Content based Lecture Video Retrieval using Textual Queries: to be Smart University

Cu Vinh Loc, Nguyen Thanh Nhan, Truong Xuan Viet, Tran Hoang Viet, Le Hoang Thao, Nguyen Hoang Viet
Can Tho University, Vietnam
{cvloc, nhannguyen, txviet, thviet, lhthao, nhviet}@ctu.edu.vn

2021 13th International Conference on Knowledge and Systems Engineering (KSE) | 978-1-6654-9975-0/21/$31.00 ©2021 IEEE | DOI: 10.1109/KSE53942.2021.9648820

***Abstract*—The amount of lecture videos is rapidly growing due to the popularity of massive online open courses in academic institutions. Thus, the efficient method for lecture video retrieval in various languages is needed. In this paper, we propose an approach for automated lecture video indexing and retrieval. First, the lecture video is segmented into keyframes in a manner that the duplication of these frames is minimal. The textual information embedded in each keyframe is then extracted. We consider this issue as a matter of text detection and recognition. The text detection is solved by our segmentation network in which we propose a binarization approach for optimizing text locations in an image. For text recognition, we take advantage of VietOCR to recognize both English and Vietnamese text. Lastly, we integrate a vector-based semantic search in ElasticSearch to enhance the ability of lecture video search. The experimental results show that our approach gives high performance in detecting and recognizing the text content in both English and Vietnamese as well as enhancing the speed and accuracy of lecture video retrieval.**

***Index Terms*—lecture video, keyframe extraction, text detection, text recognition, query-by-text**

I. Introduction

In recent years, more and more universities are utilizing the opportunity of digital recording technology to record their lectures and deliver them online for their students to retrieve independent of time and location. This technology is also applied for implementing the modern distance education and e-learning system. The educational institutions consider online learning and teaching as a mandatory part of their training system in accordance with current situation. The usage of videos as a learning and teaching resource has become common in an online environment. As a result, a huge amount of e-lecturing has been produced on the Web. The recent survey by University of the People [1] has been conducted based on the Youtube educational channels in 2019, which is shown in Table I. The figures show that the interest of users are more towards videos-based learning.

TABLE I
Statistics on YouTube Educational Channels

Channels	Subscribers	Video views (times)
TED Ed	10.2 million	1.6 billion
Crash-Course	10.2 million	1.5 billion
Khan Academy	5.24 million	1,7 billion

The popular video search engine like Bing, Youtube and Vimeo is mostly conducted and replied on textual metadata. The metadata consists of short description, title, genre, author and etc. [23]. This kind of information is added to the video to make sure high quality search results. However, the process of metadata creation is manually produced, expensive and time consuming. Besides, the metadata describes less information, and it is likely not to cover entire content within a video. Thus, the users might loss desired videos regarding to their requested information. To find out the most relevant videos corresponding to the user's query and to automatically generate the metadata, we have found that there have been a lot of existing works to solve this issue including text detection and graphic localization [14], [15], [23], feature extraction [7], [11], [16], text transcript from the audio component [12]. These methods work well for lecture videos in English, but they fail in recognizing text in Vietnamese language.

The content of a lecture video mostly contains speech, graphic, printed text [14] and text in subtitle. The textual information that appears in the lecture slides contains almost all the content of a teaching topic. Thus, in this work, we focus on developing a system for video retrieval based on the printed text and text in subtitle of a video. The proposed approach is able to apply for text content in both English and Vietnamese. In addition, the system not only assists the learners in searching the required lecture content more efficiently, but also provides a solution in application architecture to build a smart university. Unlike the existing works, the authors make use of OCR technique for text detection and recognition. We propose a segmentation network to detect text in the extracted keyframe, which is hereafter referred to image, in which we adjust the binarization process for heatmap generation instead of using standard binarization. The keyframe extraction and semantic search integration are also novel points proposed in our approach.

The paper is organized as follows. The existing works are reviewed in the next section. In section III, the proposed method is presented. Section IV highlights the experimental results. Conclusion and future enhancements are discussed in section V.

II. Related Work

Content-based video retrieval is a multidisciplinary and active research field. The existing works in this field are classified into spatial and temporal domains [15] for content representation in the lecture videos. The spatial domain focuses on feature vectors obtained by evaluating the different

pixels of video frames. Meanwhile, the temporal domain is performed by dividing the video into shots, frames, and scenes.

