International Journal of Electrical and Computer Engineering (IJECE)
Vol. 10, No. 6, December 2020, pp. 6019~6025
ISSN: 2088-8708, DOI: 10.11591/ijece.v10i6.pp6019-6025

Video content analysis and retrieval system using video storytelling and indexing techniques

Jaimon Jacob[1], M. Sudheep Elayidom[2], V. P. Devassia[3]
[1]Govt. Model Engineering College, India
[2]Cochin University of Science and Technology, India
[3]St. Joseph's college of Engg and Technology, India

Article Info

Article history:

Received Aug 7, 2019
Revised May 5, 2020
Accepted May 20, 2020

Keywords:

Video content analysis
Video indexing
Video searching
Visual story telling

ABSTRACT

Videos are used often for communicating ideas, concepts, experience, and situations, because of the significant advances made in video communication technology. The social media platforms enhanced the video usage expeditiously. At, present, recognition of a video is done, using the metadata like video title, video descriptions, and video thumbnails. There are situations like video searcher requires only a video clip on a specific topic from a long video. This paper proposes a novel methodology for the analysis of video content and using video storytelling and indexing techniques for the retrieval of the intended video clip from a long duration video. Video storytelling technique is used for video content analysis and to produce a description of the video. The video description thus created is used for preparation of an index using wormhole algorithm, guarantying the search of a keyword of definite length L, within the minimum worst-case time. This video index can be used by video searching algorithm to retrieve the relevant part of the video by virtue of the frequency of the word in the keyword search of the video index. Instead of downloading and transferring a whole video, the user can download or transfer the specifically necessary video clip. The network constraints associated with the transfer of videos are considerably addressed.

Corresponding Author:

Jaimon Jacob,
Govt. Model Engineering College,
Kerala, India
Email: jaimon@mec.ac.in

1. INTRODUCTION

Nowadays, social media and networking tools are very common and accessible to all users. Video messages are predominantly using in these platforms for conveying their views, concepts and ideas. As an outcome of this exponential growth in video communication, the amount of video data transmitting in communication networks also increases in the exponential rate.

Based on the white paper released by CISCO indicates visual networking index (VNI), clearly estimating the surge in traffic of global IP by three times in 2022 compared with that in 2017 [1], because of this growth in video data. According to these studies, the video data transmitting over the internet is mostly contributed by video file sharing, video games and video conferences, and by 2022 will constitute 82% of the total data traffic. The importance of video traffic handling can be understood from these studies and developing an envisioned specific part downloading provision from entire video results in improved handling of video traffic.

The technique of video storytelling is the action of describing the video like a story. It analyses the contents of each frame and identifying significant video clips from a long video. In the first stage, a context-aware multimodal embedding learning framework is implemented for extracting the contextual

meaning of event dynamics in each frame. In the second stage, a Narrator model is used for generating a story from the video frames.

Video indexing formulates a method of creating an index from a video which is given as input using the video storytelling technique. It is similar to the index appear in textbooks. When a user searches for a keyword, the video search engine checks the contents of the index file and confirm that video is relevant to the search keyword using the word count available in the index file. If it is found relevant, the range of frames that is significant to the search keyword will be extracted from long video and send to the searcher. Thus, instead of transferring the entire video, the appropriate part of the video will only need to send and thus reduces the video traffic drastically.

In this paper, an innovative model using ResBRNN-kNN video storytelling and Wormhole indexing techniques is proposed for the content analysis of the video and its retrieval. Hence, when a search keyword is raised, an appropriate video clip can be identified, extracted from a long video and can be transferred to the seeker, resulting in decreased video traffic by avoiding the need for entire video transfer.

This paper is organised as follows. In section 2, contemporary works related to this area is described. The proposed work is described in detail in Section 3. Results of implementation is described in section 4. Section 5 describes the conclusion and future scope for improvement.

