Papers*, pages **44–52**, Osaka, Japan, **December, 2016**.

[15] A. Kumar , R. K. Goel , *"An Efficient Algorithm for Text Localization and Extraction in Complex Video Text Images"*, IEEE International Conference on Information Management in the Knowledge Economy,**2013**.

[16] T. Hao K. Huang, F. Ferraro, N.. Mostafazadeh, I. Misra, J. Devlin, A.Agrawal, R. Girshick, Xiaodong He, "*Visual storytelling".* Annual Conference of the North American Chapter of the Association for Computational Linguistics", *arXiv:**1604.03968v1** [cs.CL]*, **2016**.

[17] T. Kim, M. OhHeo, S. Kyoung-WhaPark ,B. T. Zhang *"GLAC-Net: GLobal Attention Cascading Networks for Multi-Image Cued Story Generation*", arXiv:**1805.10973v3, Feb 2019**.

[18] X. Wu, Fan Ni , S. Jiang, "*Wormhole: A Fast Ordered Index for In-memory Data Management",* arXiv**:1805.02200v2** [cs.DB], **May 2018**.

Authors Profile

First Author- Jaimon Jacob, attained the degrees B.Tech in Computer Science and Engineering from University of Calicut in 2003, M.Tech in Digital Image processing from Anna University, Chennai in 2010, MBA in Information Technology from Sikkim Manipal University in 2012, M.Tech in Computer and Information Science from Cochin University of Science and Technology in 2014. Currently working as Asst. professor in Computer Science and Engineering, Department of Computer Science, Govt. Model Engineering College. Thrikkakara, Ernakulam, Kerala. Four International Conference papers and Two National Conference research papers published. Author passionate in research area "video processing". Associate with professional bodies ISTE,IETE and IE.

Second Author-Prof.(Dr.) Sudeep Ilayidom attained the degrees B.Tech, M.Tech, PhD. Currently Working as Professor, Division of Computer Engineering ,School of Engineering, Cochin university of Science and Technology. Ernakulam, Kerala. Published a Text book on "Data mining and warehousing" by Cengage Fifty Five research papers published in the related area Data mining. A well known musician in Malayalam Film Industry. Passionate ion research area Data Mining, Big Data and related areas.

Prof.(Dr.) V.P.Devassia attained the degrees B.Sc. Engineering from MA College of Engineering, Kothamangalam, in 1983, M.Tech in Industrial Electronics from Cochin University of Science and Technology, Ph.D in Signal Processing from Cochin University of Science and Technology in 2001. Worked as Graduate Engineer(T) in Hindustan Paper Corporation Ltd, Design Engineer, HMT Limited, Principal, Govt. Model Engineering College, Ernakulam. Author passionate in research area Signal Processing. associate with professional bodies ISTE,IETE and IE.