EDUZONE – A Educational Video Summarizer and Digital Human Assistant for Effective Learning

Tandin Wangchen
Department of Computer Science and Software Engineering
Sri Lanka Institute of Information Technology
Sri Lanka
it19098838@my.sliit.lk

P. Navodya Tharindi
Department of Computer Science and Software Engineering
Sri Lanka Institute of Information Technology
Sri Lanka
it19408316@my.sliit.lk

K. C. C. Chaveena De Silva
Department of Information Technology
Sri Lanka Institute of Information Technology
Sri Lanka
it19114736@my.sliit.lk

W. D. Thushan Sandeepa
Department of Computer Science and Software Engineering
Sri Lanka Institute of Information Technology
Sri Lanka
it19161334@my.sliit.lk

Nuwan Kodagoda
Department of Computer Science and Software Engineering
Sri Lanka Institute of Information Technology
Sri Lanka
nuwan.k@sliit.lk

Kushnara Suriyawansa
Department of Computer Science and Software Engineering
Sri Lanka Institute of Information Technology
Sri Lanka
kushnara.s@sliit.lk

2022 7th International Conference on Information Technology Research (ICITR) | 979-8-3503-3203-2/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICITR57877.2022.9993155

***Abstract*—The availability of technology and the expansive nature of the internet have created a surge in the demand for online learning. Despite so many advantages, there are some existing drawbacks related to online learning. The lengthy recorded video lectures of different subjects and modules in a static manner, are extremely tedious for the learner to understand the contents available. And lack of assistance for academic-related problems of students is also stated as a major issue that comes with online education. EDUZONE provides a reliable solution to mitigate and overcome these challenges. This tool is educational assistance that generates a summarized version of the video lectures which depicts the overall idea of the whole video with the capability of a lecture notes generator along with a digital human which helps to clarify students' problems and build an efficient conversational flow. The summarized video content can be used by the learners for revisions and as a quick reference before any examinations. In addition to generating short and precise content, EDUZONE also indexes any specific topics to make it easier to find content and generate class notes, highlighting all the important content. Overall EDUZONE can be considered a time-efficient educational assistant which helps students with their studies.**

Keywords—video summarization, digital human, class notes, video indexing, video lecture, OCR, VSU Model

I. INTRODUCTION

Online learning provides an opportunity for those who do not have access to quality in-person learning. Due to the COVID-19 pandemic, most universities have led to an unprecedented transformation in teaching and learning by shifting to online. It was clear that most institutes have undertaken a considerable effort to address the issues of online delivery while presenting many obstacles to both educators and students [1].

Hence, recorded lecture videos can be considered an important online learning resource for higher education. Educational videos differ in content, presentation style, and duration [2]. For example, lecture videos mostly present PowerPoint slides relevant to the lecture topic and have long durations of more than 30 minutes. Students usually avoid watching long educational videos even though they contain great information [3]. However, the challenge of quickly finding the content of interest in a lecture video is a major limitation of this format. Also, processing these lengthy videos is challenging, as it requires lots of time. The ability to automatically summarize and extract the content and thus, search and navigate a video based on queried content would be invaluable to students. Lecture videos which are containing handwritten content can be difficult to recognize [4]. Thus, it is difficult to note down summarized lecture notes including unclear figures, graphs, and math expressions. Furthermore, Students who are more academically sound displayed a higher chance of going through more educational assistance to have accurate answers to their problems [5]. But with online learning, it is impractical to make sure each student is being supervised fully.

This research provides a solution to the previously mentioned e-learning-related problems. EDUZONE includes four main functions such as video summarization, video indexing, lecture notes, and digital human. Video summarizer helps the student to depict only the compulsory contents of the whole lengthy video. The summarized content of the video will be efficient and effective as only the necessary content is going to be extracted. Video indexing allows lecture videos to be segmented into topic units. This organizes the videos for browsing and provides search capabilities. The lecture note gives students a way to effectively record important parts that the lecturer wrote on PowerPoint slides or the whiteboard. The digital human is a 3D virtual assistant that has the possibility to provide personalized and quick services to students because it helps to build an efficient conversational flow between the user and the virtual persona.