2. RELATED WORKS

Most of the contemporary works in the area of video content analysis and indexing are based either LSTM-Long short-term memory or convolutional neural network (CNN). In [2], Venugopalan et al., proposed a frame work which is based on deep image description models for translating videos to natural language using deep recurrent neural networks. In the first stage, extracting the fc7 features [2] for each frame, the LSTM network is fed by mean pool of the features across the entire video at every time step. Until the LSTM picks the end-of-sentence tag, it outputs one word at each time step based on the video features (and the previous word).

The work proposed in [3] is grounded video description, which contains three modules to perform language generation namely, grounding module, region attention module and language generation module. The visual hints from the video is perceived by the grounding module, the visual clues to form a high-level inscription of the visual content is dynamically attended by region attention and the language decoding is done by the language generation module.

In paper [4], a convolutional relational machine (CRM) which is an end-to-end deep convolutional neural network is presented for identifying group actions exploiting the information by spatial relations of individual persons in image or video. It generates an intermediary activity map (spatial representation) based on activities of individuals and groups. The reduction of incorrect prophecies in the activity map is the responsibility of a multi-stage enhancement component. Finally, group activities are identified by the refined information from the accumulation component. Experimental outcomes exhibit the beneficial involvement of the data mined and signified in terms of the activity map.

A novel video captioning framework is proposed in [5], which assimilates a soft attention mechanism and bidirectional long-short term memory (BiLSTM) to produce enhanced global depictions for videos and improved recognition of lasting gestures in the videos. The long-short term memory is used as a decoder to fully explore global contextual information for producing video captions. The following are the benefits of the method proposed in the paper: 1) the BiLSTM construction systematically conserves visual data and global temporal and 2) the mechanism of soft attention distinguish and focus on principal targets from the convoluted content by help of a language decoder.

Video description generation using audio and visual cues is proposed in [6], this system relies on deep models like CNN and LSTM-RNN. The system construction of visual-only portrayal system is almost similar to that of the audio-only system differing only in the representation module feature. The feature encoding in visual-only system uses CNN while the bag-of-acoustic-words is used in the audio-only system. The two stages in the structure execution are training and test phase. The LSTM-RNN model is pre-trained using related auxiliary data and fine-tuned on the target domain data or trained using target domain training data in the training phase. This method utilizes LSTM-RNN to process incoming video frames for visual and acoustic encoding by model sequence dynamics and connecting it directly to an acoustic feature extraction module and a CNN.

LSTM-long short-term memory networks are a variant of RNN-recurrent neural network, able to learn the long-term dependencies. In [7], CNN features are extracted from video frames, then the single feature vector generated, that conveys the meaning of the entire video. LSTM used for creating the video descriptions from the merged images of separate frames. A sequence decoder is used to

create narration from the mean-pooled vector. LSTM is used as a sequence decoder, but it fails to consider the entire time-based information while generating the narration.

Semantic features used for recognising the action, scene and objects in [8]. It uses a semantic concept space to model the events using the semantic characteristic features. In the realm of video event interpretation and cataloguing, the advantages of concept-based event representation (CBER) is utilized by this method. Web video query using semantic signature illustration is implemented in [9], specifically for complex events in video query examples. To compute the variance in semantic existence of events, this method uses the off-the-shelf concept indicators.

An innovative algorithm proposed in [10] to manage the complex dynamics in videos from the real world. In this algorithm, captions are created using end-to-end sequence-to-sequence model. LSTM is the state-of-the-art technology in recurrent neural networks for generating captions for videos. In the algorithm, LSTM model is used to generate the description of an event in a video clip. To connect a sequence of video frames with a sequence of words, LSTM model is trained on video sentence pairs. The dynamic structure of the frames in a video sequence can be analysed and can be learned is the major merit of this model.

In [11], various video indexing techniques are compared to their merits and demerits. A detailed portrayal of video indexing presented in this paper and its significance in content-based information retrieval, event feature extraction and multimodal video indexing. An innovative algorithm proposed in [12], video indexing and retrieval using video content analysis. Various semantic features like motion features, edge and keyframe texture used for video content analysis. These semantic features are abstracted to characterise a video using a feature vector.

