Cross Modal Localization of Moments in Video

Sahla Sherin O[1], Ani Sunny[2]

[1]*Master of Technology, Dept. of Computer Science and Engineering, Mar Athanasius College of Engineering, Kothamangalam*

[2]*Assistant Professor, Dept. of Computer Science Engineering, Mar Athanasius College of Engineering, Kothamangalam*

---***---

Abstract – *There has been growing interest nowadays in locating particular moments inside a video in reacting to a given sentence rather than simply retrieving the entire video, known as temporal moment localization. The current program focuses primarily on sparse frame-level characteristics such as visual expression, blurring the critical indications for finding the desired moment. In our proposed method, the video along with the sentence is given as input, and the extraction of the noun is per- formed first by natural language processing tools. Then the extracted noun words are used to know the relation between the objects and the video moment obtained with the maximum relevant objects and the relationship will be saved. In this work we developed relation network which is composed of convolutional neural network and long-short term memory network for relation recognition. The project introduces the first truly end-to-end object detector, which determines where to look next and when to guess, rather than the conventional expensive search and locate system.*

Key Words: Natural Language Processing, Deep Learning, Convolutional Neural Network, Temporal Moment Localization, Long-Short Term Memory

1.INTRODUCTION

A key challenge in computer vision is temporary localization of events or actions of interest. Recent development in deep learning and computer vision has advanced the task from action recognition to detection[7,8]. Recently there has been increasing interest in identifying the questions using natural language instead of pursuing a predefined collection of actions or events. Localization of events supports many real- time applications to find some critical moments, such as po- lice investigation in CCTV videos, etc. Localizing moments in a longer video via natural language is linked to other vision tasks such as video retrieval, video review, video description and question answering, and the retrieval of natural language objects[1]. Temporal moment localization refers to recognizing specific moment from a video in response to a textual query.

At the same time there are many problems occurs in temporal moment localization. First, recognization of relevant objects and interactions from the video is important since uncompressed videos may include complicated scenes and a large number of objects. But in the input query only a few objects are listed. Hence it is extremely difficult to differentiate between the video moment containing the related object and the interactions from other scenes. Comprehension of critical question information in the second case is highly difficult. Many keywords in the question language hold the vital semantine hints to get the desired moment back. The natural language anatomy helps one to deal not only with an open collection of objects or behaviors but also with the connections and relationships between the individuals. It is a highly demanding mission, since it requires both vision and language comprehension. Previous tasks are concentrated on the number of objects present in each frame, but they do not consider the relation between the objects. Present approaches to problems of localization, whether spatial or temporal, are often based on a proposal and classification. Here candidate regions are first generated by a method and then fed to a classifier to obtain the probabilities of the target classes being contained. These methods of generating action proposals make predictions based on the ”actionness” score of each short snippet in the videos[9,10,11].

The detection in many of past works is based on the artifacts defined in the textual query. They don't know how connected the objects are. Hence these approaches contribute to independent identification of the video moment being queried. Be- cause one item may be seen in different frames, it also brings in more moments. We are implementing a Cross-Modal Temporal Moment Localization with Relation Network (CTLRN) to address those limitations. The software mainly focuses on the relation between the items and the demand. The model is trained to identify the specific objects that are identified in the query, and to recognize the different relationships or behaviors between the objects.

The remainder of the paper is set out as follows. The second section offers a thorough overview of the prior studies presented in the literature on the research subject. The characteristics of the project are described in the proposed system, third part. This section also gives the project comprehensive design, solution process and architecture diagrams. And finally it ends with the project's importance relative to the previous work, room for future work and clearly state the points with the original target.

2. RELATED WORKS

Activities consist of a diverse mix of actors, actions, and objects over different time periods. Earlier research focused on classifying video clips showing a single incident, where the videos had been cut. C.V.Jawahar [2] introduced a video- retrieval model focused on text queries. The method allows for search based on video-contained textual information. The videos are annotated based on the textual contents which are defined from the frames. Rather of using OCR the model suggests an method that enables the text to be extracted within the video by matching at the image level. There are two phases to the system suggested.