Nguyen *et al.* [14] have proposed a multi-modal analysis of graphics and text. The authors classify the information contained within the lecture videos into speech, text and graphical elements. The speech is transcribed by using VoxSigma. Text elements are recognized by classical Optical Character Recognition (OCR). Meanwhile, the visual words are extracted by utilizing SIFT detector. Another method based on OCR and Automatic Speech Recognition (ASR) has been proposed in the work [23]. In this work, the authors make use of Connected Component (CC) to segment the content of a slide frame. The textual content is then recognized by using OCR, and the ASR is used to provide speech-to-text information from lecture videos. Hao *et al.* [7] have proposed near-duplicate video retrieval which is conducted through 4 steps: keyframe extraction by using the method of shot-based sampling; feature extraction by Local Binary Pattern (LBP); hash coding learning by hash function; and the computation of video similarity by utilizing Hamming distance.

A combination of Convolutional Neural Network (CNN) and conventional handcrafted descriptors has been presented in the work [11] in which the authors propose a Nested Invariance Pooling (NIP) for computing invariant representation. To improve the discrimination of global descriptor, the hybrid pooling moment within NIP is designed. The features extracted by this approach are robust against geometric transformation. A two-phase approach has been proposed by Ashok Kumar *et al.* [15]. For offline phase, the authors segment video frames by applying block level keyframe extraction. The text in the right frame is recognized by OCR. The extracted text is then processed to get meaningful keywords. For online phase, the authors propose a strategy for ranking matched videos. Poornima *et al.* propose a method to extract useful keyframes, and the Fuzzy C Means algorithm is then applied to extract features from the obtained keyframes. The extracted features are fed to a deep learning scheme for optimizing the user query. Another scheme for e-lecture video retrieval [12] is based on machine learning text classification algorithm. The summary of extracted text and keywords are then used for training the text classification model.

Balasubramanian [4] *et al.* propose a model to extract key phrases from the video contents. The extraction process is conducted based on the features of video transcripts and slide content. The extracted key phrases make the retrieval process better. Waykar and Bharathi [22] propose a scheme based on probability extended nearest neighbor (PENN) in which multiple modalities like texture-based content features and OCR are combined for video retrieval. The OCR technique is used for text recognition, and local vector pattern is utilized to extract features from the video content. The process of video retrieval is then performed by using PENN classifier. Another method by Daga [5] *et al.* extracts features from the keyframes. These features consist of semantic word, text, and local gabor pattern vectors. The authors utilize K-Nearest Neighbour Naive Bayes and Naive Bayes classifier to retrieve lecture videos from the feature database.

III. Proposed Method

The workflow of our approach is depicted in Figure 1.

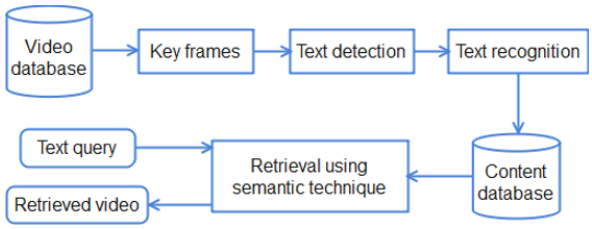


Fig. 1. The workflow of entire system.

A. Keyframe extraction

The lecture videos consist of large data with intensive and redundant information, and complex hierarchical structure. The video contains frames, shots and scenes. A number of frames, shots and scenes form shot, scene and video respectively. In other words, the lecture video is a series of still images, and each of these images is called a frame. The most popular frame rate for video is 24 fps, 30 fps, and 60 fps. For lecture videos, a lecture slide may consist of several content items, and the time that the transition from a slide to another slide may be different from others. Thus, it is difficult to fix a frame rate for entire video. If we choose a fixed frame rate, it is possible to obtain a lot of redundant information (e.g. overlapping data in a sequence of successively related frames) or to lose frames. In order to solve the mentioned issues, the conventional approach of splitting video into frames does not give high performance. Instead, the scene change detection is an effective approach to minimize the duplicated frames and lost frame as well. The main steps of the proposed method are depicted as follows:

- Step 1: Convert the frames of a video to gray level ones.
- Step 2: Generate median binary pattern (MBP)-based frame features from each of the obtained gray frames.
- Step 3: Compute the feature difference between each and every successive frame features.
- Step 4: Compute the mean of all feature differences, which is considered as a threshold.
- Step 5: Compare the frame difference with the obtained threshold to choose the right frames.