II. BACKGROUND AND LITERATURE SURVEY

To contextualize our research, we draw upon literature on the implementation and evaluation of existing systems and tools similar to our application. The followings are literature surveys done separately for each component which are Video summarization, Video indexing, Lecture notes, and Digital human.

A. Video Summarization

The video summarizer creates a content-summarized video that can be used by the learners for revisions and as a

979-8-3503-3203-2/22/$31.00 ©2022 IEEE

quick reference before examinations. The summarized video will be efficient and effective as only the content that needs to be shown will be extracted.

Juho Kim et al. conducted a study to identify in-video drop-outs during online lecture videos [6]. They concluded that the lengthy video had a heavy drop-out as the viewers were either tired or bored by the lengthy video. This analysis was done on the online lecture video where teacher interaction keeps the viewers intact. While the offline video does not have much interaction with the viewer, making them less intact with the video, which leads to even higher drop-out in the lengthy video.

In 2016, Ke Xhang and Wei-Lun Chao [7] came up with a novel supervised learning technique for automatically summarizing videos by selecting keyframes or sub-shots. However, summarization of the offline education video can be very heavy for the CPU and GPU, as offline educational videos generally from one to four hours. To process such a lengthy video, the GPU and CPU are loaded with tons of images and methods to the cache creating it unreliable as a product.

But Selahattin Akkas et al. [8] created a Single Shot Multibox Detector model from TensorFlow Detection Model Zoo. That captured images from the video efficiently on CPU and GPU. This model was built for the Indy Car series, where they needed to predict tasks to increase race safety and develop better strategies to win the race. Therefore, this proposed system overcomes the above issue of CPU and GPU overloading while having a smooth summarized video that is effective and efficient for the viewers.

B. Video Indexing

By using automatic segmentation, lengthy lecture videos can be divided into topic units, making it easier to find what you're looking for, and improving the learning experience overall. A considerable amount of effort has progressed into indexing videos based on visual content, audio content, and text. However, very few studies have been done on the difficulty of indexing lecture videos into topic units.

Scene changes are the most used mechanism for video indexing. Research done by HongJiang Zl1ang et al suggested a system for recognizing gradual transitions based on both statistical analysis and motion [9]. However, because lecture videos include few scene changes and those that do not correspond with topic transitions, this process of indexing based on scene changes is ineffective for video lectures.

Another focused method for indexing videos and keyframe detection is applying OCR (Optical Character Recognition) technologies. Research done by L. S. Kate et al. presented technology for video search in lecture video archives by applying ASR (Automatic Speech Recognition) on lecture audio and OCR on video content to extract metadata [10]. They describe a method for texts in slides, content-based video indexing, and video recovery in vast video archives in that study. Research done by Che X et al. proposed a method for indexing lecture videos based on the synchronized slides that accompany them using slide matching and OCR. But they assumed that the slides and video streams are in synchronization, which may not always be the case. Furthermore, it is only accurate for particular types of lecture videos because their method is primarily reliant on matching slide content to video, [11].

Furthermore, indexing methods based on closed captions or transcribed text have been the main motivator for working in this field of Topic Detection and Tracking (TDT). The narrative segmentation task in TDT refers to the task of indexing a stream of data (transcribed speech) into topically consistent stories. Michael Chau et al. conducted a study to identify topic changes based on multiple linguistic features [12]. This research is a potentially successful solution that can be improved further because transcribed text extracted from lecture videos provide rich content information for the detection of topic changes.