Method of sequence-to-sequence with mechanism of temporal attention is used in paper [13], to generate an image caption automatically considering the temporal dynamics of each frame in the image given as input. Recurrent neural network (RNN) is used for generating the captions. In [14], an innovative content-based video indexing and reclamation algorithm is proposed. This framework uses the correspondence-latent dirichlet allocation (corr-LDA) probabilistic framework for video content analysis and indexing. The major merit of this proposed algorithm is automatically annotating the videos which are stored in video database and return meaningful description based on semantic relations between scenes.

The merits and demerits of different assessment metrics such as WMD, CIDEr, SPICE, ROUGE, METEOR and BLEU used in the video descriptions presented in [15] in terms of deep learning models, data sets used, no of classes and different domain. The effective usage of linguistic knowledge which is extracted from a huge text corpus, in describing a video content is presented in the paper [16]. An innovative video summary framework is suggested in [17]. It uses attentive encoder decoder network (AVS) to retrieve the semantic information from the video frames, using the BiLSTM bidirectional long term memory.

In [18], a novel framework for video summarization is presented for domain-specific videos. The algorithm considers the characteristics features most relevant to the particular domain and generating a summary which describes the contextual meaning f input video. In [19], address various tasks involved in video content analysis and indexing such as to develop an interactive environment with objects for video, to manage video content management tasks, to define a characteristic architecture, to develop automated tools and techniques for video content recognition and representation, to apply the techniques of knowledge representation for index construction development and methods for restoration. A JEDDi-Net (joint event detection and description network) is proposed in paper [20] that solves the task of dense video description in possible ways. The model continuously codes a three-dimensional convolutional layer input video stream, suggests and generates varying time events on the basis of pooled characteristics.

A novel video description model has been proposed in paper [21] to allow use of bounding box annotations and produce grounded subtitles. In this study, the video facts are compared with the subtitle sentence by referencing the noun of the sentence in one of the frames of a video. In [22], a new dense framework for video captioning was proposed, which specifically models the temporal dependency of occurrences in the video and utilizes visual and language contexts for coherent storytelling from past events. From the detailed literature survey, it is observed that video storytelling technique-based video content analysis and retrieval is not addressed.

3. PROPOSED WORK

The whole video is considered in all video search algorithms mentioned in recent literature, althougha portion of the video may be of interest to the stakeholders. However, the proposed model only allows content wise retrieval of the interesting video frames, significantly reducing the bandwidth.

A righteous concept to speed up the recovery of video data using a video index is proposed in this approach. Video index is analogous to the page index that looks in textbooks, where every other important word or phrase is arranged in a sorted order together with the dense and frame. The major functional units appear in this concept are VIG for generation of the video index, ATC for converting audio to text, TE for extraction of the text, and VCG for generating video caption as illustrated in Figure 1.

ICG generates captions to each image after image content analysis. Audio to text converter generates the keywords based on the audio available with the video. Text extractor extracts the textual information present in the images. The textual information available from these three functional units is given as input to video index generator to generate the video index which contains major keywords, dense and range of frames where this major keyword appears. Video search engine examines the content of the video index table and confirm whether this video is relevant to the search keyword by checking the dense in the index table. If it is found relevant, the significant part of the video clip will be extracted from the long video using the frame range information available in the index table.