B. Text detection

Detecting text in the keyframes/ images with complex background or multi-orientation is a challenging task, and the conventional OCR technique fails in most of these cases. To enhance the ability of localizing text embedded in the complex images, there have been many proposed approaches such as text proposal [19], region awareness [3], pixel aggregation [21], PixelLink [6], textbox++ [10] and etc. Similar to the existing works in which the segmentation-based approaches are popularly used to detect text in the image with complex background or text with different orientation, we also take

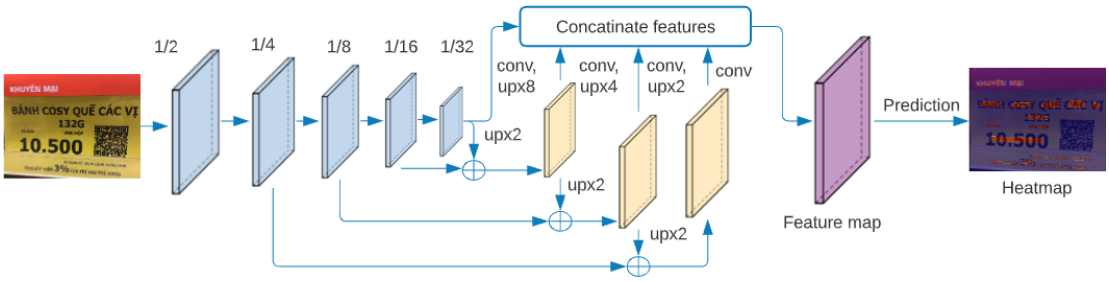


Fig. 2. The architecture of text detection network.

advantage of this approach for text detection task, and the method of semantic segmentation [17] is utilized .

The architecture of our network is presented in Figure 2, where the sign of $\oplus$ depicts the element-wise sum of the upsampling layers, the 1/2, 1/4,..., 1/32 illustrate the scale ratio compared to an original image, the sign of "up$\times$i" demonstrates upsampling with ratio i, and the "Prediction" operation consists of a 3×3 convolutional operator and two deconvolutional operators. The workflow of the network can be briefly described as follows. The input image is fed into a network consisting of downsampling and upsampling process. The operations of dowsampling and upsampling process are based on the work presented in [17]. However, it differs from the existing work in the upsampling process. In this work, we recover the dimension of the feature maps by making use of element-wise sum operator as depicted in Figure 2.

The downsampling process consists of several convolutional layers for feature extraction. The convolutional layers of the downsampling process are partly relied on the VGG-16 network [18] in which the dense layers are converted to the convolutional layers. The dimension of feature map generated by this stage is reduced. Thus, the size of this feature map is necessary to be reconstructed to the original one, and the reconstruction is performed by the process of upsampling (depicted by "up" in Figure 2). In the common segmentation task, the feature map is used to predict the probability map , and the probability map is then used to generate the heatmap by applying a threshold procedure. By this approach, we have observed that the text regions with low contrast against a background in the images are missed by computing connected components on the obtained heatmap.

To solve this issue, in this work, we use the obtained feature map to predict segmentation probability map and threshold map, and these two maps are then combined to generate an enhanced map. We then utilize both adaptive threshold and dilated convolution on the enhanced map to obtain the heatmap, and this process is depicted by.

$$H_{i,j} = \frac{1}{1 + e^{-k(P_{i,j} - T_{i,j})}} \tag{1}$$

where H is the heatmap, T is the adaptive threshold map, P is the probability map, (i, j) indicates the pixel coordinates in the map, and k depicts the amplifying factor.

In the training process, the probability map, the threshold map and the heatmap have training label in which the probability map and the heatmap share one training label. We have observed that the threshold map can be obtained by learning without supervision, and the threshold map emphasize the edges of text regions. To obtain better result, we utilize the edges of text regions to supervise the threshold map. The ground-truth for the probability map is inspired by the work presented in [20]. In the inference process, the bounding boxes surrounding the text regions can be obtained by the probability and heat map. The loss function L of the network can be described by.

$$L = log_{10}(L_B) + \alpha \times L_P + \beta \times L_T \tag{2}$$

where α and β are weight factors, L_P is Dice loss, and L_B is Active Contour loss. L_T is a loss which is computed by L_1 distance between the prediction and ground-truth inside the dilated text polygon.

