Classification of Important Segments in Educational Videos using Multimodal Features

Junaid Ahmed Ghauri[a], Sherzod Hakimov[a] and Ralph Ewerth[a,b]

[a] TIB – Leibniz Information Centre for Science and Technology, Hannover, Germany
[b] L3S Research Center, Leibniz University Hannover, Germany

Abstract

Videos are a commonly-used type of content in learning during Web search. Many e-learning platforms provide quality content, but sometimes educational videos are long and cover many topics. Humans are good in extracting important sections from videos, but it remains a significant challenge for computers. In this paper, we address the problem of assigning importance scores to video segments, that is how much information they contain with respect to the overall topic of an educational video. We present an annotation tool and a new dataset of annotated educational videos collected from popular online learning platforms. Moreover, we propose a multimodal neural architecture that utilizes state-of-the-art audio, visual and textual features. Our experiments investigate the impact of visual and temporal information, as well as the combination of multimodal features on importance prediction.

Keywords

educational videos, importance prediction, video analysis, video summarization, MOOC, deep learning, e-learning

1. Introduction

In the era of e-learning, videos are one of the most important medium to convey information for learners, being also intensively used during informal learning on the Web [1, 2]. Many academic institutions started to host their educational content with recordings while various platforms like Massive Open Online Courses (MOOC) have emerged where a large part of the available educational content consists of videos. Such educational videos on MOOC platforms are also exploited in search as learning scenarios, their potential advantages compared with informal Web search have been investigated by Moraes et al. [3]. Although many platforms pay a lot of attention to the quality of the video content, the length of videos is not always considered as a major factor. Many academic institutions provide content where the whole lecture is recorded without any breaks. Such lengthy content can be difficult for learners to follow in distant learning. As mentioned by Guo et al. [4] shorter videos are more engaging in contrast to pre-recorded classroom lectures split into smaller pieces for MOOC. Moreover, pre-planned educational videos, talking head, illustrations using hand drawings on board or table, and speech tempo are other key factors for engagement in a video lecture as described by Zolotykhin and Mashkina [5].

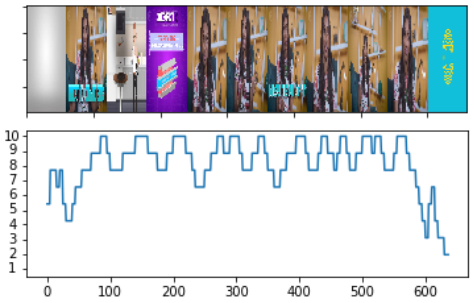

Figure 1: Sample video with annotations of importance scores for each segment

In this paper, we introduce computational models that predict the importance of segments in (lengthy) videos. Our model architectures incorporate visual, audio, and text (transcription of audio) information to predict importance scores for each segment of an educational video. A sample video and its importance scores are shown in Figure 1. A value between 1 and 10 is assigned to each segment indicating the score of a specific segment whether it refers to an important information regarding the overall topic of a video. We refer to it as the *importance score* of video segments in educational domain, similar to the annotations provided by TVSum dataset [6] on various Web videos. We have developed an annotation tool that allows annotators to assign importance scores to video segments and created a new dataset for this task (see

Proceedings of the CIKM 2020 Workshops, October 19–20, Galway, Ireland
EMAIL: junaid.ghauri@tib.eu (J.A. Ghauri); sherzod.hakimov@tib.eu (S. Hakimov); ralph.ewerth@tib.eu (R. Ewerth)
ORCID: 0000-0001-9248-5444 (J.A. Ghauri); 0000-0002-7421-6213 (S. Hakimov); 0000-0003-0918-6297 (R. Ewerth)

Section 4). The contributions of this paper are summarized as follows:

- Video annotation tool and an annotated dataset
- Analysis of influence of multimodal features and parameters (history window) for educational video summarization
- Multimodal neural architectures for the prediction of importance scores for video segments
- The source code for defined the deep learning models, the annotation tool and the newly created dataset are shared publicly[1] with the research community.

The remaining sections of the paper are organized as follows. Section 2 presents an overview of related work in video-based e-learning and computational architectures covering multiple modalities in educational domain. In Section 3, we provide detailed description of model architectures. Section 4 presents the described annotation tool and the created dataset. Section 5 covers the experimental results and discussions on the findings of the paper and Section 6 concludes the papers.

