MANAS Journal of Engineering
ISSN 1694-7398 | e-ISSN 1694-7398

Volume 11, Issue 2, (2023) Pages 229-234
https://doi.org/10.51354/mjen.1369636

Extracting book titles from book recommendation videos using a deep learning approach

Bartu Sarımehmetoğlu[1,*], Hamit Erdem[2]

[1] Başkent University, Institute of Science, Department of Electrical and Electronics Engineering, Ankara/Turkiye
bartusari96@gmail.com, ORCID: 0000-0002-4778-0580
[2] Başkent University, Institute of Science, Department of Electrical and Electronics Engineering, Ankara/Turkiye
herdem@baskent.edu.tr, ORCID: 0000-0003-1704-1581

ABSTRACT

Extracting text from images and videos is an emerging field of research with a wide range of applications, including video search, video editing, and translation. Nowadays, book promotion videos in different languages are shared on social media and especially on YouTube. In this study; It is recommended to take book titles through book promotion videos. The developed system takes video as input and separates the names of the books. The viewer can select the desired book by clicking on the detected book titles and watch the relevant part of the video. This application result in time saving by the viewer. In order to achieve this application, a deep learning-based system was developed to retrieve the names of books from videos. YOLO-based method was used in the study. Different YOLO algorithms were used in the study, and YOLOv5 was found to be more successful. This study contributes to the field of text extraction and video analysis by developing a deep learning-based approach to extract book titles from book promotion videos.

ARTICLE INFO

Research article

Received: 1.10.2023
Accepted: 21.11.2023

Keywords:
Book promotion videos,
book titles,
deep learning,
YOLOv5,
time saving

*Corresponding author

1. Introduction

With the rapid development of visual media, obtaining different data from visual image and video media is increasing. Part of the work in this field is to obtain texts in the form of text from videos and images. Text information obtained from the environment can be used in areas such as image searching, mobile robot applications, instant translation, and industrial automation [1]. In the texts obtained from the video and image; players' names, venue names, warning signs, advertising signs, traffic-related explanations and other important information can be extracted [2]. Text extraction from images involves three main steps: detection, localization, and recognition. These steps can be performed individually or end-to-end. In recent studies, deep learning methods are also used in extracting text from images along with other methods [3,4,5]. Automatically extracting book titles from book trailers can help viewers find books they are interested in more quickly and easily. This can be achieved using deep learning methods. Similar studies have been conducted on determining the book title from the book cover or video. Histogram method was used in the study conducted by In Seop Na [6]. In this study, the pictures of 100 books were investigated, and the name of the book was found by extracting meaning from the book cover. In another study, images of book covers were used as input in a similar manner [7]. Matlab software was used in the relevant study. As in other fields, deep learning method has been applied to extract the book title [8]. In the study, a dataset consisting of 10 books was used for the "Optical Character Recognition (OCR)" application. In recent studies [9], the image matching method was applied. 1400 images obtained from 200 books were used in the study.

In this study, a YOLO (You Only Look Once) based method, one of the deep learning methods, was used. The developed software takes the video promotional video as input and quickly lists the names of the books in the promotional video under the video. The viewer can select the book they are interested in and watch only the relevant part of the video. This allows them to use their time efficiently without having to watch the entire video. This process aims to highlight the important content of the video and eliminate the need for users to scan or watch long videos.

Social media is now a popular platform for video book promotions in many different languages. The person who is interested in the subject has to watch the entire video without knowing how many books and which books are in these videos. In this study, it writes the names of the introduced books at the bottom of the video as an option. The process performed is a deep learning-based text extraction process from the video. The user can watch only the relevant part of

the video by clicking on the name of the book that he is interested in. This saves time instead of watching the entire video. To perform the process, text extraction from video based on YOLOv4, YOLOv5 and YOLOv7 was performed and the success of the process was tested according to standard success criteria.