C. Text recognition

In this work, we adopt VietOCR [2] to recognize text because this approach outperforms the popular methods like tesseract OCR and EasyOCR. In addition, the VietOCR is capable of recognizing both Vietnamese and English text. This approach consists of two main components such as AttentionOCR and TransformerOCR. The AttentionOCR is designed by a combination of CNN and Attention Seq2Seq. The input of this module is an image, and the output of this CNN network is a feature map. The feature map then becomes an input of a long short-term memory (LSTM) model. The dimension of the feature map is flatten in accordance with the size of LSTM model. At each time, the LSTM model needs to predict what the next word in the image will be.

The architecture of TransformerOCR consists of encoder and decoder module. The encoder module is used to learn the representation vector of a sentence with the expectation that this vector carries the perfect information of that sentence. Meanwhile, the decoder module performs the function of converting the representation vector into a target language. Training the AttentionOCR and TransformerOCR is similar to Seq2Seq model, they both use cross entropy loss to optimize the model instead of using CTCLoss like CRNN model. It means that at a time the model predicts a word, and compares

it with the corresponding label. This process is to calculate the loss value and update the weight of the model.

D. Semantic search

After locating and extracting the video content, the extracted texts are collected and stored in a JSON format. We adopt MongoDB for storing and searching the video content because of its effectiveness. The structure of JSON format for storing the video lecture is depicted in Figure 3. To optimize the time consuming for searching and ranking the lecture videos, we integrate a search engine like Elasticsearch to MongoDB. This is capable of keyword and semantic search.

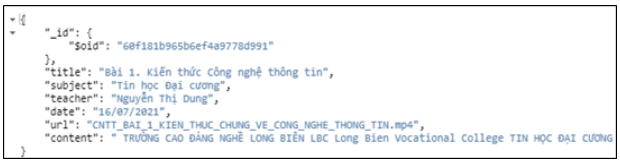


Fig. 3. The structure of JSON format for storing the lecture video.

IV. Experimental results

To locate the text in the lecture videos with complex background, the proposed text detection network is trained on the public dataset entitled VinText [13]. This dataset contains Vietnamese scene text images, and it consists of 2,000 images (about 56,000 text instances). The training set contains 1,200 images. The test set includes 500 images, and the validation set consists of 300 images. The text appeared in these images is highly diversified in the direction such as horizontal text, multi-directional text, and curve text. Regarding the initialization of network parameters, we optimize the loss parameters with $\alpha = 1$ and $\beta = 10$. The learning rate is initialized by 0.0007. The weight decay, high momentum and batch size are set to 0.0001, 0.9 and 16 respectively. We train the network on the mentioned dataset for 100 epochs.

For lecture video dataset, we evaluate the system on the public videos which are downloaded from the Youtube channel. These videos contain the lecture content in various fields. A total of 150 lecture videos are included. For the assessment of the system, we utilize the standard measures like recall, precision and f-measure metrics which are used in most scene text detection research. To prove the performance of our approach, we illustrate the gained results based on the following aspects.

A. Frame extraction

We use 150 lecture videos as described above to assess the keyframe extraction. The threshold to choose the right frame is set to the mean of histogram difference. For each video, we collect necessary information for this evaluation such as the number frames and actual frames of a video. Due to the nature of lecture videos, most of lecture videos contains several successive frames whose content is duplicated. Thus, selecting the right frame or keyframe is a main part of this evaluation. Table II depicts the results of sample lecture videos. The figures show that the frame duplication has been eliminated by applying our approach. There are few videos whose keyframes are lost (e.g. the extracted frame of video 7). However, the content of the obtained keyframes till reflects the entire content of a video, and the missing keyframes do not much affect the process of video searching.

