Automated Video Segmentation

Arjun Sadashiv
AM.EN.U4EAC20012
Electronics and Computer Engineering
Dept.of Electronics and Communications
Amrita Viswa Vidhyapeetham
Amritapuri Campus

Arun Gangadharan
AM.EN.U4EAC20013
Electronics and Computer Engineering
Dept.of Electronics and Communications
Amrita Viswa Vidhyapeetham
Amritapuri Campus

Jibin Thomas Daniel
AM.EN.U4EAC20030
Electronics and Computer Engineering
Dept.of Electronics and Communications
Amrita Viswa Vidhyapeetham
Amritapuri Campus

Sanat Jayakrishnan
AM.EN.U4EAC20058
Electronics and Computer Engineering
Dept.of Electronics and Communications
Amrita Viswa Vidhyapeetham
Amritapuri Campus

***Abstract*— Online lecture videos have become an overwhelming juggernaut in the online content creation spheres of popular video streaming sites. Especially after the recent Covid-19 pandemic that shifted the physical classroom dynamic to that of a virtual one. Without the organic interaction between teacher and student that would have occurred in a physical classroom setting, many a learning experience was stifled, and information could not be conveyed efficiently. This led to a large number of students relying on the Internet to learn and a large number of them flocked to the video lectures that were available on streaming platforms like MITOCW, Coursera, NPTEL, and YouTube to name a few. Many of these lecture videos are too lengthy and might cover multiple concepts. So, it is imperative to segment the videos for increased adoption among the student community. The main objective of our project is to create an automatic 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. To the best of our knowledge, no work has been explored using the audio from the video lectures for this task. 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).**

I. Introduction

In a regular household, with the average family, education of the children tends to be of paramount importance. In today's day and age, where electronics and the internet run rampant in every household, education and access to information have become incredibly easy. One such method of extracting information is by simply watching educational videos in which a lecturer explains a concept and understanding it. This has become the lifeline of many a student in the far-reaches of the world. But in today's busy world, ease of use and convenience also tend to play a major role in the way students tend to curate the video lectures they require or select to parse through. Convenience usually translates how quickly one can go through one such video and the segmentation of a lecture usually helps a student in increasing their ease of use. But segmentation of video lectures can only occur if the streaming platform the video is posted on permits and if the creator of the video does segment their lecture. Now this is not the case most of the time, video lectures do not come pre-segmented.

This is an inconvenience we aim to resolve using our project. By understanding and reviewing similar works done in the same field we seek to create a model which automatically segments a video based on the topics covered within the lecture and its correlating audio.

In this review, we highlight some papers and summarise the contributions made by them as they all will help us in our journey to build an audio-analysing model.

II. Topic Transition in Educational Videos Using Visually Salient Words

The rapid growth of online courses and Open Educational Resources (OER) has led to a need for methods to efficiently consume this multimedia content. One of the most challenging areas of research is to automatically identify topic transitions in educational videos. Topic transitions are the points in a video where the topic changes. For example, in a lecture video on types of discrete random variables, the topics can vary from Binomial RV, Geometric RV to name a few.

Manually identifying topic transitions is a time-consuming and expensive process. A new approach to automatically identify topic transitions in educational videos is proposed in this paper. The approach analyses the visual content of a lecture video to determine the transition points.

A. Dataset Used

10 NPTEL (National Program of Technology Enhanced Learning) educational videos with a duration of about 1-1.5 hours each, with a total of 12 hours of video content was selected. These kinds of videos have a large amount of diversity in lighting conditions, slide orientations, lecturer positions, etc. Handwritten text, printed text, along with writing on slides are different scenarios that challenge word recognition which are found throughout the dataset.