C. Lecture Notes

The lecture note generator is essentially a function that creates a class note using lecture videos. S. S. Alrumiah et al. [2] developed a Latent Dirichlet Allocation (LDA) based subtitles summarization model. The subtitles of educational videos were summarized using an LDA-generated keywords list. Their process follows three phases, which are pre-processing, keyword generation using LDA, and summarization based on the subtitles.

To identify visual text, A. Kumar et al. [13] proposed an efficient method for text extraction: complex video text images. It operates in three stages; edge detection, text localization, and text segmentation. S. Shetty et al. [14] provide a new method to detect and recognize the text from video frames with Optical character recognition (OCR). H. V. Shin et al. [15] introduced Visual Transcripts, a readable and interactive representation of blackboard-style lecture videos. They separate the visual material of a lecture video into discrete figures using a variation of the standard line-breaking algorithm. The transcript text is then structured using the temporal relationship between the figures and the transcript sentences. The researchers C. Xu et al. [16] describe a revolutionary lecture-note-generating approach in this study. The visual entities are first retrieved from presentation slides of lecture videos, and then semantic similarity detection is used to associate visual things with their corresponding descriptive speech sentences. Then, a placement optimization approach is proposed to fit the visual things and speech texts into a note.

D. Digital Human

Digital humans are 3-dimensional virtual objects that look almost equally like human beings. These are independent 3D objects existing in virtual worlds, similar to voice bots or traditional chatbots that we are commonly using today [17].

In 2008, Doering et al. [18] studied the effect of conversational agents on communication. They considered the communication of agents and participants when agents help participants in designing an online portfolio. The collected data for four weeks showed that participants' and agents' conversations were not only related to the domain of e-portfolio designing but also a range of different subjects. In conclusion, the researchers point out three things that need to be improved. Improving the intelligence of agents regarding the specific domain, learner-developed conversational agents, and agents that can satisfy realistic and humanistic expectations of users are the main three facts proposed by them.

When considering engineering education, virtual agents are providing an interesting medium because these allow active and authentic learning experiences and visual collaboration. In 2013, Soliman and Guetl proposed a system [19] that has a pedagogical agent in the virtual world. Thus, this proposed system is also included lacking tuition support from the teacher for clarifying doubts. The research explains a prototype design of an Intelligent Processes Automation (IPA) that can communicate with a learner in natural language and support their understanding of the virtual experiments.

III. Research Objective

The main objective of this research is to provide an educational assistant software solution for video lectures under four components. It aims at improving video lecture delivery and learners' engagement. Following are the four components provided by EDUZONE.

- Lecture video summarizer that summarizes the video including only important topics, points, and concepts based on their audio, video, and emotions.
- Video indexer which segments lecture videos by topic units displays with a navigable visual index.
- Lecture note generator that identifies hand-drawn obscure contents and visual contents in lecture videos and generates an automated class note with visual content.
- Digital human assistant which helps users to clarify lecture content and have a human-like conversation using a verbally infected 3D virtual avatar.

All those features are provided for university students and academic staff who are facing difficulties with online learning and searching for better solutions.

IV. Methodology

EDUZONE combined necessary tools which help to optimize recorded lecture delivery and student engagement. Unlike other products which are dependent on the use of manual processes that demand user interaction, our product can be considered as an automated low processing power product. EDUZONE provides four main functionalities. Video summarization, video indexing, lecture notes, and digital human. The following sections explain how the four main functionalities blend in to create EDUZONE.

A. Video Summarization

According to the technique depicted in Fig. 1, the original video was divided into two files: a wav file for decoding the audio files and a file of images taken once every third of a second. These files are produced on a temporary file that will be removed after the procedure is finished. The noise and the factors to eliminate the white spaces will be executed and interpreted in the folder.

The audio is compressed and decompressed using a Python module called audiostm without interrupting the video's flow. When an image needs to be eliminated, it is either copied or cropped out of the frame, resulting in a smooth video transaction.

After identifying the white space, it moves to the following process, where the new audio is processed again.