TABLE II
The results of frame extraction

Lecture video	Total frame	Frame ratio (frame/second)	Actual frame	Extracted frame
Video 1	12,146	30	16	17
Video 2	13,973	30	18	18
Video 3	20,068	30	28	28
Video 4	20,039	29	27	27
Video 5	17,645	30	28	28
Video 6	29,827	30	37	37
Video 7	17,162	29	31	29
Video 8	17,978	30	24	26
Video 9	14,620	29	30	30
Video 10	16,080	30	31	31

The evaluation of keyframe extraction on 150 lecture videos is measure by the precision, recall and F-score of 95.22%, 82.83%, and 88.59% respectively where the true positive (TP) is defined by the number of keyframes correctly extracted, the false negative (FN) is the number of keyframes in actual frame which doesn't present in the extracted frame, and the false positive (FP) is the number of extracted frames which doesn't present in the actual frame. By this evaluation, we have observed that the proposed method presented in section III-A gives high performance in extracting keyframes from the lecture videos.

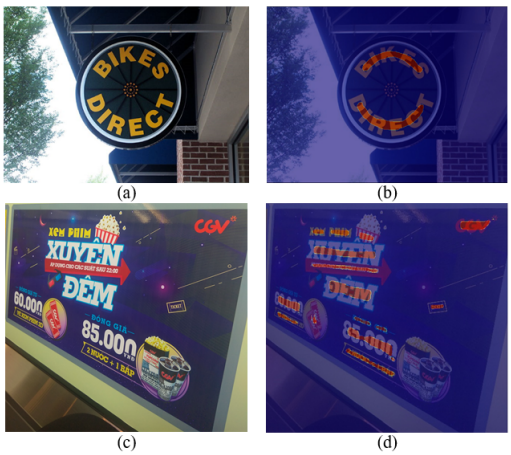
B. Text detection

Besides the VinText dataset, we estimate the text detection on other public datasets like ICDAR13 [9] and ICDAR15 [8] to compare the performance of our approach with the existing works. These datasets consist of high resolution images, and the text embedded in the images is in English.

1) Qualitative results: The heatmap generated by our approach is depicted in Figure 4 (b) and (d) . The proposed approach for text detection is capable of identifying the text regions with complex background and varying orientation. The generated heatmaps are then used to compute the bounding boxes surrounding the text content. We have observed that the proposed network trained on the VinText dataset gives high probability in segmenting text and non-text areas within an image. In addition, this approach gives high performance to identify text in English and Vietnamese language.

The other heatmaps generated from a keyframe of the lecture video are illustrated in Figure 5. We have observed that the proposed network has ability to detect the text content appearing in the lecture videos in most of the cases.

2) Quantitative results: We present the effectiveness of our approach by evaluating the proposed text detection on three public datasets. For ICDAR13 and ICDAR15 dataset, we compare the results of text detection with typical existing



(a) (b) (c) (d)

Fig. 4. The generated heatmap: (a) and (c) are the original images; (b) and (d) are generated heatmap.

(a) (b)

Fig. 5. The heatmaps of lecture video: (a) and (b) are polygon (green line) and bounding boxes (red line) surrounding the text content.

approaches, and the results of evaluation corresponding to these datasets are depicted in Table III and Table IV.

TABLE III
THE EVALUATION OF TEXT DETECTION ON THE ICDAR13 DATASET

Approach	Precision (%)	Recall (%)	F-score (%)
Zhu and Uchida [26]	83.00	84.00	84.00
Zhang *et al.* [24]	78.00	88.00	83.00
EAST [25]	84.86	74.24	79.20
Our approach	**89.47**	84.20	**86.76**

TABLE IV
THE EVALUATION OF TEXT DETECTION ON THE ICDAR15 DATASET

Approach	Precision (%)	Recall (%)	F-score (%)
Deng *et al.* [6]	82.89	81.65	82.27
EAST [25]	84.64	77.22	80.76
Our approach	**89.35**	79.80	**84.31**

The figures show that the proposed approach outperforms the existing works by an F-Score of 86.76%, 84.31% corresponding to ICDAR13 and ICDAR15 dataset. We have observed that the positions of located text are precise in most of the cases. However, our text detection approach might fail in cases such as the small logo can be misclassified as text content; the small text content is missed; or part of the text is missed because of low resolution.

In addition, we have conducted the experimentation of our approach on the VinText dataset which is Vietnamese scene text images. The precision, recall, and F-Score of this evaluation are 89.84%, 84.65%, and 87.17% respectively. Detecting the text regions in the scene image is more difficult than that of the lecture videos. Because the lecture videos often contain less complicated background, and the contrast between text and background is low. By experiment, we have observed that the proposed method well identifies the text regions in the lecture videos, which is illustrated in Figure 5. The performance of detecting text in the lecture videos is also evaluated by using the standard measures. Specifically, the values corresponding to the precision, recall, and F-score are 95.10%, 87.34%, and 91.05%.