2. Problem definition and related work

Nowadays, video book promotions are widely carried out on social platforms in many different languages. Images taken from 4 different videos, as examples of these videos, are given in Fig. 1. These videos are used to convey the content and features of the books to the viewers. In these videos, the presenter briefly introduces approximately 5 to 10 books in approximately 25-30 minutes. For viewers, it is important to save time by being able to see which books are in these videos without watching the entire video and automatically going to the book they want to watch. Because, by using deep learning methods, studies such as extracting important information from video book trailers and automatically summarizing these videos can be done. Thus, viewers can reach the books they are interested in faster, make book selection easier, save time and improve their book reading experience.

Figure 1. *Images taken from 4 different book promotion videos on YouTube [10]*

Automatic summarization of video book trailers can enable viewers to access the books they are interested in faster, make book selection easier, and improve their book reading experience. Deep learning is a powerful tool for automatically summarizing video book trailers. Deep learning methods make sense of the content of videos using image processing, natural language processing and machine learning techniques. These techniques are used to detect objects and events in the video, analyze speech and text, and summarize the content of the video using this information. Some previous studies have proposed a method that uses image processing techniques to detect objects and events in video [11]. These studies produce summaries that summarize key scenes or conversations in the video. Some other studies have used natural language processing techniques to analyze speech and text in video. These studies produce video summaries that provide an overview of the content in a more general way [12]. In more complex studies; image processing, natural language processing and machine learning techniques were used together [13]. These studies produce summaries that more accurately and concisely summarize the content of the video.

2.1. Extracting text from video and image

Extracting text from images is a research field that has gained great importance today. Studies in this field offer the ability to automatically recognize and extract text in digital images and documents. This process involves a combination of image processing and optical character recognition (OCR) techniques. The process of extracting text from an image basically consists of image loading, pre-processing, character recognition and text extraction steps. Image processing techniques includes processes such as contrast enhancement, noise removal and edge detection to make text in images more distinguishable. Optical character recognition (OCR) methods recognize and extract text through steps such as character segmentation, character recognition and extraction of results. The issue of extracting text from the image is used in many application areas such as digitization of documents, research and text mining, and electronic document management. Studies carried out in this field focuses on obtaining more accurate and reliable text extraction results by using advanced algorithms, deep learning techniques and large datasets.

Deep learning techniques, which are among the methods used in text extraction from images, are an approach that can analyze complex data structures and produce high-performance results. For this reason, the use of deep learning methods in text extraction studies from videos and images is becoming increasingly common. Deep learning methods can achieve a higher success rate in extracting text from video and images than traditional methods. Therefore, deep learning methods form the basis of research in this field. For example; in the study [14], a deep learning-based system was developed to extract text from a video. This system was able to detect text in the video with an accuracy rate of over 95%. In the study [15], a deep learning-based system was developed to extract text from an image. This system was able to detect text in the image with an accuracy rate of over 98%. The development and improvement of these techniques aims to obtain more accurate and efficient text extraction results. Text extraction steps are as follows.

- Input image
- Location determination
- Verifying
- Segmentation
- Recognition
- Character sequence

3. Materials and methods

In this study, a YOLO-based study, one of the deep learning algorithms, was conducted to extract book names from the book promotional videos. These videos are collected from YouTube. The performance of the proposed system tested with unseen video during training phase. The study was carried out using Python language as software. The display card used in the study is GeForce RTX 3060.

3.1. YOLO-based image extraction

YOLO-based methods developed to obtain real-time and precise results, provide great success in the important task of extracting text from images. YOLO is a fast deep learning model that performs object detection and classification simultaneously. It divides the image into small cells and uses a customized CNN to detect objects in each cell. In this way, it can accurately identify and extract regions containing text by estimating the bounding boxes and class probabilities of objects. YOLO-based methods can improve text detection and extraction capabilities by training on a large dataset. The methods provide high accuracy, real-time performance and general applicability, making it possible to automatically process and understand text-based data.

