AUTOMATED VIDEO SEGMENTATION PROJECT

Arjun Sadashiv, Arun Gangadharan M
Jibin Thomas Daniel, Sanat Jayakrishnan
Roll No: AM.EN.U4EAC20012, AM.EN.U4EAC20013, AM.EN.U4EAC20030, AM.EN.U4EAC20058

Guide: Ms. Anuraj K, Designation: Asst. Prof. Sr. Grade
Co-Guide: Dr.Vivek Venugopal, Designation: Asst. Prof. ECE Department of Electronics and Communication Engineering Amrita School of Engineering, Bengaluru

Objective

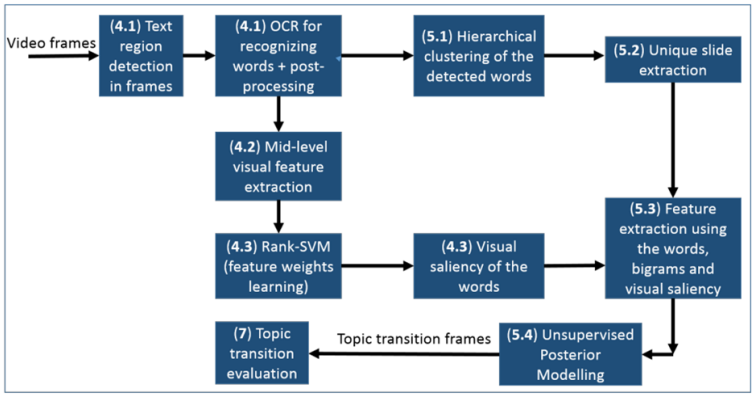
- The main objective of our project is to create an automatically applied lecture video segmentation system that splits lectures into bite-sized topics using audio cues.
- Prior works in this regard have considered the video frames and their output from an Optical Character Recognition (OCR) system for segmentation.
- The performance of our proposed system will be computed using metrics such as Normalized Mutual Information (NMI), Mean over Frames (MoF), F1-score, and Intersection over union (IoU).

Literature Survey

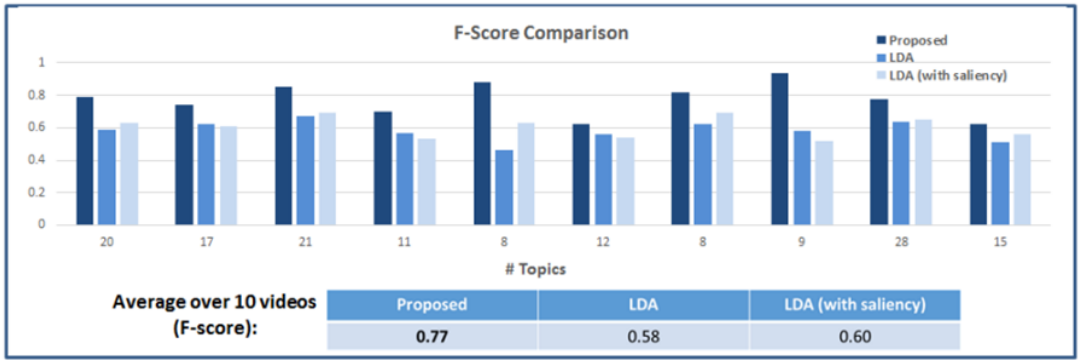
Sl.No	Title	Author	Year	Paper Type
1	Topic Transition in Educational Videos Using Visually Salient Words	Ankit Gandhi et al.	2015	Educational Data Mining
2	Localizing Moments in Video with Natural Language	Lisa Anne Hendricks et al.	2017	ICCV 2017
3	Automatic Hierarchical Table of Contents Generation for Educational Videos	Debabrata Mahapatra et al.	2020	Companion of the The Web Conference 2018
4	End-to-End Learning of Visual Representations from Uncurated Instructional Videos	Miech, Antoine and Alayrac et al.	2020	2020 IEEE/CVF CVPR
5	Unsupervised Audio-Visual Lecture Segmentation	Darshan Singh S et al.	2022	2023 IEEE/CVF WACV

Topic Transition in Educational Videos Using Visually Salient Words

- A visual saliency algorithm is used in order to find the topic transition points in an educational video.[1]
- A saliency score is assigned to each word extracted from the video. (Using Ranking-SVM)
- Using the words along with their saliency scores, the probability that a given slide is a topic transition slide is found.
- The proposed algorithm obtained an F-score of 0.17 higher than that of Latent Dirichlet Allocation based methods.
- Significant improvement in topic navigation using the proposed algorithm is found through demonstration.