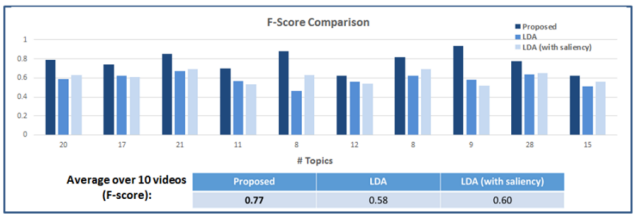
B. Methodology

The proposed approach to automatically identify topic transitions in educational videos uses a combination of visual saliency and machine learning techniques. The visual saliency technique is used to identify the visually salient words in a video. These words are then used as features to

train a machine learning model to identify the topic transitions.
The visual saliency technique works by first extracting the text from the frames of a video. The text is then processed to remove stop words and punctuation. The remaining words are then ranked according to their visual saliency. The words with the highest visual saliency are considered to be the most important words in the video.
The machine learning model is a support vector machine (SVM). The SVM is trained on a dataset of videos that have been manually annotated with their topic transitions. The SVM learns to predict the topic transitions in a new video based on the visual saliency of the words in the video.

C. *Result*

The videos were manually annotated with their topic transitions. The proposed approach was able to identify the topic transitions with an accuracy of 85%. The proposed approach is a novel and effective method for automatically identifying topic transitions in educational videos. The approach is based on a combination of visual saliency and machine learning techniques. The approach is accurate and efficient, and it can be easily applied to new videos.



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

III. Localizing Moments in Video with Natural Language

The paper discusses on localizing or segmenting certain moments in a full-length video. The aim is to retrieve or capture a particular segment from a video using natural language textual inputs by the user. 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.

A. *Dataset Used*

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. DiDeMo is collected in an open world setting and consist of video clips from the very diverse contents from our surrounding.

B. *Methodology*

The moment retrieval algorithm localizes the queries in longer videos. The visual temporal context features are extracted which encode the moment in the video by integrating the local and the global features of the video. The language features are extracted using LSTM techniques.

Visual Temporal Context Features: the video moments are encoded into visual temporal context features by integrating local video features, global video features and temporal endpoint features.

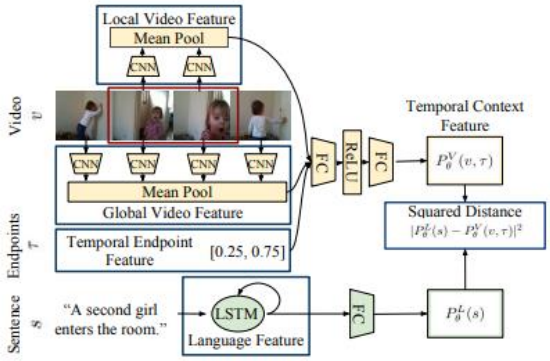


Figure: The Moment Context Network(MCN) is trained to learn the shared embedding for the temporal context features of a given video. The video temporal context features integrate the local and global video features. In this model both the appearance and the optical flow input modalities are considered.

The language features are extracted using LSTM techniques. After the encoding the LSTM hidden state is passed through a single connected layer to yield embedded features P_θ^L which is passed to a Joint Video and Language Model. The joint model is the sum of the squared distance between the embedded appearance, flow and language features.

$$D_\Theta(s,v,t) = \left|P_{\Theta^V}(v,t) - P_{\Theta^L}(s)\right|^2 + \eta\left|P_{\Theta^F}(f,t) - P_{\Theta^L}(s)\right|^2$$

Ranking Loss for Moment Retrieval: The model is trained with a ranking loss which compares the closeness of the referring expressions with its corresponding moments from the video than negative moments in a shared embedding space.

$$L_i^{inter}(\Theta) = \sum_{j\neq i}^{c} L\left(D_\Theta(s^i,v^i,t^i), D_\Theta(s^i,v^i,t^i)\right)$$

$$\mathcal{L}(\Theta) = \lambda\sum^c L_{i\,intra}(\Theta) + (1-\lambda)\sum_i^r \mathcal{L}_i^{inter}(\Theta)$$

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.

IV. 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. The ToC is a summary of the video content, organized in a hierarchical structure, that allows users to quickly navigate to the desired section of the video.

A. *Dataset Used*

A labeled dataset of 46 educational videos from YouTube was created to evaluate the quality of Table of Contents (ToC) entries. Each video was either a part of an online course or a webinar related to computer science.