There has also been extensive work recently in identifying events in longer untrimmed videos. Jiyang Gao[3] introduced a temporal activity localization model for finding specific moment from an untrimmed video with respect to the textual query. A novel Cross modal Temporal Regression Localizer (CTRL) was proposed for the joint modeling of text query and video clips. The approach generates alignment score for candidate clip along with localization regression test. It takes advantage of a CNN model to extract the clips and LSTM network visual features to extract embedded sentences. A cross- modal processing module is designed to model the text and the visual characteristics together. Finally a multilayer network is equipped for the visual alignment of sentences and the regression of clip locations. In 2017 Hendricks[1] also introduced another method for localizing moments in video with natural language. This proposes a Moment Context Network (MCN) that includes a global video framework to include a temporal context and a temporal endpoint function to show when a moment in a video occurs. Existing methods for video retrieval based on natural language retrieve a full video based on text string data, but do not define when a moment occurs in a video. A collaborative video-language model in which referencing words and video features from the corresponding moments are near in a shared embedding space is proposed for the user to understand.

From the proposed methods[3,1] one considers proposed temporal regions just before and after the proposed region and one considers video context in the form of a global-context function that represents the entire video. Both can implicitly provide suitable contextual moment in their context apps, and do not evaluate appropriate context for the query. In 2018 Hendrick[4] suggested another approach for addressing this problem, Moment Localization with Latent Context (MLLC), which models video context as a latent variable. The latent variable helps the model to attend different video contexts that are dependent on unique query-input pairs. It also provides versatility in terms of location and contextual moment duration. Recently several existing works [3,1] leverage one temporal sliding window approach over video sequences to generate video segment candidates, which are then independently combined [3] or compared [1] with the given sentence to make the grounding prediction. In order to tackle the limitations, Jingyuan Chen [5] introduced a novel Temporal Ground Net(TGN) model, that takes full advantage of fine-grained interactions between video frames and words in a sentence. TGN sequentially processes video frames, where at each time step it rely on a novel multi modal interactor to exploit the evolving fine-grained frame by-word interactions. Then, TGN works on the yielded interaction status to simultaneously score a set of temporal candidates of multiple scales and finally localize the video segment that corresponds to the sentence. Dongliang He [6] introduced another strategy for temporal moment localization. The approach proposed an end-to-end reinforcement learning (RL) based framework for grounding natural language descriptions in videos. Here an agent iteratively reads the description and watches the entire video as well as the temporary grounding clip and its boundaries, and then it determines where to move the temporal grounding boundaries based on the policy. At each time step, the agent makes an observation that details the currently selected video clip, and takes an action according to its policy. The environment is updated accordingly and offers the agent a reward that represents how well the chosen action performs.

3. PROPOSED SYSTEM

The proposed system mainly consists of three parts. (1) Language Processing Module (2) Video Processing Module and (3) Output Module. Fig - 1 shows the architecture diagram of the proposed system.

