Automated Whiteboard Lecture Video Summarization by Content Region Detection and Representation

Bhargava Urala Kota, Alexander Stone, Kenny Davila, Srirangaraj Setlur, Venu Govindaraju

University at Buffalo, State University of New York, USA

ICPR 2020
25th INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION
Milan, Italy 10 | 15 January 2021

Abstract

Lecture videos are a useful resource for students and educators across the world. Our goal is to summarize lecture videos by extracting regions of text content from every frame, extracting features from text regions to represent content in each frame. Finally, summaries of videos can be produced by comparing text features across frames locally (within a temporal window) and globally (across the entire video) to obtain a smaller subset of frames (called *keyframes*) which contain all of the text in the video. Extracted frames and text features will facilitate content-based lecture video search.

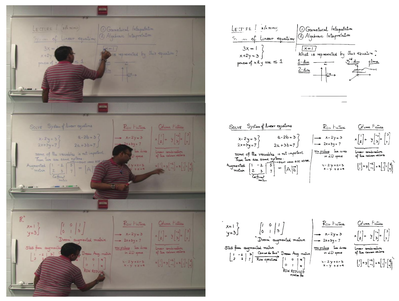

Figure 1: Example lecture video frames with corresponding generated summary[2] keyframes on the right

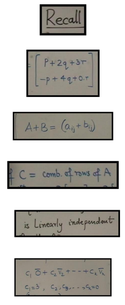

Figure 2: Examples of different kinds of text content in lectures

Challenges

- Current search engines primarily support meta-data based search and retrieval of lecture videos, effective *video summarization* techniques are needed to extract key content and condense this data into an easily searchable form.
- Lecture content is often *loosely structured* and exhibits large variances in semantic grouping. Further, background noise, illumination changes and occlusions are also present.
- Lectures could have typeset text on slides; handwritten content on white/chalkboards or digitally rendered, which adds to the challenge for extraction and feature representation of the content.
- Techniques to learn *vector embedding* for words are known. They must be extended *for structured text* such as math expressions needing extensive transcription annotations.

Methodology

- Overview: We extract localization information and representative features from content regions in every frame (using a detector and feature extraction neural network). We then form tracklets based on feature similarity and geometric displacement. We use partial region features to detect growing, occluded content and refine tracklet information. Tracklet information is then used to identify the right video frame to segment leading to keyframe summaries. Unique content regions are presented as key object summaries.

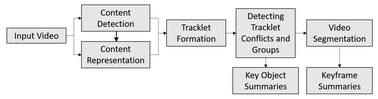

Figure 3: Overview of our lecture video summarization pipeline

- Content Detection: We finetune PSENet on whiteboard lecture video data. PSENet is chosen because i) it allows to detect text of arbitrary shapes ii) it is able to disambiguate closely grouped text
- Content Representation: We use the backbone layers of the detector network and add additional layers to extract meaningful content representation trained under Multi-Similarity loss. We also try low-level image descriptors such as Histogram of Gradients.

Figure 4: Architecture of our text feature extraction network

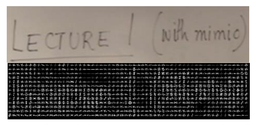

Figure 5: Low-level feature extraction for fine-grained feature description of content region

- Video Segmentation: Detected regions and features can be used to build tracklets which describe lifetime and spatial location of content. This information is used for video segmentation. Tracklets need to be robust to occlusion, writing and erasing events to obtain accurate segmentation.

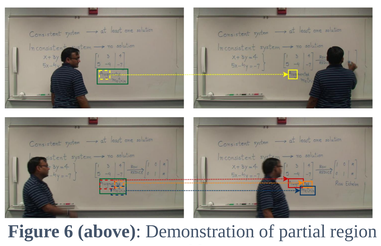

Figure 6 (above): Demonstration of partial region tracklets

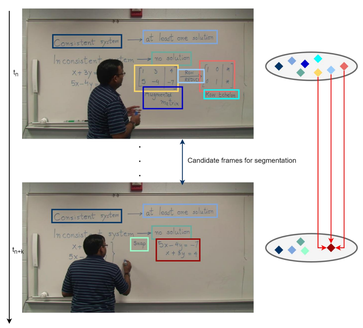

Figure 7 (right): Demonstration of segmentation by determining and resolving conflicting region tracklets. Red arrows indicate conflict in the tracklet graph

Conclusions and Results

- Higher level descriptors learnt using neural network fall short of HoG in terms of performance metrics possibly due to lack of training data and pooling layers
- Our partial region feature method and weighted conflict detection method provides state-of-the-art compression (in terms of number of summary frames) with near s.o.t.a. content recall and precision.

Method	AVG N_f (σ)	AVG Global R	AVG Global P	AVG Global F	AVG Per Frame R	AVG Per Frame P	AVG Per Frame F
AccessMath [7]	17.29 (4.54)	96.28	93.56	94.90	95.73	92.21	93.93
Maximum Content Sum [15]	34.42 (10.15)	**96.49**	94.51	**95.49**	**96.13**	91.95	93.99
Prior work 1 [8]	19.43 (5.32)	92.33	94.16	93.23	91.69	93.45	92.56
Prior work 2 [10]	21 (5.17)	95.80	92.88	94.32	95.40	92.44	93.90
Xu et al. [17]	12.29 (2.14)	95.89	86.28	90.83	94.18	85.15	89.44
Combined Area-Feature (best compression)	**10.29** (1.67)	95.01	93.84	94.41	93.98	93.69	93.84
Combined Area-Feature (best f-measure)	15.57 (1.99)	95.61	**94.56**	95.08	94.60	**93.77**	**94.17**

Future Work

- Robust representation of structured text – exploring graph vectorization techniques as a possibility.
- Apply our framework to presentation videos, multi-camera recordings with camera motion, zoom, pan etc.
- User studies focusing on retrieval of queried content from video summaries.
- Exploration of different types of summaries – skims, key topics, composite frames, transcription-based to enable search/retrieval.

References

[7] K. Davila and R. Zanibbi, “Whiteboard video summarization via spatio-temporal conflict minimization,” in International Conference on Document Analysis and Recognition (ICDAR), 2017.

[8] B. Urala Kota, K. Davila, A. Stone, S. Setlur, and V. Govindaraju, "Automated Detection of Handwritten Whiteboard Content in Lecture Videos for Summarization", 2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR), pp 19-24, 2018.

[10] B. Urala Kota, K. Davila, A. W. Stone, S. Setlur, V. Govindaraju, "Generalized framework for summarization of fixed-camera lecture videos by detecting and binarizing handwritten content." International Journal on Document Analysis and Recognition (IJDAR) (2019): 1-13.

[15] C. Choudary and T. Liu, “Summarization of visual content in instructional videos,” IEEE Transactions on Multimedia, vol. 9, no. 7, pp. 1443–1455, 2007.

[17] F. Xu, K. Davila, S. Setlur, and V. Govindaraju, “Content extraction from lecture video via speaker action classification based on pose information,” in 2019 International Conference on Document Analysis and Recognition (ICDAR). IEEE, 2019, pp. 1047–1054.