2. Related work

Various studies have been conducted that address the quality of online education, create personalized recommendations for learners, or focus on highlighting the most important parts in lecture videos. Student interaction with lecture videos offers new opportunities to understand the performance of students or for the analysis of their learning progress. Recently, Mubarak et al. [7] proposed an architecture that uses features from e-learning platforms such as watch time, plays, pauses, forward and backward to train deep learning models for predictive learning analytics. In a similar way, Shukor and Abdullah [8] used watch time, clicks, completed number of assignments for the same purpose. Another method by Tang et al. [9] is a concept-map based approach that analyzes the transcripts of videos collected from YouTube and visual recommendations to improve learning path and provide personalized content. In order to improve student performance and enhance the learning paradigm, high-tech devices are recommended for the classroom setting and content presentation. For instance, instructors or presenters can highlight important sections which can be saved along with the video data and later be used by students when they are going through the video lectures.

Research in the field of video summarization addresses a similar problem, where important and relevant content from videos is classified to generate summaries (for instance, [10, 11] and [12]). All of these methods are based on TVSum [6] and SumMe [13] datasets that consist of Web videos. The nature of these datasets is very different to videos from the educational domain. These datasets can be a good source of visual features but spoken words or textual content are relatively rare or not present at all. Inspired from video summarization work, Davila and Zanibbi [14] presented a method to detect written content in videos, e.g. on whiteboards. This research focuses on a sub-task which only takes into account the lectures in which the written content is available, and also addresses only the topic of mathematics. Xu et al. [15] focused on another kind of technique where speaker pose information can help in action classification like writing, explaining, or erasing. Here, the most important segments are *explaining*, which could be an indication of an important segment in educational videos.

Another important aspect of e-learning is student engagement for different types of online resources. Guo et al. [4] analyzed various aspects for MOOC videos and provided a number of related recommendations. Shi et al. [16] analyzed the correlation of features and lecture quality by considering visual features from slides, linguistic elements and audio features like energy, frequency, pitch, etc. to highlight important and emphasized statements in a lecture video. As suggested by YukiIchimura [17], one of the best practices in MOOCs is to offer information on which parts of a lecture video are difficult or need more attention, which could potentially lead to a more flexible and personalized learning experience. In order to perform such tasks by machines, they need to incorporate multimodal information from educational content. To deal with multimodal data is not easy and this is also true for multimodal learning, as explained by Wang et al. [18]. If user interaction data are available for videos along with visual, textual information, then the task can be solved by multimodal deep learning models.

3. Multimodal Architecture

In this section, we describe the proposed model architecture that predicts importance scores for each video segment by fusing audio, visual and textual features. Each video contains audio, visual and textual (subtitles) content in the three different modalities. To join