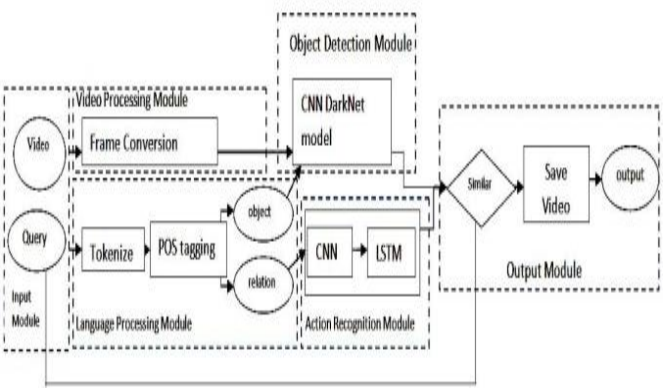


Fig -1 System Architecture

3.1 Language Processing Module

The input sentence contains relevant information for the retrieval of the moment from the video. The extraction of relevant knowledge can be achieved through the application of Natural Language Processing (NLP). NLP was originally used for computer translation and speech recognition. The basic principle of using NLP in the extraction of knowledge is to analyze grammatical structure at sentence level, then construct grammatical rules within sentence for some useful details. NLP techniques rely on syntactic and semantine information, often manually encoded for a given domain. Initially the NLP is used for machine translation, speech recognition and also information representation. Work on knowledge extraction uses NLP techniques to preprocess documents and to extract the knowledge that underlies them. The approaches mentioned in this section basically rely on word recognition in the documents, and use NLP techniques such as automated Part-of-Speech Tagging to preprocess textual data and Term Extraction to extract the useful information. NLP techniques can be seen as an automated, simplified indexing method that extracts linguistically relevant structures from the document's full textual content. These techniques are ideal for extracting keywords from a sentence.

The input sentence contain keywords like the objects present in the video and also how the objects are related to each other. To extract these information we need to apply NLP techniques to the sentence. The most important NLP techniques are the marking of Part-of-Speech (POS) and extraction of words. The Part-Of-Speech tagging (POS-Tagging) aims to automatically assign part-of-speech tags (i.e. morpho-syntactic categories like noun, verb, adjective ...) to descriptive terms. The system's input module (Fig -1) determines what input is given to the machine. Here an untrimmed video and a textual query is given as the input. From the textual query the system will extract different keywords that can identify the required seg- ment from the video. The video is processed frame by frame.

The language processing module is implemented by using NLTK tools. NLTK (Natural Language Toolkit) is a leading forum for the development of Python programs for human language data work. It offers interfaces for many corporate and lexical tools which are simple to use. It also provides a suite of text-processing libraries for grouping, tokenization, halting, tagging, sorting, and semantic reasoning. Firstly, the input sentence is tokenized. The process is known as tokenisation of the word. The problem of splitting a string of written language into its component words is word tokenisation. After that the words are added to the part-of-speech tagging. The fig-2 also shows NP chunking. From the tagged part noun and verb is extracted as object and relation respectively. Consider fig -2 with an example sentence "A person walking with a book". Here the objects are person and book and the relation between them is walking.

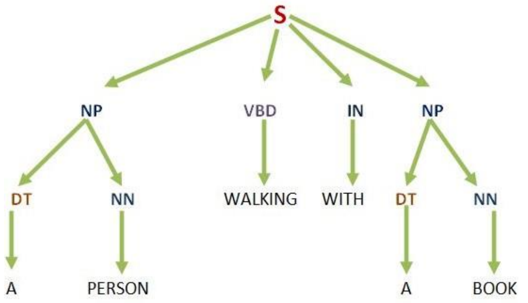


Fig -2 POS Tagging

3.2 Video Processing Module

Video data is processed, and the frames are discarded. That is the video that is translated into many frames and stored in a folder. Frame by frame scanning is done to complete image processing. The processing of the video is further divided into two parts: module for object detection and module for the identification of reference. The object detection module helps identify various objects that are identified in the input query. The relation recognition module find relation between the objects which is a major part of the system.

In object detection module the input is frame or images which is extracted from the input video and it is stored in a folder. The first step is pretreatment. Preprocessing refers to all the raw data transformations until they are fed into the machine learning algorithm or profound learning. For example, the training of a convolutionary neural network on raw images could lead to poor classification results [16]. It's also essential to speed up training by preprocessing. Preprocessing can involve steps to resize, label, segment and morphology. The object detection network is built by CNN. DarkNet is the basis for the network proposed for object detection. It is where the whole image is connected to a single neural network. This network divides this picture into regions and predicts bounding boxes and probabilities for each region. Such bounding boxes are weighted according to predicted probabilities. This network has varying advantages over other object detection framework. The network shows only those objects that have confidence greater than 0.5. Fig- 3 displays model DarkNet for detection of artifacts. It has 24 convolutionary layers and two fully connected layers. Alternating 1 x 1 convolutional layers eliminates the space characteristics from previous layers. We used 4 convolution layers and 2 fully connected layers in the pre-trained model to improve performance.