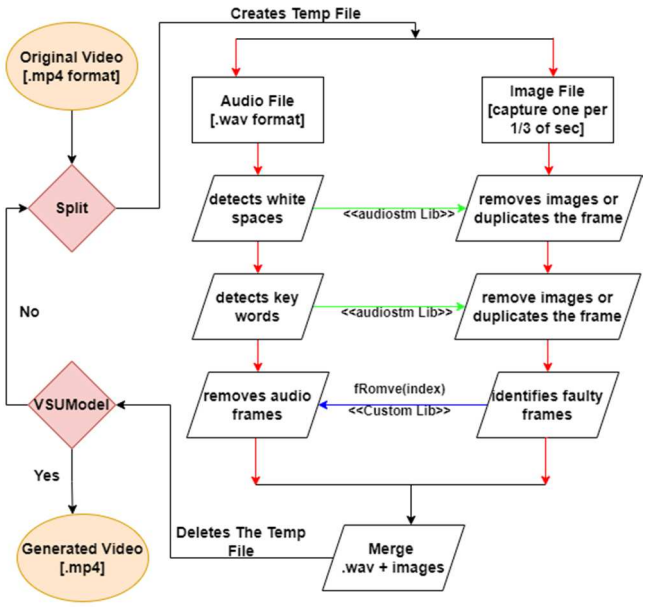

Fig. 1. Flow chart of the Video Summarizing process

This process identifies some keyword that is coded into the program. If the system catches any identified keyword during the process, it will try to capture the audio frame. The unwanted images will be removed with the help of audiostm and capture audio frames.

After the images are finished, the new images are further filtered to remove the faulty frames. The pre-dataset that has been added to the software finds the defective frames. This filtration corresponds to the audio frame that needs to be removed.

During the merge between audio and images, the temporary file will be deleted while triggering the VSUModel, which checks the smoothness of the video. If the model passes, the summarized video will be generated. However, if the model fails, then the process will be redone.

B. Video Indexing

Video indexing provides automatic video indexing while providing a navigatable index to the video by topic changes throughout the lecture video. We focus on development using video transcripts because of the minimum time to process and ability to hope high accuracy.

As Fig. 2 explains, the process starts by using the previously summarized recorded lecture as input and generating a text transcript from the video. After splitting the video into identified several clips, using generated transcript as an input, used topic splitting algorithm will identify topic transitions using machine learning techniques and index the video by identified topic units. For transcribing videos the audio is generated from the video using python MoviePy library.

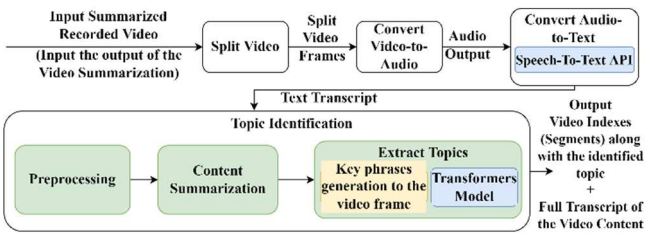

Fig. 2. System flow of Video Indexing process

Finally using the converted audio, the text will generate using Google Speech-To-Text API. This generated text transcript is used for the topic model.

To represent the document, we are using a transformer-based embedding technique. Embedding is a numerical representation of the data that has meaningful distances between the points. Word embedding is a type of word representation that allows words with similar meanings to have similar representations. Objects can be donated with a proper name and have some physical and abstract existence. We are using Google NLP API to extract those entities for every text transcript. The Data Preprocessing step of topic modeling removes all the unwanted things from the text which do not carry information about topics. And using NLTK, text content is summarized to improve accuracy and to Fine-tune. By using Transformers, the topics/key phrases are extracted. The Transformer uses layered self-attention and pointwise, completely linked layers to comply with the encoder-decoder design as a whole. The transformer's attention function is calculated by translating a query to an output and a collection of key-value pairs. The result is then calculated as a weighted total of the values, with each value's weight determined by the query's compatibility function with the connected key. And after the topics/ keywords identification most important topic will be returned from the most repeated for that specific text piece. And finally, those returned topics along with their indexed clip will be available to the user. Users can search by topic and navigate to that indexed lecture clip. And the generated full transcript is available to the user for references and usage which enhances accessibility.