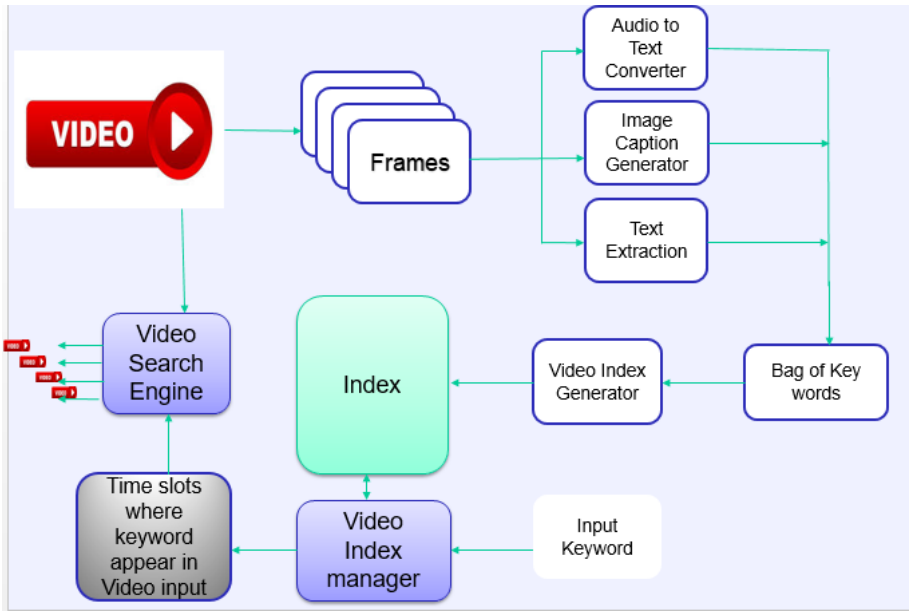

Figure 1. Process flow-video recouping model using video indexing

Video content analysis contemplate three types of information present in any video, *viz* audio, text, video. This information can be used for analyzing and to generate an error free video index. Input videos may be integrated with combination of one or more information. Absence of any one or two types of information content doesn't affect the accuracy of video index it generates. Video content analysis comprises three major blocks, Audio to Text converter (ATC), text extractor (TE), and the image caption generator (ICG).

3.1. Audio to text converter

In this module, the recurrent neural network-based technique is used to generate the textual information from the audio available with the video. Audio is sliced into chunks with 20 ms size and feed into this module as input. It will generate texts corresponding to the audio chunks using the ability to memorize and compute the estimates.

3.2. Text extractor

Text extractor generates the textual information present in the images, given as input to this module. It consists of three stages, line edge detection mask for edge generation, projection profiles for text localization, and segmentation of text & recognition of text as illustrated in [23]. In this method, an efficient algorithm is proposed to localize and extract the texts from graphics as well as scenes present in videos images. A localized video image frame is converted into intensity-based edge map by SOBEL edge operator in this algorithm.

3.3. Image caption generator

The video storytelling method [24], is used for video caption generation and it involves two subtasks: (a) identify the most relevant video clips from the input video, (b) create a narration from the video clip identified as significant. Embedding learning with multimodal semantic is used as a context-aware framework for identifying the relevant part of the video clip. A local-to-global two-phase process is used for training the Embedding. The first phase prototypes specific pairs of clip-sentence to know an embedded locale. The second stage consider the whole video as a segment sequence. The video's temporal dynamics is captured by Residual bidirectional RNN (ResBRNN) and then integrates past and future relevant information into the multimodal embedding space. The temporal coherence is preserved, and thus increasing the corresponding embedding variety.

In the second stage, the stories are created using a Narrator model. The narrator extracts a series of relevant clips from it, due to an input video clip. A sequence of sentences apt for the clips in the multimodal embedding space is retrieved for generating a storyline. It's difficult to discover the right clips because there's no simple description of what visual elements are necessary to shape a perfect story. To this end, the narrator has been formulated as a reinforcement learning agent examining the video input sequentially and then studies a strategy of choosing clips to maximize the outcome. We construct the outcome as the linguistic measure between thus received story and the reference stories written by humans. By maximizing the literal metric directly, the narrator learns to identify interesting, broad clips that form a nice narrative.