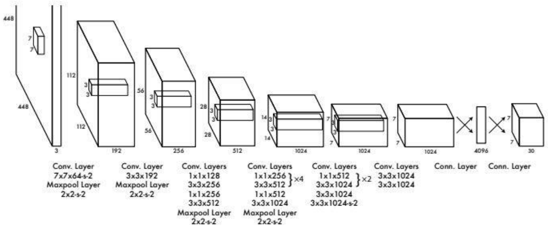


Fig -3 DarkNet Model

Relation recognition or action recognition is an important module in this system. The module is made of CNN and LSTM as backbone. The technique of action recognition is divided into two parts: Firstly, we remove CNN features from the video frames. Secondly, features that reflect the time interval sequence of operation are fed to the LSTM. CNN is a dominant outlet for representing and classifying pictures. In the case of video data, each frame is represented by the CNN features, followed by the sequential information between them using LSTM (Fig-4). A video is a mixture of frames that shift between frames 30 and N per second. This documents all the tiny modifications in each frame as CNN detects hidden patterns in pictures. Those sequential form changes are learned via RNN for action recognition in a picture. We learned through RNN to recognize motion in a picture. Training a deep-learning model for image representation requires thousands of images and needs also high processing power such as GPU to change the weight of the CNN model [18].

In this process, parameters of the pre-trained CNN model are used for extraction of features, which is trained on a large- scale ImageNet [17] dataset of more than 15 million images. Video is also sequential data in which visual object movements are expressed in several frames, so that frame sequence helps to understand the meaning of an event. RNNs may display such sequences but they forget the earlier inputs of the sequence in case of long-term sequences. This problem is known as the vanishing gradient problem that can be solved via a special form of RNN called LSTM. For bidirectional LSTM, the output at time t depends not only on the preceding frames of the series, but also on the frames ahead.

Bidirectional RNNs are fairly simple, stacking two RNNs atop each other. Another RNN moves in the direction of and another goes backward. The combined output is then determined according to both RNNs' hidden state. The module utilizes several layers of LSTM, and the system has two layers of LSTM for both forward and backward passes. The related functions from the input query are applied to the file processing module to find the features from the input file. If the object from the object detection module is matched to the input query by the relation from the relationship recognition module in a frame, then the frame is stored in a folder. After the entire video has been processed, the frames detected are converted to a video and saved.

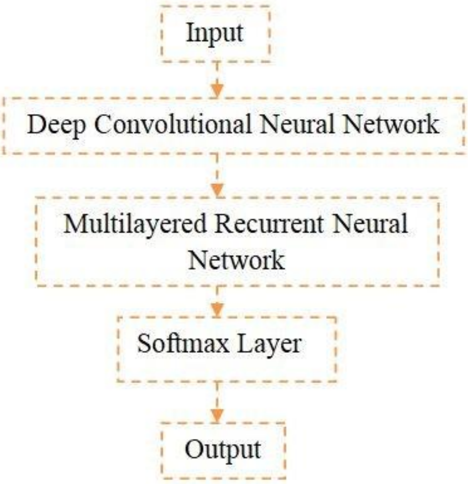

Input
Deep Convolutional Neural Network
Multilayered Recurrent Neural Network
Softmax Layer
Output

Fig - 4 Relation Network Model

4. EXPERIMENT

4.1 Data Description

COCO Dataset: COCO dataset is for module detection of artifacts. COCO is a data collection for the identification, segmentation and captioning of large objects. COCO has several features such as object segmentation, recognition in context etc. COCO contains 330K images (greater than 200K labeled). The dataset contains photos of 91 objects types [13].