C. *Lecture Notes*

Lecture Notes Generator is a system that creates well-organized class notes using lecture slides-based videos (Fig. 3). This system has the potential to provide fast and efficient services to the students as the generated class notes can be used by the learners for revision and as a quick reference before any exam. Users can interact with the system through the user interface. Lectures have to upload their videos and lecture slides to the system. When the user requests the lecture note related to the lecture video, the proposed system analyzes the lecture slide and video and generates a well-organized class note. As the input to this system, The user can input video. As the first step, split the video into small segments. one segment duration is 5 seconds. Then generate audio files from the video segment using the moviePy python library. These audio files are created in (.wav) format, so we have to convert them to mp3 format for the next step. Then the transcript is created using the audio files. For this, the IBM Watson speech-to-text API is used. Then generate the subtitle file. The next stage is to segment the input video into frames.

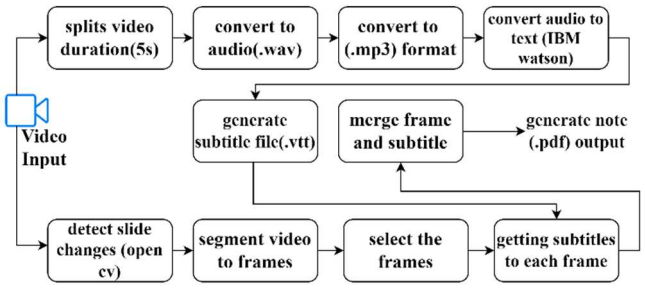

Fig. 3. System flow of Note Generator process

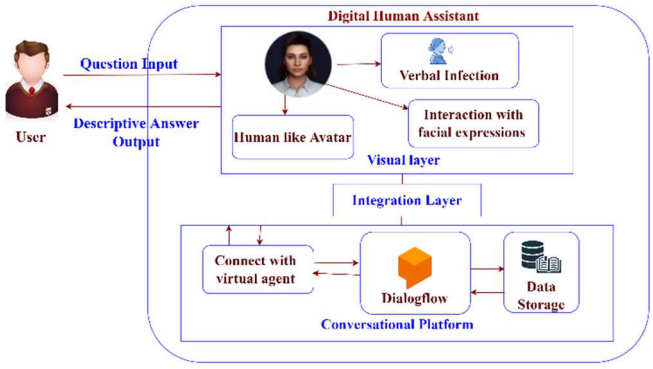

Fig. 4. System diagram of Digital Human Assistant

When the lecturer switches to the following PowerPoint presentation, this program uses computer vision with OpenCV to identify animations and segment videos into frames. After that, the selection of the correct frame and the selection of the corresponding subtitles are carried out. Then merge those frames and subtitles and generate the PDF lecture note as the output.

D. *Digital Human*

The Digital human is a 3D virtual assistant that is a combination of AI with a lifelike face and verbal inflection. Digital human technology has the potential to provide quick and personalized services to students because it helps to build an efficient conversational flow between user and virtual persona. When the user requests the answer to a question relevant to the lecture topic, the proposed digital human model will analyze the question and will give the best answer.

Fig. 4 displays the proposed system diagram for digital human assistants. The digital human assistant has three layers which are the visual layer, conversational platform, and integration layer. The visual layer provides the actual digital human and background. It has a digital human persona, which has a human-like expression when conversing with students. The conversational platform or chatbot layer enables the conversation with the digital human.