Result



Average over 10 videos (F-score):	Proposed	LDA	LDA (with saliency)
	0.77	0.58	0.60

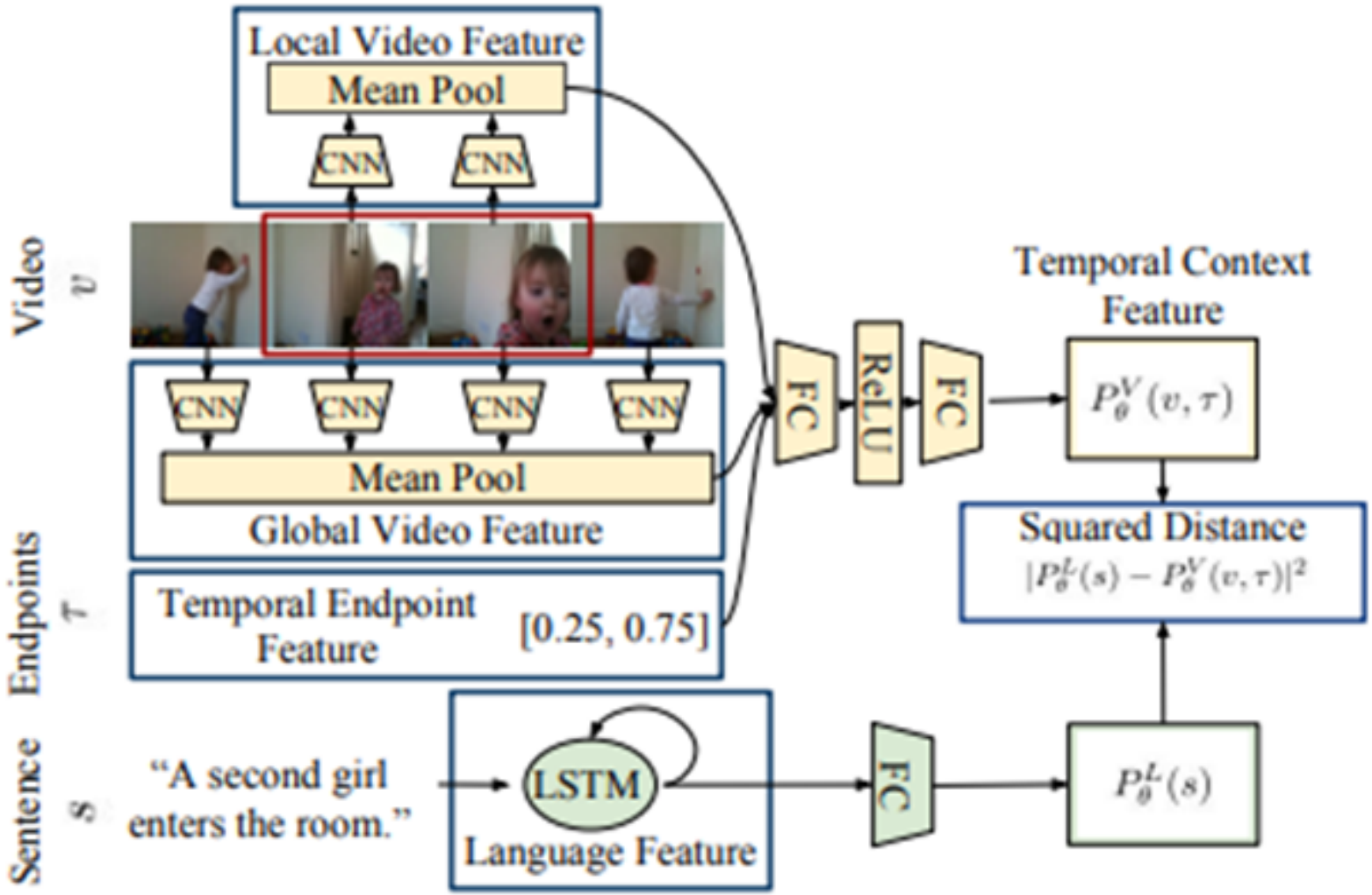
A figure showing the visual saliency scores of words on a sample slide