3.4. Video index generator

There are three distinct modules there to generate a story from the given video. These modules are ATC, TE, and VCG. The textual information generated by these modules is provide as input to the VIG and it uses Wormhole [25] algorithm for indexing, which guarantees worst-case time complexity $O(\log L)$ in search of a keyword with a definite length L. Along with the dense information for each keyword, video frame range will also store in this index table, which can be used by VSE while checking the relevance. Index table for the video will follow the format of keyword, dense, lower frame no, upper frame no (keywords in sorted order, dense of keyword, the lower frame number and upper frame number where the dense is found high).

4. RESULTS AND DISCUSSION

The proposed work was implemented in Python IDE and tensor flow deep learning framework. Video captions were generated using video story telling method using video story dataset, a new dataset that prepared to enable the study. VIG uses wormhole algorithm to warrant least time complexity. 86% accuracy was achieved in a video *campfire in a forest* using the stories generated by the ResBRNN-kNN method [24]. Figure 2 represent the set of shots in the video clip-Campfire in a forest used for generating story. The output of the video caption generator is as follows.

"The nature of the forest is shown while the hiker and dog are on the hiking trail. A girl talks to the camera while laying down her tent. They are preparing the campsite and starting up the trail on a hike. The campers explore the woods and hike up some rocks. The couple hike along a forest trail with their dog. The campers take turns swinging over the water. The friends start a campfire. A man is cooking his food on the fire. The campers are exploring the area around their campsite and the surrounding woods. They get back in the car and drive again."

Figure 2. Set of shots in video clip-campfire in a forest

The video index generator generates the index as shown in Table 1. Based on the story generated by video caption generator. Third column in the table represent the percentage of true positive to evaluate the accuracy.

Table 1. Video index corresponding to the video shown in Figure 2

Key word	Frame no range	True Positive
Camera	15-28,32-59,78-99,124-145	91%
Camper	12-24,35-49,56-64	89%
Campfire	140-164	79%
Car	19-24,145-161,165-199	84%
Dog	23-35,67-89,97-123	92%
Food	134-168	80%
Forest	1-9,21-31,62-87,99-120	74%
Friend	25-39,61-99, 103-123	87%
Girl	25-39,61-99, 103-123	68%
Hiker	12-24,35-49,56-64	81%
Man	15-39,61-69, 101-123,141-158	89%
Nature	1-24,34-45,65-89,102-156	81%
Water	20-31,62-87,109-129	81%
Wood	111-123,134-145,165-196	78%

5. CONCLUSION

As per the literature survey presented, the entire video needs to be transferred to the searcher while a portion of the video may be of interest to the stakeholder. In this paper, a video retrieval system is proposed to retrieve a portion of the long video which is interested, using video content analysis, video storytelling and video indexing concepts. The information present in the video in the forms of images, audio and text considered for video content analysis and subsequently for video indexing. Hence, more accuracy is ensured.

Text generation from audio input is implemented using the RNN based audio recognition model. Text extraction from the video images through various stages like edge generation based on mask-based line edge detection, text localization based on projection profiles, segmentation of text & recognition of text. Video storytelling method is implemented using the video story data set to generate the captions from video images. In video search engine, successful implementation of this framework, would facilitate enormous improvements in data traffic by reducing information exchange size. Furthermore, the intended portion of the video only needs to be viewed from the user's point of view. As a continuation of the work, the same algorithm can be implemented and tested in a news video which contains several contextual contents in various topics