[1] https://github.com/VideoAnalysis/EDUVSUM

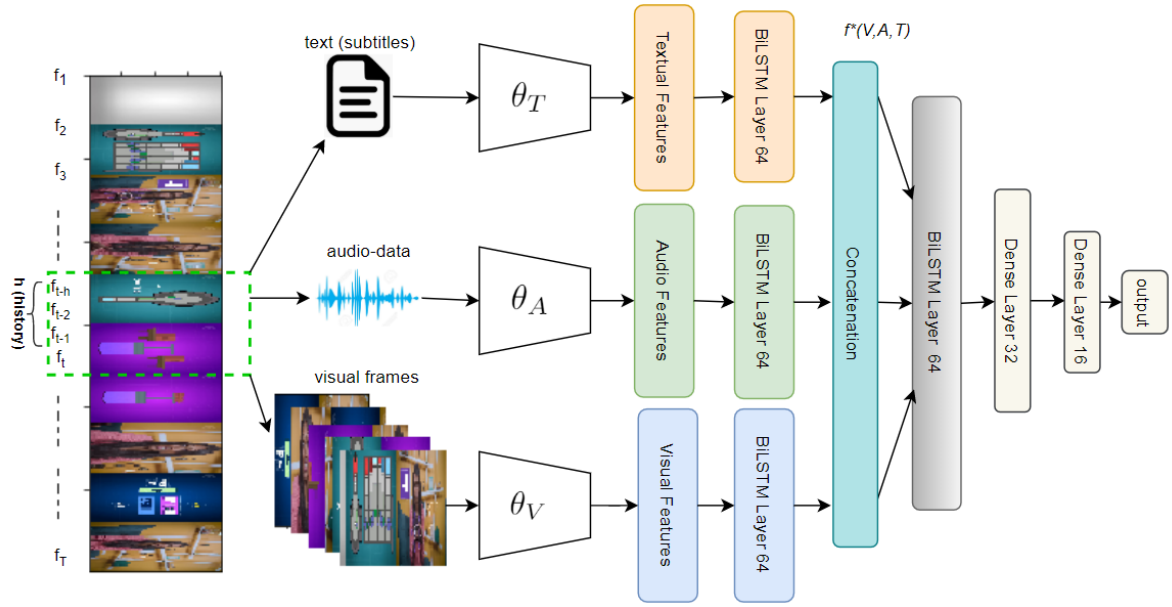

Figure 2: Multimodal architecture for classification of important segments in educational videos

different modalities we adapt and extend ideas from Majumder et al. [19], who apply fusion to three kinds of modalities available in videos: visual, audio, and text. The overall architecture is depicted in Fig. 2. In order to deal with the temporal aspect of videos, we use Bidirectional Long Short-term Memory (BiLSTM) layers to incorporate information from each modality [7, 12, 20]. We use state-of-the-art pre-trained models to encode each modality in order to extract features. After the extraction of feature embeddings for each modality, they are fed into separate *BiLSTM* layers. The outputs of these layers are then concatenated in a time-oriented way and then fed into another *BiLSTM* layer, which has 64 units. The output is fed into two dense layers with size of 32 and 16, respectively. Lastly, the output from the last dense layer is fed into a softmax layer that outputs a 10-dimensional vector indicating the importance score of a given input video frame belonging to a certain segment. In addition to the current frame, the model also includes history information that consists of n previous frames according to the setting of history window size parameter. Our experimental results show different configurations and corresponding results, where we evaluate different history windows sizes. Next, we describe the feature embeddings for each modality and the corresponding models to extract them.

Textual Features: The textual content is based on subtitles provided for each video. The text features are extracted by encoding words in subtitles using BERT (Bidirectional Encoder Representations from Transformers) [21] embeddings. *BERT* is a pre-trained transformer (denoted as θ_T) that takes the sentence context into account in order to assign a dense vector representation to each word in a sentence. The textual features are 768-dimensional vectors that are extracted by encoding subtitles of videos. Later, these features are passed to a layer with 64 *BiLSTM* cells.

Audio Features: The audio content is utilized by means of various features that represent the zero crossing rate, energy, entropy or energy, spectral features (centroid, spread, flux, roll-off) and others. In total, there are 34× n_a features, where n_a depends on the window size and step size which are 0.05 and 0.025 % of the audio track length in a video. The combination of the rate of change of all these features yields a total number of 68 features. We use *pyAudioAnalysis* [22] toolkit (denoted as θ_A) to extract these features. These features are fed into a layer with 64 BiLSTM units. We keep the same number of units in the BiLSTM layer of all modalities.

Visual Features: We explored different visual models like Xception [23], ResNet-50 [24], VGG-16 [25] and Inception-v3 [26] pre-trained on ImageNet dataset. Visual content of the videos is encoded using one of the visual descriptors mentioned above, denoted as θ_V.

Our ablation study in Section 5 provides further details on the importance of choice of visual descriptors. Once the features are extracted, they are fed into a *BiLSTM* layer with a size of 64.

Consider a video input of T sampled frames, i.e., $V = (f_t)_{t=1,\ldots,T}$, f_t is the visual frame at point in time t. The variable T depends on the number of selected frames per second in a video. The original frame rate is 30 per second (fps) for a video. The input video is split into uniform segments of 5 seconds from which we select 3 frames per second as a sampling rate. The input of the model are the current frame (f_t) at time step t and the preceding frames ($f_{t-1}, f_{t-2}, \ldots, f_{t-h}$) according to the selected history window size h. The features from a modality are extracted as defined above and passed to the respective layers. The model outputs an importance score for the given input frame (f_t).

