**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.

**Dataset**

* 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.

**Method**

The proposed method consists of three steps:

1. Shot segmentation: The video is first segmented into shots, which are consecutive frames that are visually similar.
2. Text extraction: Textual information is extracted from each shot, such as the speaker's text, slide titles, and captions.
3. ToC generation: A hierarchical ToC is generated using the textual information from each shot.

Shot boundaries (SB) detection

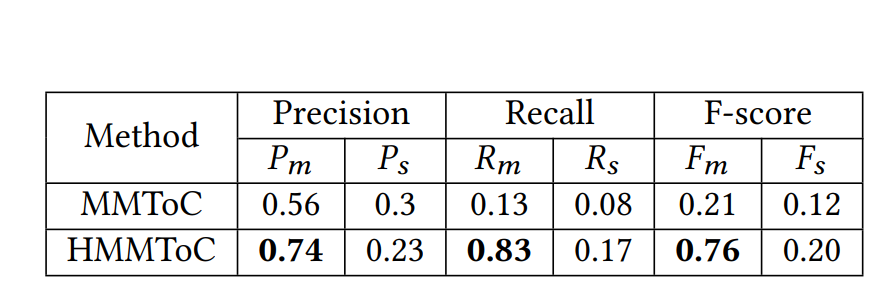
* A novel method for shot boundary detection is proposed.
* The method consists of three steps: visual content representation, continuous signal construction, and classification of shot boundaries.
* In the visual content representation step, the shape of objects in an image is captured by computing 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 SB or not.

Text extraction

* 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.

**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.



**Experimental results on titles of ToC**

A table with numbers and symbols

Description automatically generated

**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.