REFERENCES

[1] CISCO, "Cisco Visual Networking Index: Forecast and Trends, 2017–2022," *White paper Cisco public*, 2019.
[2] S. Venugopalan, et al., "Translating videos to natural language using deep recurrent neural networks," *arXiv: 1412.4729*, 2014.
[3] L. Zhou, et al., "Grounded Video Description," *arXiv: 1812.06587v2*, 2019.
[4] S. M. Azar, et al., "Convolutional relational machine for group activity recognition," *arXiv: 1904.03308v1*, 2019.
[5] Y. Bin, et al., "Describing Video with Attention-Based Bidirectional LSTM," in *IEEE Transactions on Cybernetics*, vol. 49, no. 7, pp. 2631-2641, 2019.
[6] Q. Jin and J. Liang, "Video Description Generation using Audio and Visual Cues," *ICMR '16 Proceedings of the 2016, ACM on International Conference on Multimedia Retrieval*, pp. 239-242, 2016.
[7] C. Zhang and Y. Tian, "Automatic video description generation via LSTM with joint two-stream encoding," *2016 23rd International Conference on Pattern Recognition (ICPR)*, Cancun, pp. 2924-2929, 2016.
[8] J. Liu , et al., "Video event recognition using concept attributes," *IEEE Workshop on Applications of Computer Vision (WACV)*, pp. 339-346, 2013.
[9] M. Mazloom, et al., "Querying for Video Events by Semantic Signatures from Few Examples," *Proceedings of the 21st ACM International Conference on multimedia*, pp. 609-612, 2013.
[10] S. Venugopalan, et al., "Sequence to Sequence-Video to Text," *Proceedings of the 2015, IEEE International Conference on Computer Vision (ICCV)*, pp. 4534-4542, 2015.
[11] M. Ravinder and T. Venugopal, "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, pp. 672-676, 2016.
[12] N. Laokulrat, et al., "Generating Video Description using Sequence-to-sequence Model with Temporal Attention," *Proceedings of International Conference on Computational Linguistics: Technical Papers*, Osaka, Japan, pp. 44-52, 2016.

[13] Q. You, et al., "Image captioning with semantic attention," *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1-10, 2016.
[14] R. R. Iyer, et al., "Content-based video indexing and retrieval using corr-lda," *arXiv: 1602.08581*, 2019.
[15] N. Aafaq, et al., "Video description: A survey of methods, datasets, and evaluation metrics," *ACM Computing Surveys (CSUR)*, vol. 52, no. 6, pp. 1-28, 2019.
[16] S. Venugopalan, et al., "Improving lstm-based video description with linguistic knowledge mined from text," *Proceedings of Empirical Methods in Natural Language Processing*, pp. 1961-1966, 2016.
[17] Z. Ji, et al., "Video summarization with attention-based encoder-decoder networks," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 6, pp. 1709-1717, 2019.
[18] V. Kaushal, et al., "A Framework towards Domain Specific Video Summarization," *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 666-675, 2019.
[19] S. W. Smoliar and H. J. Zhang, "Content based video indexing and retrieval," *IEEE multimedia*, vol. 1, no. 2, pp. 62-72, 1994.
[20] H. Xu, et al., "Joint event detection and description in continuous video streams," *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 25-26, 2019.
[21] J. Aneja, et al., "Convolutional image captioning," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5561-5570, 2018.
[22] J. Mun, et al., "Streamlined dense video captioning," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6581-6590, 2019.
[23] A. Kumar and R. K. Goel, "An Efficient Algorithm for Text Localization and Extraction in Complex Video Text Images," *2013 International Conference on Information Management in the Knowledge Economy*, pp. 14-19, 2013.
[24] J. Li, et al., "Video Storytelling," *arXiv: 1807.09418v1*, 2018.
[25] X. Wu, et al., "Wormhole: A Fast-Ordered Index for In-memory Data Management," *arXiv: 1805.02200v2*, 2018.

BIOGRAPHIES OF AUTHORS

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, Govt. Model Engineering College, Kerala. Author passionate in research area "video processing". Associate with professional bodies ISTE,IETE and IE.

Prof. (Dr.) Sudheep Elayidom attained the degrees B. Tech, M. Tech, Ph. D. Currently Working as Professor, Division of Computer Engineering, School of Engineering, Cochin university of Science and Technology. Ernakulam, Kerala. A well-known musician in Malayalam Film Industry. Passionate in 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. 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 as LM-ISTE, FIETE, FIE and C.Eng. IE(I).