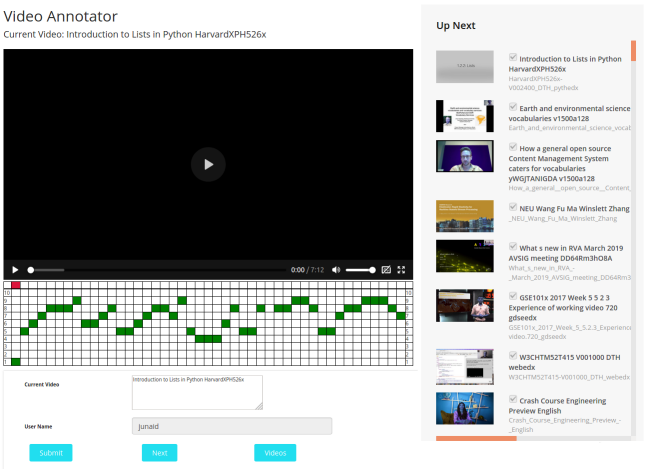

Figure 3: Screenshot of the Web-based annotation tool for labeling video segments

4. Dataset and Annotation Tool

We present a Web-based tool to annotate video data for various tasks. Each annotator is required to provide a value between 1 and 10 for every 5 second segment of a video. A sample screenshot of the annotation tool is shown in Figure 3. The higher values indicate the higher importance of that specific segment in terms of information it includes related to a topic of a video.

We present a new dataset called EDUVSUM (Educational Video Summarization) to train video summarization methods for the educational domain. We have collected educational videos with subtitles from three popular e-learning platforms: Edx, YouTube, and TIB AV-Portal[2] that cover the following topics with their corresponding number of videos: computer science and software engineering (18), python and Web programming (18), machine learning and computer vision (18), crash course on history of science and engineering (23), and Internet of things (IoT) (21). In total, the current version of the dataset contains 98 videos with ground truth values annotated by the main author who has an academic background in computer science. In the future, we plan to provide annotation instructions and guidance via tutorials on how to use the software for human annotators.

5. Experimental Results

In this section, we describe the experimental configurations and the obtained results. We use our newly created dataset consisting of 98 videos for the experimental evaluation of model architectures. The dataset is randomly shuffled before dividing it into disjoint train and test splits using 84.7% (83 videos) and 15.3% (15 videos), respectively. The videos are equally distributed among the topics of the dataset. The dataset splits and frame sampling strategy are compliant with previous work in the field of video summarization (Zhang et al. [10], Gygli et al. [13] and Song et al. [6]).

We evaluated different configurations of model architectures as classification and regression tasks. The experimental configurations include varying visual feature extractors, history window sizes, audio features, and textual features. In our experiments, we sampled 3 frames per second in order to not include too much redundant information where variation the between consecutive frames is low. This sampling rate corresponds to 10% of the original frame rate of the video which has 30 frames per second. Additionally, we analyzed the effects of multimodal information by including or excluding one of the modalities. The results are given in Table 1. All models are trained for 50 epochs over the training split of the dataset using *Adam* optimizer. To avoid over-fitting we applied dropout with *0.2* on *BiLSTM* layers. Due to many configurations of experimental variables, we listed the best performing four models for each visual descriptor along with the respective history window sizes and input features from specific modalities or all.

Each trained model outputs an importance score for every frame in a video. We computed Top-1, Top-2 and Top-3 accuracy on the predicted importance scores of

[2] https://av.tib.eu/

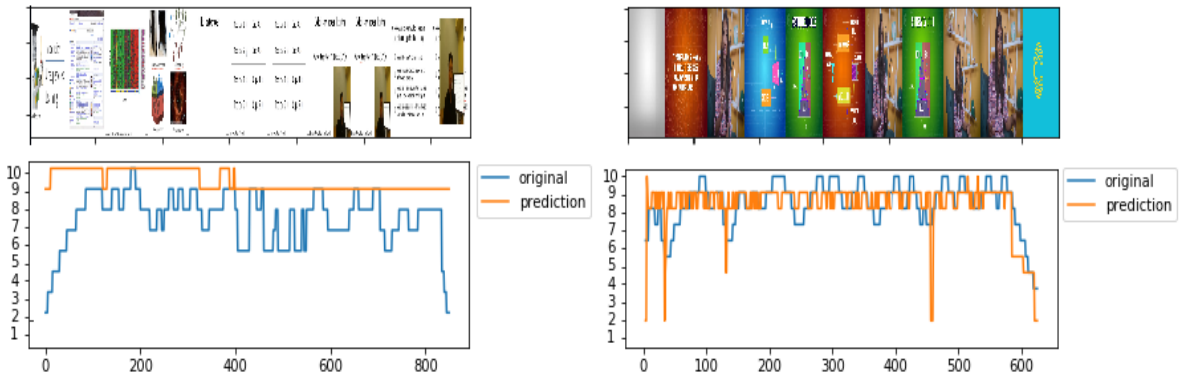

Figure 4: Predictions of VGG-16 model for two videos. Left: model prediction with low accuracy (18%), Right: model prediction with high accuracy (34%)