UCF101: UCF101 is an action recognition dataset having 101 actions with 13320 videos [15]. For which 9537 videos are used for training and 3783 for testing. In the dataset 101 activities are categorized into 25 groups in which each category can consist of 4-7 action videos. The types include contact between human-objects, body motion alone, musical instrument playing and human-human interaction.

4.2 Experimental Settings

Evaluation Protocols: In order to evaluate the performance of our method and the baselines, we follow the evaluation metric "R@k, IoU=θ " setup in [12]. More specifically, for each sentence query, we calculate the temporal Intersection over Union (IoU) between the predicted moment candidates and ground truth. Then for each IoU larger than m, we compute the percentage of top-n results. In the following paragraphs, we use R (k,θ) to denote "R@k, IoU=θ ". Table 1 and 2 shows performance comparison of our system CTLRN with state-of-arts MCN and SLTA with R@1 and R@5 respectively.This is illustrated in Chart -1 and Chart -2.

Table -1: Performance comparison of CTLRN with R@1 with state-of-arts.

Methods	R@1		
	IoU=0.3	IoU=0.5	IoU=0.7
MCN	32.59%	11.67%	2.63%
SLTA	38.96%	22.81%	8.25%
CTLRN	**40.58%**	**25.01%**	**11.25%**

Table -2: Performance comparison of CTLRN with R@5 with state-of-arts.

Methods	R@5		
	IoU=0.3	IoU=0.5	IoU=0.7
MCN	89.52%	54.21%	14.56%
SLTA	94.01%	72.39%	31.46%
CTLRN	**96.05%**	**81.20%**	**33.28%**

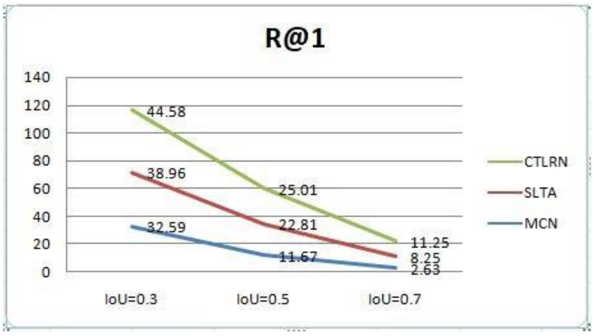


Chart -1: Performance comparison of CTLRN at R@1

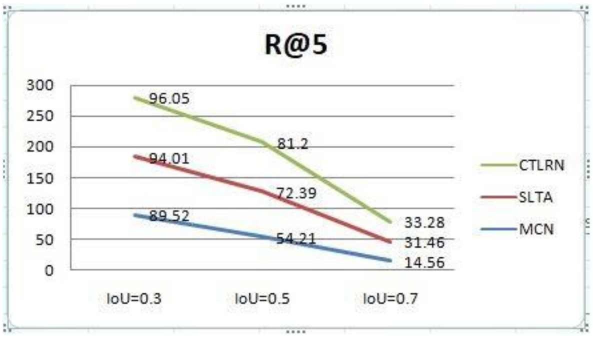
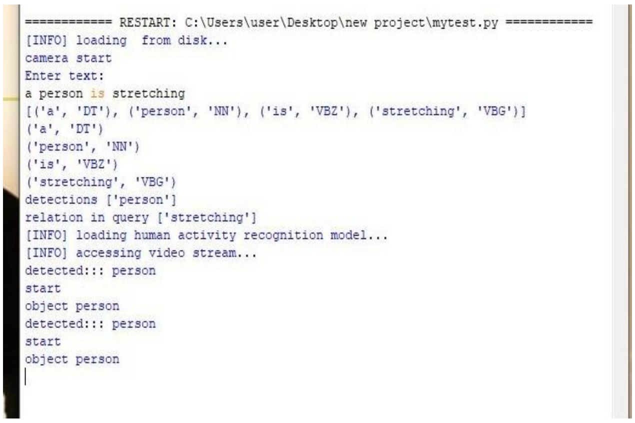