C. Text recognition

By implementing the process of text detection presented in section III-B, the extracted text regions are depicted in Figure 6. The extracted text regions are then fed into the network presented in section III-C for text recognition.

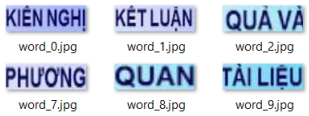


Fig. 6. The text regions extracted from the lucture videos.

We also compare the performance of the VietOCR with other approaches like tesseract OCR and EasyOCR. This comparison is conducted on 200 text regions extracted from the lecture videos, and the results are presented in Table V. The figures show that the VietOCR outperforms the others in recognizing Vietnamese language, and it is also good in recognizing the texts in English.

TABLE V
THE COMPARISON OF TEXT RECOGNITION

Approach	Precision (%)	Recall (%)	F-score (%)
TesseractOCR	84.65	89.84	87.17
EasyOCR	91.15	86.37	88.70
VietOCR	**95.46**	89.65	**92.46**

D. Semantic search

The text content recognized in section IV-C is organized and stored in a JSON format. To speed up the process of searching and ranking the lecture videos, we integrate the semantic search of ElasticSearch into MongoDB. The MongoDB is in charge of storing the content of lecture video, which is under the JSON format. Meanwhile, the ElasticSearch is responsible for receiving the query from the users and building the index according to the defined structure. In this work, we develop a web-based lecture video retrieval system in which all components like frame extraction, text detection, text recognition, and video retrieval are integrated. The appearance of lecture video retrieval is depicted in Figure 7.

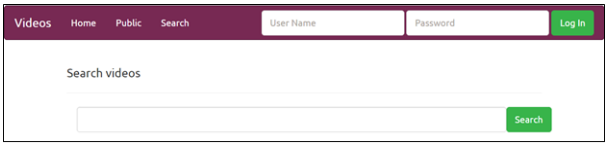


Fig. 7. The appearance of lecture video retrieval.

Various text queries are taken for estimating the precision, recall, and F-score. The average values corresponding to these measures are 99.14%, 96.05% and 97.57%. Unfortunately, we have not found the existing work for Vietnamese lecture video retrieval, so we can't perform the quantitative comparison with the others. Figure 8 depicts an exemplary result of lecture video searching and ranking. The input is a text query of the user. The right column illustrates the retrieved lecture videos whose content is relevant to the input text query, and the system is able to retrieve the videos in a very fast manner.

Fig. 8. The illustration of lecture video searching and ranking.

V. Conclusion

We have proposed various stages for lecture video retrieval. First, the median binary pattern is used to extract keyframes to meet a two-fold task: the extracted keyframes can cover the content of whole video sequences; and the duplication of keyframes is minimal. This method is conducted by measuring the characteristic difference between adjacent frames. Second, the segmentation-based text detection is developed to identify the text regions within the extracted keyframes. This approach is based on the principle of fully convolutional networks in which we use adaptive threshold and dilated convolution to optimize the binarization process for heatmap generation, which is utilized to compute the bounding box surrounding the text regions, and optimize the loss function of the network. Third, we make use of VietOCR to recognize text from the detected text regions. Lastly, we integrate the semantic search of ElasticSearch to MongoDB for optimizing the process of lecture video searching and ranking. The proposed approach gives high performance in terms of time complexity, video retrieval, text detection and recognition. The keyframe extraction, text detection and recognition are the primary research points in future work.