Comparison of results obtained through the proposed method, LDA and LDA with visual saliency. The proposed method stood out by outperforming LDA by 17%.

Localizing Moments in Video with Natural Language

- The paper discusses on localizing or segmenting certain moments in a full-length video.[2]
- For this they proposed a model known as Moment Context Network (MCN).
- MCN model localizes the input queries in the video by integrating local and global features from the video over time.
- The aim is to retrieve or capture a particular segment from a video using natural language textual inputs by the user.
- The dataset used or created for the project is named Distinct Describable Moments (DiDeMo) dataset which consists of over 40,000 pairs of localized video moments and its natural language. The dataset consists of over 10,000 unedited videos with 3-5 pairs of descriptions and distinct moments per video

Moment Context Network (MCN)



Result

- The MCN model was able to localize a diverse set of moments which includes moments requiring understanding temporal indicators like "first" and moments with camera motion.

Automatic Hierarchical Table of Contents Generation for Educational Videos

- The paper proposes a new method for automatically generating a hierarchical table of contents (ToC) for educational videos.[3]
- The ToC is a summary of the video content, organized in hierarchical structure, that allows users to quickly navigate to the desired section of the video.
- The proposed method consists of three steps:
  - Shot segmentation: The video is first segmented into shots, which are consecutive frames that are visually similar.
  - Text extraction: Textual information is extracted from each shot, such as the speaker's text, slide titles, and captions.
  - ToC generation: A hierarchical ToC is generated using the textual information from each shot.

Result

Method	Precision		Recall		F-score	
	P_m	P_s	R_m	R_s	F_m	F_s
MMToC	0.56	0.3	0.13	0.08	0.21	0.12
HMMToC	**0.74**	0.23	**0.83**	0.17	**0.76**	0.20

Experimental results on titles of ToC

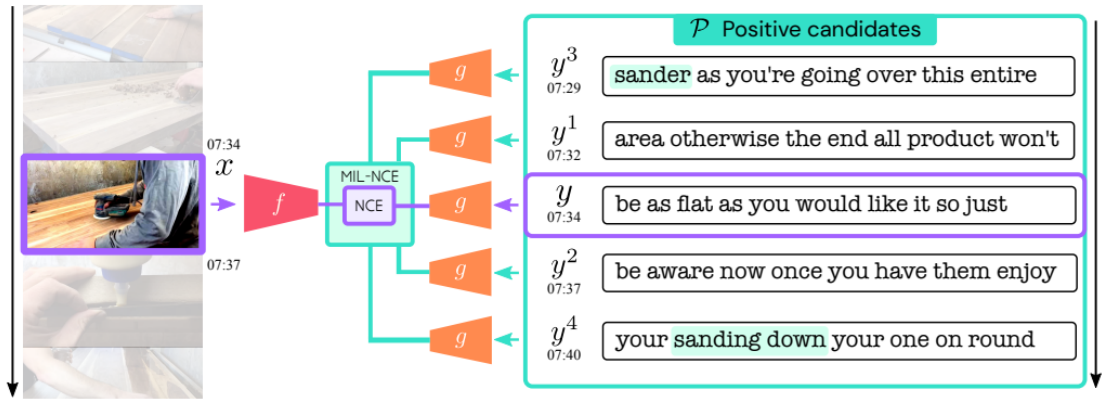
Method	Precision		Recall		F-score	
	P_m	P_s	R_m	R_s	F_m	F_s
MMToC	0.83	0.22	0.21	0.14	0.31	0.14
HMMToC	**0.86**	0.23	**0.92**	0.181	**0.85**	0.21

Experimental results on timings of ToC

- The performance results of HMMToC(Hierarchical Multi-Modal Table of Contents) and MMToC, two competing algorithms, were evaluated on the labeled dataset.
- HMMToC was found to have a higher precision, recall, and F-Score than MMToC for both the timing and title criteria.