each frame by treating it as a classification task. The best performing model for Top-1 accuracy is *VGG-16* with a history window size of 2 achieving an accuracy of 26.3, where only visual and textual features are used for training. The model with Top-2 accuracy is *ResNet-50* with the history window of 3 that is trained on visual, audio, textual features and it achieves an accuracy of 47.3. The best performing Top-3 model is again *VGG-16* with a history window 3, visual and audio features, and it achieves an accuracy of 67.9.

In addition, we compute the Mean Absolute Error (MAE) values for each trained model by treating the problem as a regression task. Each model listed in Table 1 includes an average MAE value based on either each frame (avg_{fra}) or segment (avg_{seg}). We performed the following post-processing in order to compare the values against ground truth where every segment (5 second window) of a video contains an importance scores between 1 or 10. As explained above, trained models output an importance scores for each frame in a video. For the calculation of avg_{fra}, every frame that belongs to the same segment is assigned the same value in the ground truth videos. For calculation of avg_{seg}, predicted importance scores of each frame belonging to the same segment are averaged. This average value is then assigned as a predicted value to a segment. The avg_{seg} is an average MAE between predicted importance score of a segment and ground truth. Based on the presented results in Table 1, the model that uses *VGG-16* for visual features together with audio features and history window of 3 performs with the least error for both frame and segment-based calculation of the average MAE.

Table 1
Average accuracy and Mean Absolute Error (MAE) values for different visual descriptors and history window (h) sizes. Modalities: Visual (V), Audio (A), Textual (T). avg_{fra} stands for average MAE value based on all frames in a video, avg_{seg} stands for average MAE for each segment in a video.

Visual Features	h	Accuracy %		MAE			V	A	T
		Top-1	Top-2	Top-3	avg_{fra}	avg_{seg}			
Inception-v3	3	22.34	32.01	55.94	1.93	1.84	✓	✓	✓
	2	22.34	30.98	55.94	1.93	1.84	✓	✓	✓
	3	22.34	30.98	55.94	1.93	1.84	✓	✓	×
	2	22.34	47.3	55.94	1.93	1.84	✓	✓	×
	2	23.95	43.48	60.2	1.82	1.74	✓	×	✓
	3	23.48	44.07	64.29	1.73	1.66	✓	×	✓
VGG-16	1	22.43	47.29	66.33	1.92	1.84	✓	✓	✓
	2	22.37	37.47	57.92	1.87	1.81	✓	✓	✓
	3	25.55	46.19	**67.92**	**1.51**	**1.49**	✓	✓	×
	2	22.91	45.08	58.93	1.83	1.79	✓	✓	×
	2	**26.26**	41.92	63.09	1.6	1.57	✓	×	✓
	3	25.65	41.28	63.21	1.65	1.62	✓	×	✓
Xception	1	23.1	39.13	57.33	1.88	1.8	✓	✓	✓
	3	22.34	30.98	55.94	1.93	1.84	✓	✓	✓
	2	22.72	47.17	59.74	1.88	1.8	✓	✓	×
	1	22.42	47.2	67.12	1.86	1.78	✓	✓	×
	3	24.04	37.99	59.76	1.82	1.74	✓	×	✓
	2	22.65	44.45	62.39	1.86	1.78	✓	×	✓
ResNet-50	3	22.6	**47.31**	67.11	1.9	1.82	✓	✓	✓
	2	22.39	37.03	57.53	1.92	1.84	✓	✓	✓
	3	24.27	37.66	59.74	1.76	1.71	✓	✓	×
	2	22.75	37.25	57.34	1.85	1.81	✓	✓	×
	2	22.69	31.59	56.66	1.85	1.8	✓	×	✓
	1	22.67	31.61	57.39	1.81	1.78	✓	×	✓

5.1. Discussion

For a deeper analysis of errors made by the trained models, we plot ground truth labels along with predictions and select two videos with relatively low (left video) and high (right video) accuracy. These plots are shown in Figure 4. The video on the left side has low accuracy (18%) because the predicted values are far off from the ground truth. The reason could be the fact that frames in the video have less visual variation and

the model predicts the same or similar values for those frames. Another reason could be that the visual features are not well suited for the educational domain, since we use pre-trained models on ImageNet dataset where the task is to recognize distinct 1000 objects. On the other hand, the video on the right side has relatively high accuracy (34%). Even though the importance scores for frames are not exact, we can observe that the model predicts lower importance scores when ground truth values are also lower, and the same pattern is observed when importance scores are increased as well. As shown in Table 1, the best model obtains an error of 1.49 (MAE) on average, but it is observable that most of the important segments (regardless of the predicted values) are detected by the trained model.