References

[1] https://www.uopeople.edu/blog/best-educational-youtube-channels-for-college-students/.
[2] http://vietocr.sourceforge.net/usage.html.
[3] Y. Baek, B. Lee, D. Han, S. Yun, and H. Lee. Character region awareness for text detection. *CVPR*, 2019.
[4] V. Balasubramanian, S. D G, and N. Kanakarajan. A multimodal approach for extracting content descriptive metadata from lecture videos. *Journal of Intelligent Information Systems*, 02 2015.
[5] B. Daga, A. Ghatol, and V. M. Thakare. Semantic enriched lecture video retrieval system using feature mixture and hybrid classification. *Advances in Image and Video Processing*, 2017.
[6] D. Deng, H. Liu, X. Li, and D. Cai. Pixellink: Detecting scene text via instance segmentation. *CoRR*, 2018.
[7] Y. Hao, T. Mu, R. Hong, M. Wang, N. An, and J. Y. Goulermas. Stochastic multiview hashing for large-scale near-duplicate video retrieval. *IEEE Trans. on Multimedia*, 2017.
[8] D. Karatzas, L. Gomez-Bigorda, A. Nicolaou, S. Ghosh, A. Bagdanov, M. Iwamura, J. Matas, L. Neumann, V. R. Chandrasekhar, S. Lu, F. Shafait, S. Uchida, and E. Valveny. Icdar 2015 competition on robust reading. In *Int. Conf. Document Analysis and Recognition (ICDAR)*, 2015.
[9] D. Karatzas, F. Shafait, S. Uchida, M. Iwamura, L. G. i. Bigorda, S. R. Mestre, J. Mas, D. F. Mota, J. A. Almazàn, and L. P. de las Heras. Icdar 2013 robust reading competition. In *Int. Conf. on Document Analysis and Recognition*, 2013.
[10] M. Liao, B. Shi, and X. Bai. Textboxes++: A single-shot oriented scene text detector. *IEEE Trans. on Image Processing*, 2018.
[11] J. Lin, L.-Y. Duan, S. Wang, Y. Bai, Y. Lou, V. Chandrasekhar, T. Huang, A. Kot, and W. Gao. Hnip: Compact deep invariant representations for video matching, localization, and retrieval. *IEEE Trans. on Multimedia*, 2017.
[12] L. Medida. An optimized e-lecture video retrieval based on machine learning classification. 09 2019.
[13] N. Nguyen, T. Nguyen, V. Tran, M.-T. Tran, T. D. Ngo, T. H. Nguyen, and M. Hoai. Dictionary-guided scene text recognition. In *CVPR*, 2021.
[14] N. V. Nguyen, M. Coustaty, and J.-M. Ogier. Multi-modal and cross-modal for lecture videos retrieval. In *Int. Conf. on Pattern Recognition*, 2014.
[15] A. K. P M, R. Ambati, and L. Raj. *An Efficient Scene Content-Based Indexing and Retrieval on Video Lectures*. 2021.
[16] N. Poornima and B. Saleena. An automated approach to retrieve lecture videos using context based semantic features and deep learning. *Journal of Engineering and Applied Sciences*, 2020.
[17] E. Shelhamer, J. Long, and T. Darrell. Fully convolutional networks for semantic segmentation. *IEEE Trans. on PAMI*, 2017.
[18] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *Int. Conf. on Learning Representations, ICLR*, 2015.
[19] Z. Tian, W. Huang, T. He, P. He, and Y. Qiao. Detecting text in natural image with connectionist text proposal network. In *ECCV*, 2016.
[20] W. Wang, E. Xie, X. Li, W. Hou, T. Lu, G. Yu, and S. Shao. Shape robust text detection with progressive scale expansion network. *CoRR*, 2019.
[21] W. Wang, E. Xie, X. Song, Y. Zang, W. Wang, T. Lu, G. Yu, and C. Shen. Efficient and accurate arbitrary-shaped text detection with pixel aggregation network. *Int. Conf. on Computer Vision (ICCV)*, 2019.
[22] S. Waykar and C. Bharathi. Multimodal features and probability extended nearest neighbor classification for content-based lecture video retrieval. *Journal of Intelligent Systems*, 2016.
[23] H. Yang and C. Meinel. Content based lecture video retrieval using speech and video text information. *IEEE Trans. on Learning Technologies*, 2014.
[24] Z. Zhang, C. Zhang, W. Shen, C. Yao, W. Liu, and X. Bai. Multi-oriented text detection with fully convolutional networks. In *CVPR*, 2016.
[25] X. Zhou, C. Yao, H. Wen, Y. Wang, S. Zhou, W. He, and J. Liang. East: An efficient and accurate scene text detector. In *CVPR*, 2017.
[26] A. Zhu and S. Uchida. Scene text relocation with guidance. In *Int. Conf. on Document Analysis and Recognition (ICDAR)*, 2017.