End-to-End Learning of Visual Representations from Uncurated Instructional Videos

- In this paper the authors propose a new learning approach, MIL-NCE, which can deal with misalignments in narrated videos.[4]
- This is aimed to make strong video representations, without having to do any manual annotation.
- A 3D CNN backbone is used is used here, specifically the I3D implementation.
- We use the visual representations at two different levels, the first being the output of the I3D Global avg pool and the second being the output of the final I3D layer.

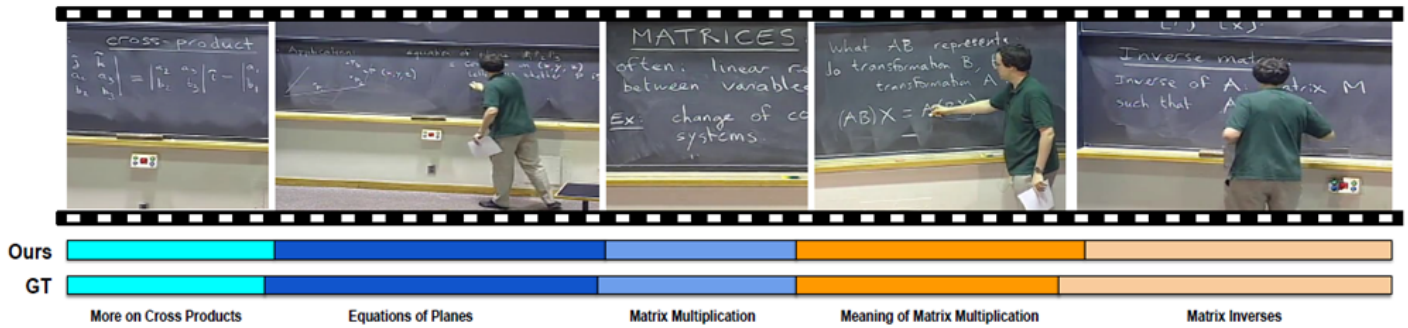


Result and Comparison

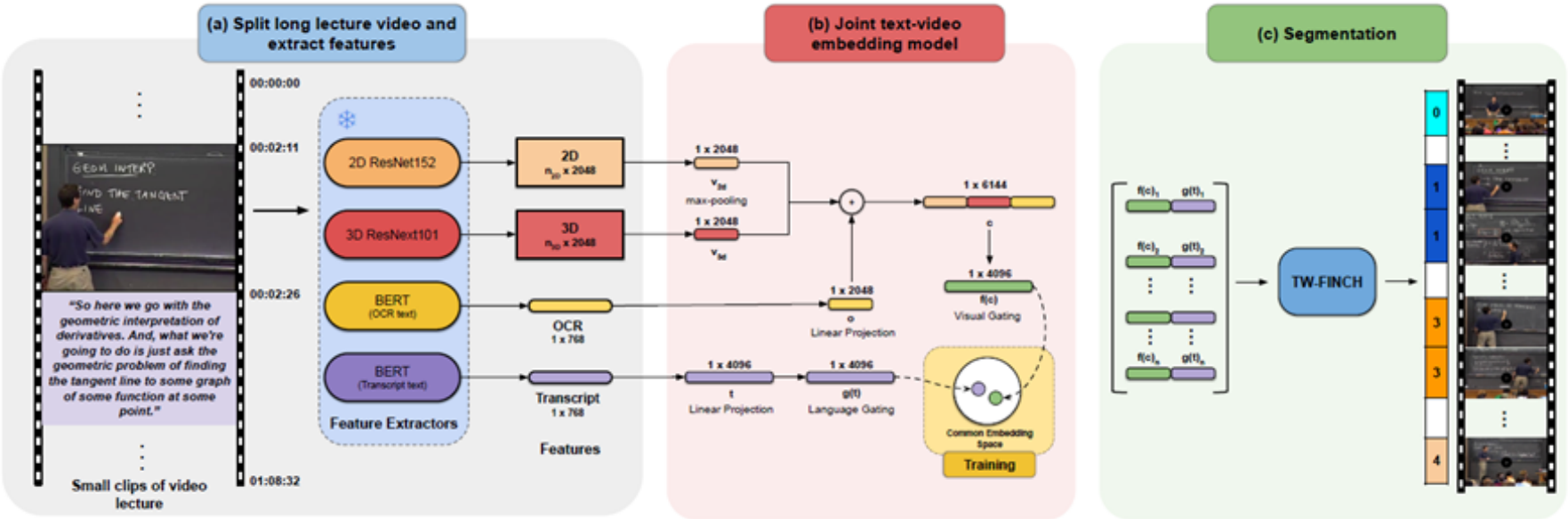
- The model is compared to other self-supervised approaches, which have a commonality with this model itself, that is the fact that it does not use any annotated video or image dataset when training. This approach outperforms prior works in the field without any fine-tuning.
- This results in a significant generalization of the representation to diverse sets of action despite being trained on uncurated instructional videos.