3.2. YOLO and CNN comparison

YOLO is a fast and effective deep learning based classification method used for object detection and classification. YOLO is a CNN-based model and detects objects by analyzing the image one at a time. One of the key differences between YOLO and CNNs is that YOLO does object detection and classification in a single gateway. This makes YOLO faster and more efficient than traditional CNNs. Another important feature of YOLO is that it is based on a pre-trained CNN model. This allows YOLO to be trained with less data and perform better. Overall, YOLO is an effective method for fast and real-time object detection. By quickly and efficiently analyzing information from image data, YOLO can accurately detect and classify objects.

3.3. Training a model in YOLO

YOLO as a deep learning algorithm is an image-based object detection algorithm. It is used to detect objects YOLO provides an effective solution for fast and real-time object detection. Its basic idea is to analyze an image layer by layer and predict object bounding boxes and class probabilities. YOLO divides the image into many small cells and places a customized CNN in each cell to predict one or more objects. The predictions for each cell are object bounding boxes that start at the center of the cell and scale according to the dimensions of the cell. Along with each bounding box, probability values are also estimated for that object's possible class label. YOLO divides the image into a grid and for each grid cell, if there is an object in that cell; it estimates the object's class, coordinates and dimensions. In this way, multiple objects can be detected on a single image and these objects can be classified and positioned. YOLO can analyze an image for multiple objects simultaneously and predicts bounding box and class probabilities for each object. This gives effective results even when objects are close together or overlapping. Moreover, YOLO is an effective and fast deep learning method for real-time object detection. It performs multi-object detection by simultaneously estimating the bounding boxes and class probabilities of objects on the image. One of the biggest advantages of YOLO is that it works faster than other object detection algorithms. Various software was used in the study. This software and their role in the training system are presented in Table I.

Table I. *Software used in the proposed system*

Software Used	Its Role in the System
OpenCV	Image processing and machine learning
Torch-PyTorch	Deep learning
Open Images Dataset	Creating a data set
EasyOCR	Optical character recognition library (text detection)
CVAT	Extraction of unwanted data
Google Colab	Model training

3.4. Performance measuring functions

During the training process of proposed model, various loss functions (Box Loss, Object Loss, Class Loss, mean Average Precision) and performance criterias (Precision, Recall, F1 score) are used to evaluate the performance of the applied model. These function and metrics which are uses for similar classification are as follows.

- Box Loss: It is a loss function used in object detection problems. It is an optimization goal to improve the accuracy and precision of the model by measuring the difference between the actual and predicted bounding boxes.
- Object Loss: It is a loss function used in object detection problems. It is an objective function to optimize the model's ability to detect correct objects by evaluating the objectness of real and predicted objects.
- Class Loss: It is a loss function used in object classification problems. It is an optimization goal to improve the model's ability to predict correct classes by measuring the difference between true and predicted classes.

- mAP (mean Average Precision): It is a value used to measure the performance of object detection or object classification models. By evaluating the similarity of actual and predicted object bins, it combines the model's metrics such as precision and recall and measures overall performance.
- Precision: It is the ratio of positive examples predicted by a classification model among true positive examples. In other words, it is a metric that measures how accurately the model classifies true positives.
- Recall: It is the ratio of true positive samples among predicted positive samples. That is, it is a measure of how many the model correctly detected and how many it missed.

3.5. YOLOv5 and YOLOv7

YOLOv5 is a deep learning based learning model used for object detection. In this study, 1000 datasets of books were trained with the YOLOv5 model to detect the books and extract the titles of the detected books. The output graphic results of the training are presented in Fig. 2 and Fig. 3. YOLOv7 is the latest version of the YOLO series, YOLOv7 is specifically designed for real-time and fast object detection. This model is particularly advantageous in applications such as autonomous vehicles, security systems and object tracking. In this study, a dataset of 1000 books was trained with the YOLOv7 model. The output graphical results of the training are given in Fig. 4 and Fig. 5.

4. Results and discussion

In this study, a system was developed to automatically summarize and make book promotional videos more understandable using the YOLO algorithm. The book names in the book promotion videos were read and sorted, and then these book names were presented to the user in the form of an interface. You can get an idea about the video through this interface, and at the same time, when you click on any book on the interface, you can automatically and directly go to that book and get an idea about the book.