Chart -1: Performance comparison of CTLRN at R@5

5. RESULT

This section provides the various results that have been obtained. We have developed the system in python. Our system mainly consists of language processing module, Object detection module and action recognition module. The result is saved as frames in a folder. After processing of entire video, the frames are converted to video and it will shown as the required output. The language processing module provides different objects and relation present in the input query. For example consider a query " A person is stretching " . The query will get objects like person and the relation stretching (Fig -5). From the query the system can understand the moment that should be retrieved from the input video. Some examples are shown in Fig- 8.



```
============ RESTART: C:\Users\user\Desktop\new project\mytest.py ============
[INFO] loading  from disk...
camera start
Enter text:
a person is stretching
[('a', 'DT'), ('person', 'NN'), ('is', 'VBZ'), ('stretching', 'VBG')]
('a', 'DT')
('person', 'NN')
('is', 'VBZ')
('stretching', 'VBG')
detections ['person']
relation in query ['stretching']
[INFO] loading human activity recognition model...
[INFO] accessing video stream...
detected::: person
start
object person
detected::: person
start
object person
```

Fig -5: Screenshot of language processing module.

After the language processing module, the object is passed to the object detection module and the relation is sent to relation recognition module. The object detection module will find the object based on the pre-trained model. Similar to that relation recognition module will find the relation with that object. Screenshot of the output from this module is shown in Fig -6 and Fig -7.

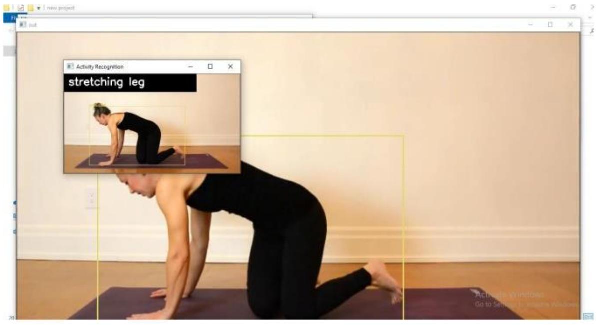


Fig -6: Screenshot of video processing module.

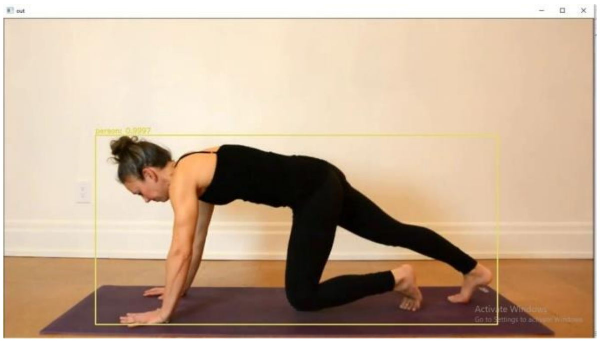


Fig -7: Figure 10: Screenshot of object detection module.

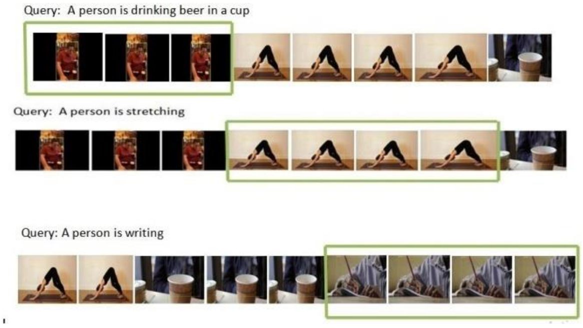

Query: A person is drinking beer in a cup
Query: A person is stretching
Query: A person is writing

Fig -8: Examples for CTLRN.