Unsupervised Audio-Visual Lecture Segmentation

- A rise in popularity of streaming videos has resulted in a drastic rise particularly in the watching of educational lecture videos. The major contribution of the authors of this paper was in the form of the AVLectures dataset. [5]
- It is a unique dataset containing over 86 courses with over 2,350 lectures covering various STEM subjects.
- Their secondary contribution comes in the form of a video lecture segmentation that splits a lecture into bite-sized components.



Lecture Segmentation



Result

- In training, the first stage encapsulates the pre-training of the model using the CwoS dataset. The second stage includes pre-training with the CwS dataset in an unsupervised manner
- In evaluation dataset all 15 courses of CwS to understand and derive the performance of the model. To measure the overall accuracy of the model metrics such as NMI (Normalized Mutual Information), MoF (Mean Over Frames), F1-score, IoU (Intersection Over Union). The paper has also highlighted the comparison between different segmentations such as Content-Aware Detector. Text Tiling, LDA (Latent Dirichlet Allocation), etc.
- The result of this Unsupervised Lecture Segmentation is the visualization of segmentation outputs for three video lectures from different courses. While some phases of detection were correct or accurate, others were not.

Block Diagram

Timeline

- **Mini Project**
  - August 2023 : Topic finalization and literature survey
  - September 2023 and October 2023 : : Implementation of automatic video segmentation model using audio data
  - November 2023 : Testing the database using deep learning approaches.
- **Project**
  - Implement the model such that both audio/visual data can be used in segmentation
  - Validate and enhance the accuracy and performance of the model.

References

[1] Ankit Gandhi, Arijit Biswas, and Om Deshmukh. "Topic Transition in Educational Videos Using Visually Salient Words". In: *Educational Data Mining*. 2015. URL: https://api.semanticscholar.org/CorpusID:14721646.

[2] Lisa Anne Hendricks et al. "Localizing Moments in Video with Natural Language". In: 2017. arXiv: 1708.01641 [cs.CV].

[3] Debabrata Mahapatra, Ragunathan Mariappan, and Vaibhav Rajan. "Automatic Hierarchical Table of Contents Generation for Educational Videos". In: *Companion Proceedings of the The Web Conference 2018*. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee, 2018. ISBN: 9781450356404. DOI: 10.1145/3184558.3186336. URL: https://doi.org/10.1145/3184558.3186336.

[4] Antoine Miech et al. "End-to-End Learning of Visual Representations from Uncurated Instructional Videos". In: *CVPR*. 2020.

[5] Darshan Singh S et al. *Unsupervised Audio-Visual Lecture Segmentation*. 2022. arXiv: 2210.16644 [cs.CV].

Thank You !!!