To evaluate the performance of the proposed system, randomly selected book promotion videos from YouTube were uploaded to the system and their performance was analyzed for different languages.

To train the model, a system compose of RTX3060 graphics card was used to meet the GPU needs of the deep learning algorithm. Each training took an average of 18 hours. 80% of the data set was used for training, 10% for testing and 10% for validation. The number of epochs is 99.

Then, different versions of CNN based YOLO algorithms (YOLO 4, YOLO 5, YOLO 7) used for training of the proposed model. A dataset of 1000 books was trained with different models under the same conditions. The system worked successfully when tested with all these versions. After training with YOLOv5 and YOLOv7 models, classification performance tested considering standard criteria. According to the results obtained, the YOLOv5 was more successful according to the precision and F1 score criteria. Additionally, the confusion matrixes related to classification results presented in Table II and III.

Table II. *Comparison of YOLOv5 and YOLOv7 models used in the study in terms of Precision, Recall and F1 Score*

Metrics	Precision $Precision = \frac{TP}{TP + FP}$	Recall $Recall = \frac{TP}{TP + FN}$	F1 Score $F_1 = 2 * \frac{precision * recall}{precision + recall}$
YOLOv5	%90	%80	%85
YOLOv7	%80	%80	%80

Table III. *Confusion matrix table for YOLOv5 and YOLOv7*

Actual Class	YOLOv5	YOLOv7
Positive	90	80
Negative	10	20

Table IV. *Comparison of YOLOv5 and YOLOv7 models used in the study as mAP_0.5 and mAP_0.5:0.95*

	mAP_0.5	mAP_0.5:0.95
YOLOv5	%87	%65
YOLOv7	%82	%62

Table V. *Comparison of YOLOv5 and YOLOv7 models used in the study in terms of Box_Loss, Obj_Loss and Class_Loss*

	Box_Loss	Obj_Loss	Class_Loss
YOLOv5	0.01	0.008	0
YOLOv7	0.03	0.011	0

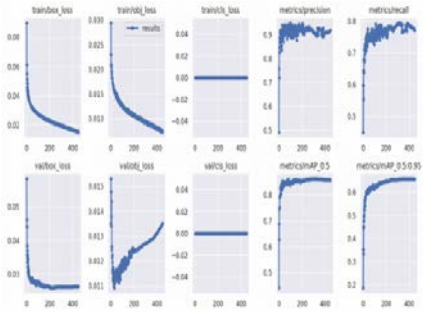

Figure 2. *YOLOv5 training results due to losses function and mAP*

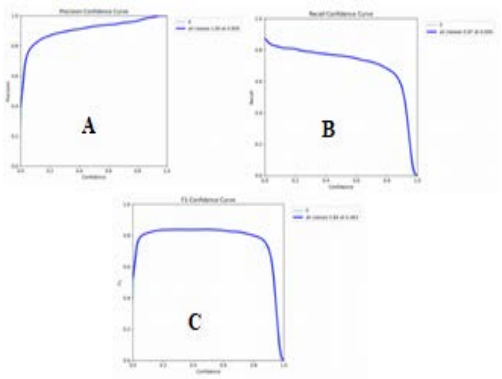

Figure 3. *YOLOv5 training results in Precision (A), Recall (B) and F1 Score (C)*

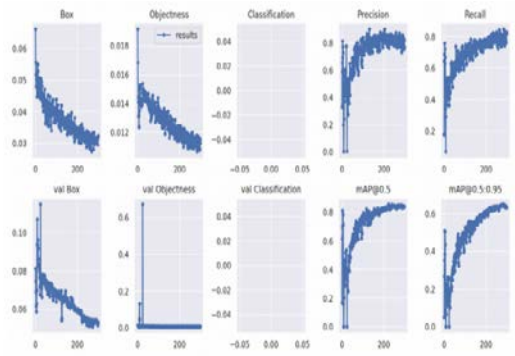

Figure 4. *YOLOv7 training results*

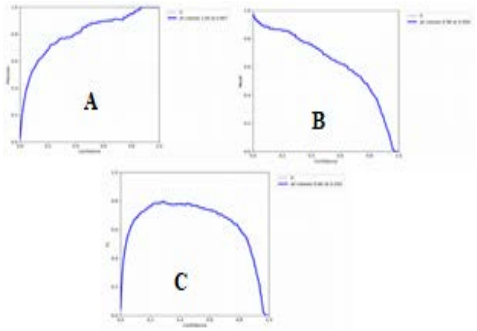

Figure 5. *YOLOv7 training results in Precision (A), Recall (B) and F1 Score (C)*