6. Conclusion

In this paper, we have presented an approach to predict the importance of segments in educational videos by fusing multimodal information. This study presents and validates a working pipeline that consists of lecture video annotation and, based on that, a supervised (machine) learning task to predict importance scores for the content throughout the video. The results show the importance of each individual modality and limitations of each model configuration. It also highlights that it is not straight forward to exploit the full potential from heterogeneous source of features, i.e., using all modalities does not guarantee a better result.

One further direction of research is to enhance the architecture for binary and ternary fusion where modalities are fused on different levels. As a second future direction, we will focus on the release of another version of the dataset that covers more topics and videos. Finally, we will investigate other types of visual descriptors that better fit to the educational domain.

Acknowledgments

Part of this work is financially supported by the Leibniz Association, Germany (Leibniz Competition 2018, funding line "Collaborative Excellence", project SALIENT [K68/2017]).

References

[1] G. Pardi, J. von Hoyer, P. Holtz, Y. Kammerer, The role of cognitive abilities and time spent on texts and videos in a multimodal searching as learning task, in: Proceedings of the 2020 Conference on Human Information Interaction and Retrieval, CHIIR '20, Association for Computing Machinery, New York, NY, USA, 2020, p. 378–382. URL: https://doi.org/10.1145/3343413.3378001. doi:10.1145/3343413.3378001.

[2] A. Hoppe, P. Holtz, Y. Kammerer, R. Yu, S. Dietze, R. Ewerth, Current challenges for studying search as learning processes, Proceedings of Learning and Education with Web Data, Amsterdam, Netherlands (2018).

[3] F. Moraes, S. R. Putra, C. Hauff, Contrasting search as a learning activity with instructor-designed learning, in: Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM '18, Association for Computing Machinery, New York, NY, USA, 2018, p. 167–176. URL: https://doi.org/10.1145/3269206.3271676. doi:10.1145/3269206.3271676.

[4] P. J. Guo, J. Kim, R. Rubin, How video production affects student engagement: An empirical study of mooc videos, in: Proceedings of the first ACM conference on Learning@ scale conference, 2014, pp. 41–50.

[5] S. Zolotykhin, N. Mashkina, Models of educational video implementation in massive open online courses, in: Proceedings of the 1st International Scientific Practical Conference "The Individual and Society in the Modern Geopolitical Environment" (ISMGE 2019), Atlantis Press, 2019, pp. 567–571. URL: https://doi.org/10.2991/ismge-19.2019.107. doi:https://doi.org/10.2991/ismge-19.2019.107.

[6] Y. Song, J. Vallmitjana, A. Stent, A. Jaimes, Tvsum: Summarizing web videos using titles, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 5179–5187.

[7] C. H. . A. S. Mubarak, A.A., et al., Predictive learning analytics using deep learning model in moocs' courses videos, Springer, Educ Inf Technol (2020). URL: https://doi.org/10.1007/s10639-020-10273-6. doi:10.1007/s10639-020-10273-6.

[8] N. A. Shukor, Z. Abdullah, Using learning analytics to improve MOOC instructional design, iJET 14 (2019) 6–17. URL: https://www.online-journals.org/index.php/i-jet/article/view/12185.

[9] C. Tang, J. Liao, H. Wang, C. Sung, Y. Cao, W. Lin, Supporting online video learning with concept map-based recommendation of learning path, in: Extended Abstracts of the 2020

CHI Conference on Human Factors in Computing Systems, CHI 2020, Honolulu, HI, USA, April 25-30, 2020, ACM, 2020, pp. 1–8. URL: https://doi.org/10.1145/3334480.3382943. doi:10.1145/3334480.3382943.