This is where the digital human is trained to recognize the questions and give the correct response. This is done with google cloud Dialogflow. The text of the conversational platform is converted into speech using react speech-to-text library. Conversely, the student's spoken question will convert back into the text to process by the conversational platform. The conversational platform and visual layer are integrated by the integration layer. The video which contains the lip-synced digital human avatar and voice is programmed to render at the right point in the response. So, it helps to build the conversational flow, and ensure that the visuals and voice are in line with the voice. The user interface is created based on customer requirements, builds the conversational flow, and ensures that the visuals are in line with the use case.

V. Result and Discussion

This section describes the evaluations used to analyze the performance of the system and the results obtained from it. We evaluated each component separately to have a better understanding of the performance, accuracy, efficiency, and user satisfaction.

A. Video Summarization

The generated summarized video should have accuracy, usability, and effectiveness. To evaluate these attributes, we tested the application with a group of testers (students). Testers were asked to test the video summarizer and give a rating compared to the original video.

The followings are the responses given by testers.

- Accuracy: 82% said it had the key features in the generated video compared with the key features of the original video.
- Usability: 93% said that the generated video had learnability, 86% said that the generated video did not have errors, and 87% were satisfied with the generated video compared to the smoothness of the video.
- Effectiveness: The group that had the generated video did perform well in the question and answers compared to the group that had the original video.

The generated video standouts to be good, but when it comes to the execution process from original to generated video, there is a time latency. This time latency is created due to figuring out the noises in the video.

B. Video Indexing

In order to index the video using topic units first step we do is to generate the text transcript from Speech-to-Text API. As we used the generated text transcript for the topic modeling, we should consider how reliable is speech-to-text output. To measure the reliability of the speech-to-text API, latency, accuracy, and Word Error Rate (WER) is selected as the test categories. In general, accuracy is the exact inverse of WER. To measure and compare the available market we used a product market analysis.

According to the market analysis by A.Roy for cxtoday, despite today's powerful AI, the average WER for speech-to-text is far from 100%. According to the publishment of benchmarks in May 2021, Microsoft had an accuracy of 81.01%, AWS had an accuracy of 83.12% (i.e., 16.88% WER) and Google had an accuracy of 84.46%. Therefore, using google Speech-to-Text is reliable to use to generate the text transcript. The performance of the video indexing process was evaluated against manual indexing by a human with twenty (20) cases. For that, we kept the skills of the manual tester unchanged by using the same user, and to test the performance of the system we selected lecture videos with different time durations.

As TABLE I shows from this evaluation, we have identified that the accuracy of the automatic video indexing is 89.59% on average. ((Automatic Segmented topics/ Actual or Manually segmented number of topics) * 100%). Therefore, our automatic indexing method can be considered a practical and effective way to identify and index the video using topic transitions. Furthermore, by this performance testing, we have identified that the manual indexing process is exhausting and time-consuming. The human took a considerable average of forty minutes in addition to the duration of the lecture video. (For a 1-hour video, a human took 1 hrs. and 40mins to sub-index the video) But the automatic segmenting process only took approximately 7 minutes which is dependent on the internet connectivity.

TABLE I. VIDEO SEGMENTATION RESULTS

Lecture Video	Duration of summarized lecture video (in minutes)	Number of Segments by Manual Indexing	Number of Segments by Automatic Indexing
A	31	8	7
B	40	8	8
C	21	5	5
D	12	4	3
E	28	7	6
F	37	9	8
G	82	20	17
H	14	4	3
I	38	8	8
J	45	10	9
K	35	7	7
L	55	12	11
M	70	16	14
N	34	8	7
O	36	9	8
P	41	7	9
Q	08	2	2
R	45	12	9
S	50	11	10
T	18	6	4

C. Lecture Notes

To measure the accuracy of the note generator we created a user survey. We took comments from thirty (33) university students. The lecture was provided to access and asked to create a summary, including the critical content. The summary they created and the note created by our system is given to them and compared and their responses were obtained through the survey. According to the answers given after considering their critical content, the result of comparing the key features above was obtained. More than 80% majority had given the highest ratings.