B. *Methodology*

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.

A novel method for shot boundary detection is proposed. The method consists of three steps: visual content representation, continuous signal construction, and classification of shot boundaries. In the visual content representation step, the shape of objects in an image is captured by computing the entropy of the edge map in a patch-wise manner. Color variations are captured by representing the image with the 2D histogram of hue and saturation values. In the continuous signal construction step, a measure is computed to capture the resemblance between two frames. This measure is a product of the similarity metrics for color and shape features. In the classification of shot boundaries step, a clustering-based technique is used to choose an adaptive threshold for classifying whether a boundary is an shot boundary or not.

Visual text extraction is performed using a commercial OCR engine and a deep learning model to generate a heat map of text regions in an image. The text regions are then filtered based on their overlap with the bounding boxes from the OCR engine. The salient features of each visual text entity (vText) are extracted, including font size, boldness, and vertical location. All the vTexts in a shot are aggregated into a single set to avoid processing repeated texts. The final ToC is generated by combining the visual and audio text information.

C. *Result*

• Timing: If the ToC entry output by an algorithm falls within a 10-second interval of the labeled entry, it is considered a hit. Otherwise, it is a mis-hit.

• Title: A ToC entry obtained from an algorithm is considered a hit if at least one of its words matches the ground truth ToC and if there is a hit with respect to timing as well.

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(Multi-Modal Table of Contents), two competing algorithms, were evaluated based on the title and timing of the ToC 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.

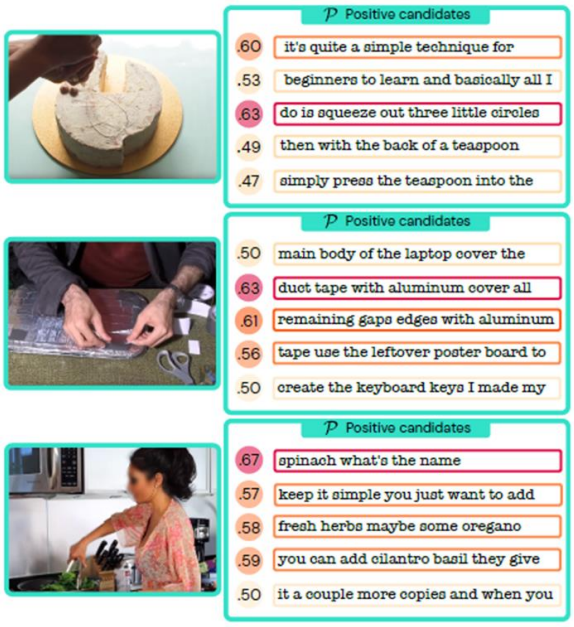
V. End-to-End Learning of Visual Representations from Uncurated Instructional Videos

Vision and language play an irreplaceable role in the way a human learn and grow. Learning by associating with a visual entity often leads to retention of that specific thing being much stronger. In today's day and age, narrated educational videos are have become a norm. Annotating such a video has become a non-scalable, difficult, and overly expensive task. The HowTo100M dataset is utilized here, this dataset enables the learning of video representations without manual supervision.

A. *Dataset Used*

The dataset utilized here for this particular model is the HowTo100M dataset which contains more than 100 million pairs of video clips and their associated narrations. It was curated and collected by searching YouTube for educational or instructional videos depicting a complex human activities.

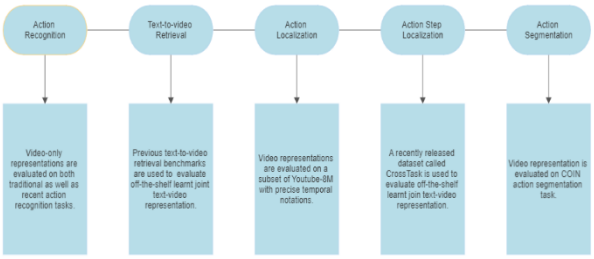
B. *Methodology*



Narration pairs taken from five different positive candidates

In this paper the authors propose a new learning approach, MIL-NCE, which can deal with misalignments in narrated videos. This is aimed to make strong video representations, without having to do any manual annotation. A 3D CNN backbone is used here, specifically the I3D implementation. The visual representations are used 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. To depict the all-roundedness of the representations, an evaluation is performed on five downstream tasks:

- Action Recognition
- Text-to-video Retrieval
- Action Localization
- Action Step Localization
- Action Segmentation



Flowchart depiction and descriptions of the downstream tasks

C. *Result*

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. On comparison to a randomly initialized I3D and an inflated I3D, a 4% improvement is obtained over the randomly initialized I3D and 1:4% over the inflated I3D.

VI. 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. It is a unique dataset containing over 86 courses with over 2,350 lectures covering various STEM subjects. Every single course contains video lectures, transcripts, OCR outputs for lecture frames, and lecture notes, slides, assignments and even related educational content. Their secondary contribution comes in the form of a video lecture segmentation that splits a lecture into bite-sized components. The clip representations use visual, audio and OCR cues and the model is trained on a pretext self- supervised task of matching the narration with temporally aligned visual content. TW- FINCH, a variation of the 1- Nearest Neighbor algorithms is a utilized on these representations. The unique position this model occupies in the sphere of research relies solely on the unique model that has been developed and trained for specific purpose.

A. *Datasets Used*

The AVLectures is primarily derived from MIT-OCW. A curated list of courses was formulated by browsing the OCW website. Non-lecture videos were removed and every frame in each lecture was processed and stored using Google Cloud API. Curation of the dataset as a whole was done by separating the dataset into two, CwS (Courses with Segmentation) and CwoS (Courses without Segmentation. The segmentation timestamps is obtained by webscraping.

B. *Methodology*

The approach highlighted in this paper has been split up into three different stages:

- Video Clip Feature Extraction:
  The three key feature types used here are OCR, 2D and 3D. The OCR feature encodes the output text from an OCR API using the BERT sentence transformer model. The Text Feature Extraction uses the same model as the one for OCR.
- Learning joint text-video:
  The model is built by first extract the visual and textual features for a video clip C and transcript (text) T using the feature extraction pipelines described above. The OCR feature is then passed through a fully-connected layer to obtain a 2048-dimensional vector. The model's parameters is trained train with the max-margin ranking loss function.
- Lecture segmentation with learning embeddings:
  The clip is extracted and transcript embeddings from the joint text-video model. All such

representations of a lecture with N clips are passed to the TW-FINCH algorithm, this is then represented as a 1-nearest-neighbour graph after encoding is done on feature similarity and temporal proximity.

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.

C. *Result*

Utilizing a model to improve the representations is clearly better (NMI 73.0 vs. 71.8), it is observed that a model pre-trained on AVLectures (rows 3-5) outperforms a model pre-trained on HowTo100M (rows 1-2) consistently. 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.

VII. Conclusion

The different papers selected and summarized in this review, all serve as stepping stones to develop our own model that will automatically segment different topics covered in a lecture by its audio cues. The different methodologies covered in the various papers highlighted in this review such as the end to learning of visual representation or an automatic generation of a table of contents based on hierarchy, all help in the understanding of the progression of work done in this field and thereby aiding us in developing our audio-based model. The unsupervised lecture segmentation and the unique dataset that comes along with it are immensely appreciated as it provides the strong foundation we were seeking, in the form of a curated segmented and non-segmented dataset. Lastly, we believe that just as the selected papers in this review have aided, our future works may aid those who decide to build upon our work and improve or create entirely new models.

VIII. 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".
https://openaccess.thecvf.com/ICCV2017

[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: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020.
https://openaccess.thecvf.com/content_CVPR_2020/html/Miech_End-to-End_Learning_of_Visual_Representations_From_Uncurated_Instructional_Videos_CVPR_2020_paper.html

[5] Darshan Singh S et al. Unsupervised Audio-Visual Lecture Segmentation.2022.
https://ieeexplore.ieee.org/xpl/conhome/10030081/proceeding