[10] K. Zhang, W. Chao, F. Sha, K. Grauman, Video summarization with long short-term memory, in: Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VII, volume 9911 of *Lecture Notes in Computer Science*, Springer, 2016, pp. 766–782. URL: https://doi.org/10.1007/978-3-319-46478-7_47. doi:10.1007/978-3-319-46478-7_47.

[11] H. Yang, C. Meinel, Content based lecture video retrieval using speech and video text information, IEEE Trans. Learn. Technol. 7 (2014) 142–154. URL: https://doi.org/10.1109/TLT.2014.2307305. doi:10.1109/TLT.2014.2307305.

[12] J. Wang, W. Wang, Z. Wang, L. Wang, D. Feng, T. Tan, Stacked memory network for video summarization, in: Proceedings of the 27th ACM International Conference on Multimedia, MM 2019, Nice, France, October 21-25, 2019, ACM, 2019, pp. 836–844. doi:10.1145/3343031.3350992.

[13] M. Gygli, H. Grabner, H. Riemenschneider, L. V. Gool, Creating summaries from user videos, in: Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VII, volume 8695 of *Lecture Notes in Computer Science*, Springer, 2014, pp. 505–520. URL: https://doi.org/10.1007/978-3-319-10584-0_33. doi:10.1007/978-3-319-10584-0_33.

[14] K. Davila, R. Zanibbi, Whiteboard video summarization via spatio-temporal conflict minimization, in: 14th IAPR International Conference on Document Analysis and Recognition, ICDAR 2017, Kyoto, Japan, November 9-15, 2017, IEEE, 2017, pp. 355–362. URL: https://doi.org/10.1109/ICDAR.2017.66. doi:10.1109/ICDAR.2017.66.

[15] F. Xu, K. Davila, S. Setlur, 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 2019, Sydney, Australia, September 20-25, 2019, IEEE, 2019, pp. 1047–1054. URL: https://doi.org/10.1109/ICDAR.2019.00171. doi:10.1109/ICDAR.2019.00171.

[16] J. Shi, C. Otto, A. Hoppe, P. Holtz, R. Ewerth, Investigating correlations of automatically extracted multimodal features and lecture video quality, in: Proceedings of the 1st International Workshop on Search as Learning with Multimedia Information, SALMM '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 11–19. URL: https://doi.org/10.1145/3347451.3356731. doi:10.1145/3347451.3356731.

[17] H. N. K. S. YukiIchimura, Keiko Noda, Prescriptive analysis on instructional structure of moocs:toward attaining learning objectives for diverse learners, The Journal of Information and Systems in Education 19 N0. 1 (2019) 32–37. doi:10.12937/ejsise.19.32.

[18] W. Wang, D. Tran, M. Feiszli, What makes training multi-modal networks hard?, CoRR abs/1905.12681 (2019).

[19] N. Majumder, D. Hazarika, A. F. Gelbukh, E. Cambria, S. Poria, Multimodal sentiment analysis using hierarchical fusion with context modeling, Knowl. Based Syst. 161 (2018) 124–133. URL: https://doi.org/10.1016/j.knosys.2018.07.041. doi:10.1016/j.knosys.2018.07.041.

[20] K. Zhang, K. Grauman, F. Sha, Retrospective encoders for video summarization, in: Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VIII, volume 11212 of *Lecture Notes in Computer Science*, Springer, 2018, pp. 391–408. doi:10.1007/978-3-030-01237-3_24.

[21] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), Association for Computational Linguistics, 2019, pp. 4171–4186. URL: https://doi.org/10.18653/v1/n19-1423. doi:10.18653/v1/n19-1423.

[22] T. Giannakopoulos, pyaudioanalysis: An open-source python library for audio signal analysis, PloS one 10 (2015).

[23] F. Chollet, Xception: Deep learning with depthwise separable convolutions, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, IEEE Computer Society, 2017, pp. 1800–1807. URL: https://doi.org/10.1109/CVPR.2017.195. doi:10.1109/CVPR.2017.195.

[24] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: 2016 IEEE Conference on Computer Vision and Pat-

tern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, IEEE Computer Society, 2016, pp. 770–778. URL: https://doi.org/10.1109/CVPR.2016.90. doi:10.1109/CVPR.2016.90.

[25] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015. URL: http://arxiv.org/abs/1409.1556.

[26] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, IEEE Computer Society, 2016, pp. 2818–2826. URL: https://doi.org/10.1109/CVPR.2016.308. doi:10.1109/CVPR.2016.308.