Unit testing was done by writing test cases to test functions one by one. The slide change points of the ten videos provided were identified and compared with the time stamp of the frames separated by the system. Video frame recognition gave a high rate of 95%. Getting the right subtitle part for a given frame was tested. The related subtitle parts were identified and matched with the subtitle released for the relevant frame through the system. It showed a result of 94%.

D. Digital Human

The digital human assistant is expected to provide the best performance by answering academic-related questions. So, we conducted a user acceptance test with fifty (50) participants to confirm whether it is fulfilling the user requirements. As the digital human aim to provide a virtual persona that has a human-like face and expressions with verbal infection, the ideal situation of the evaluation was to get the test subject to evaluate the digital human assistant compared with a real person.

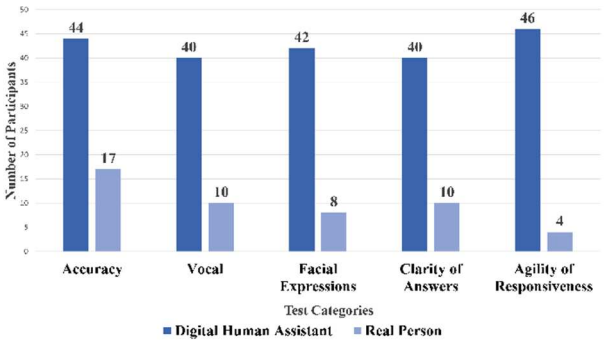

Fig. 5. Responses from user acceptance test

The study was carried out with 50 participants who were randomly selected among computing faculty students at SLIIT. Participants were requested to ask five questions which are related to academic studies from the digital human and real person. Group members participated as real person prototype and answered questions. Lastly, we ask participants to rank two prototypes they interacted with in terms of accuracy, vocal, facial expressions, clarity of answers, and agility of responsiveness.

Fig. 5 shows evaluations made with the user acceptance testing. According to responses participants found the answers given by the digital human are clearer, faster, and more accurate than the group member (real person). Students who were more interested in natural voice than digital human voice commented if they can customize the digital human voice according to their preferences for male or female, it will bring a more enjoyable and interesting experience for them. Furthermore, as facial expressions are only generated by lip-syncing movements, students were interested in having a digital persona that can animate the rest of the face such as eyes and head movements. Overall, we found participants expressed an interest in seeking answers using the digital human assistant and enjoyed the experience.

VI. CONCLUSION

EDUZONE is a solution to the modern problem of the online asynchronous learning platform. In many institutions, lectures are delivered via recorded lecture videos. They offer a wide range of advantages. Because recorded lectures are sometimes long and lack interactive elements, learner engagement is inadequate. The objective of the EDUZONE is to offer a summarized lecture video that includes only important topics, increase the searchability of the lecture video by indexing by topics while providing easy navigation through the video, provide a well-organized classroom note for user reference, and provide a live lecturer interaction via a digital human assistant. EDUZONE helps users to be more efficient and faster to adequate with the resources as the user doesn't have to spend additional time in creating or searching the information while offering an experience similar to a live lecture. As for the future, EDUZONE can be improved upon by increasing the accuracy of video summarization and video indexing. Lecture note generation can be enhanced by including handwritten whiteboard content. Furthermore, digital human assistants can be improved by training in a wide range of topics.

REFERENCES

[1] Weerathunga, Prageeth Roshan and Samarathunga, WHMS and Rathnayake, HN and Agampodi, SB and Nurunnabi, Mohammad and Madhunimasha, MMSC, "The COVID-19 pandemic and the acceptance of E-learning among university Students: The Role of Precipitating Events," *Education Sciences,* vol. 11, p. 436, 2021.

[2] Alrumiah, Sarah S and Al-Shargabi, Amal A, "Educational Videos Subtitles' Summarization Using Latent Dirichlet Allocation and Length Enhancement," *cmc-computers materials & continua,* vol. 70, 2022.