6. CONCLUSION

In this paper we deal with a problem of temporal moment localization. To better align the sentence query into the video we have introduced a new model named as Cross-Modal Temporal Moment Localization with Relation Network (CTLRN). Our system have mainly three components. That is language processing module which uses natural language processing tools for extracting relevant keywords from the query, object detection module which extracts different object from the video that is defined in the query and finally the relation recognition module which extracts different relation between the objects. The system will process the video frame by frame. If the features from video processing module matches the query features it will save the frames as the output.

To verify the effectiveness of the model, extensive experiments done on three public datasets. The system shows more accuracy over other state-of-arts. Our system takes a huge computational power. Therefore in future we are planning to increase the accuracy by including new technologies.

REFERENCES

[1] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell and Bryan C Russell, Localizing Moments in Video with Natural Language, ICCV- 2017.

[2] C.V. Jawahar, Balakrishna Chennupati, Balamanohar Paluri, Nataraj Jammalamadaka, Video Retrieval based on Textual Query, International Conference on Ad- vanced Computing and Communication, 2005.

[3] Jiyang Gao, Chen Sun, Zhenheng Yang and Ram Nevatia , TALL: Temporal Activity Localization via Lan- guage Query, ICCV-2017

[4] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, Bryan Russell, Localizing Moments in Video with Temporal Language, EMNLP-2018.

[5] Jingyuan Chen, Xinpeng Chen, LinMa, Zequn Jie, Tat-Seng Chua, Temporally grounding natural sentence in video, EMNLP-2018.

[6] Dongliang He, Xiang Zhao, Jizhou Huang, FuLi, Xiao Liu, Shilei Wen, Read, Watch, and Move: Reinforcement Learning for Temporally Grounding Natural Lan- guage Descriptions in Videos, EMNLP-2018.

[7] Simonyan, K., and Zisserman, Two-stream convolutional networks for action recognition in videos, NIPS 2014

[8] Karpathy, A. Toderici, G. Shetty, S. Leung, T. Sukthankar, and Li F, Large-scale video classification with convolutional neural networks, CVPR 2014

[9] Zhao, Y. Xiong, Y. Wang, L. Wu, Z. Tang, X. and Lin D, Temporal action detection with structured segment networks, ICCV 2017

[10] Yuan, Z.Stroud, J.C.Lu, T.and Deng,J., Temporal action localization by structured maximal sums, CVPR 2017.

[11] Lin, T. Zhao, X. Su, H. Wang, C. and Yang, M.,BSN: boundary sensitive network for temporal action proposal generation, ECCV 2018

[12] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. 2017. TALL: Temporal Activity Localization via Language Query. In Proceedings of the IEEE International Conference on Computer Vision. IEEE, 52675275

[13] https://www.google.com/url?sa=t&source=web&rct=j & url=http://cocodataset.org

[14] Gunnar A Sigurdsson, Gul Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. 2016. Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding. In Proceedings of the European Conference on Computer Vision. Springer, 510526.

[15] https://www.crcv.ucf.edu/data/UCF101.php

[16] N.Chumerin, convolutional neural network, 2015.

[17] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, ”ImageNet: A largescale hierarchical image database” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit (CVPR).

[18] C. Feichtenhofer, A. Pinz, and A. Zisserman, ”Convolutional two-stream network fusion for video action recognition,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

BIOGRAPHIES

Sahla Sherin O received Master of Technology in Computer Science and Engineering from Mar Athanasius College of Engineering, Kothamangalam affiliated to APJ Abdul Kalam Technological University in 2020. Her research interest is in Deep Learning, Data Mining and Blockchain.

Ani Sunny is currently working as assistant professor in the Department of Computer Science and Engineering, Mar Athanasius College of Engineering, Kothamangalam, Kerala, India. Her research interest is in Networks, Image Processing, Hardware and microprcessors.