[3] Carmichael, Michael and Reid, A and Karpicke, Jeffrey D, "Assessing the impact of educational video on student engagement, critical thinking and learning," *A SAGE white paper,* 2018.

[4] B. U. Kota, "Lecture Video Summarization by Detection and Representation of Content," State University of New York at Buffalo, 2020.

[5] H. A. Abu-Alsaad, "Agent applications in e-learning systems and current development and challenges of adaptive E-learning systems," in *2019 11th international conference on electronics, computers and artificial intelligence (ECAI)*, 2019.

[6] Juho Kim, Philip J. Guo, Daniel T.Seaton, Piotr Mitros, Krzysztof Z. Gajos, Robert C. Miller, "Understanding in-video dropouts and interaction peaks inonline lecture videos," in *first ACM conference on Learning@ scale*, 2014.

[7] Zhang, K., Chao, WL., Sha, F., Grauman, K, "Video Summarization with Long Short-Term Memory," in *European conference on computer vision*, 2016.

[8] S. S. M. a. J. Q. S. Akkas, "A Fast Video Image Detection using TensorFlow Mobile Networks for Racing Cars," in *IEEE International Conference on Big Data (Big Data)*, Los Angeles, CA, USA, 2019.

[9] Zhang, HongJiang and Smoliar, Stephen W, "Developing power tools for video indexing and retrieval," in *SPIE*, 1994.

[10] Kate, Laxmikant S and Waghmare, MM and Priyadarshi, Amrit, "An approach for automated video indexing and video search in large lecture video archives," in *International conference on pervasive computing (ICPC)*, 2015.

[11] Che, Xiaoyin and Yang, Haojin and Meinel, Christoph, "Lecture video segmentation by automatically analyzing the synchronized slides," in *Proceedings of the 21st ACM international conference on Multimedia*, 2013.

[12] Lin, Ming and Nunamaker, Jay F and Chau, Michael and Chen, Hsinchun, "Segmentation of lecture videos based on text: a method combining multiple linguistic features," in *37th Annual Hawaii International Conference on System Sciences*, 2004.

[13] Kumar, Anubhav and Awasthi, Neeta, "An efficient algorithm for text localization and extraction in complex video text images," in *2013 2nd International Conference on Information Management in the Knowledge Economy*, 2013.

[14] Shetty, Shashank and Devadiga, Arun S and Chakkaravarthy, S Sibi and Kumar, KA Varun, "Ote-OCR based text recognition and extraction from video frames," in *2014 IEEE 8th international conference on intelligent systems and control (ISCO)*, 2014.

[15] Shin, Hijung Valentina and Berthouzoz, Floraine and Li, Wilmot and Durand, Frédo, "Visual transcripts: lecture notes from blackboard-style lecture videos," *ACM Transactions on Graphics (TOG),* 2015.

[16] Xu, Chengpei and Wang, Ruomei and Lin, Shujin and Luo, Xiaonan and Zhao, Baoquan and Shao, Lijie and Hu, Mengqiu, "Lecture2Note: Automatic Generation of Lecture Notes from Slide-Based Educational Videos," in *2019 IEEE International Conference on Multimedia and Expo (ICME)*, 2019.

[17] D. M. Berry, "Introduction: Understanding the digital humanities," in *Understanding digital humanities*, Springer, 2012, pp. 1-20.

[18] Doering, Aaron and Veletsianos, George and Yerasimou, Theano, "Conversational agents and their longitudinal affordances on communication and interaction," *Journal of Interactive Learning Research,* vol. 19, 2008.

[19] Soliman, Mohamed and Guetl, Christian, "Implementing Intelligent Pedagogical Agents in virtual worlds: Tutoring natural science experiments in OpenWonderland," in *2013 IEEE Global Engineering Education Conference (EDUCON)*, 2013.