An example of the system output generated at the end of the study is shown in Fig. 6.

YOLOv5 exhibits less FN and FP error rates than YOLOv7. This means that YOLOv5 has the ability to more accurately predict true positives (TP) and true negatives (TN). This result shows that YOLOv5 is a more accurate classification model. After the training with the YOLOv5 and YOLOv7 models, the results obtained regarding these parameters are given in Table IV and Table V.

Based on the results, it can be said that the YOLOv5 model is better in terms of accuracy on the book dataset. As a final test, Table VI shows the results of the system for book promotion videos (in different languages). The results were interpreted based on the number of books found by the system in the video and the number of book names found correctly. The system managed to find books in the video with an average rate of 98%. This shows that the system is successful in detecting books. However, there is less margin of error in the system when finding book titles.

Table VI. *Performance of the system on randomly selected book promotion videos*

YouTube video address	Input videos	Detected books	Success rate	Performance of the system
https://www.youutube.com/watch?v=R2Zt-3spWFQ	14	13	%93	%100 **(English)**
https://www.youutube.com/watch?v=82vJPdsJyHU	17	16	%94	%100 **(Turkish)**
https://www.youutube.com/watch?v=_zCTZM92Ju0	10	10	%100	%100 **(Turkish)**
https://www.youutube.com/watch?v=M0qL4zzIuC8	12	12	%100	%100 **(English)**
https://www.youutube.com/watch?v=7d4s0kdQFeE	17	16	%94	%100 **(Turkish)**
https://www.youutube.com/watch?v=eojGtx9g0kg	6	6	%100	%100 **(German)**
https://www.youutube.com/watch?v=yHXuBmLAJN8	10	10	%100	%100 **(French)**
TOTAL	86	83	%97	%100

Figure 6. *Screenshot of the system (Titles of detected books). Source video from YouTube [10]*

References

[1]. He, J., Zhang, Z., Yang, W., Zhang, J., Zhang, Y., and Zhang, H. (2018). Scene Text Detection and Recognition: The Deep Learning Era. International Journal of Computer Vision, 129(7), 1641-1671.

[2]. Chen, D., Odobez, J.-M., and Bourlard, H. (2014). Text detection and recognition in images and video frames. IEEE Transactions on Pattern Analysis and Machine Intelligence, 37(3), 697-711.

[3]. Busta, M., Neumann, L., and Matas, J. (2017). Deep textspotter: an end-to-end trainable scene text localization and recognition framework. In Proceedings of the IEEE international conference on computer vision (pp. 2204-2212).

[4]. Liu, X., Meng, G., and Pan, C. (2019). Scene text detection and recognition with advances in deep learning: a survey. International Journal on Document Analysis and Recognition, 22, 143-162.

[5]. Chu, W.-S., Lin, Y.-H., & Yang, Y.-H. (2012). Title extraction from book cover images using histogram of oriented gradients and color information. International Journal of Contents, 8(4), 95-102.

[6]. Na, I. S. (2016). A Novel Scene Text Detection Method Using Histogram-Based Thresholding. IEEE Transactions on Image Processing, 25(2), 675-689.

[7]. Chen, Y., Liu, J., and Yu, K. (2018). Book Cover Image Recommendation Using Deep Learning. In Proceedings of the 2018 ACM Multimedia Conference on Multimedia Information Retrieval (MIR) (pp. 530-537).

[8]. Al-Omari, A. A., Al-Duwairi, M., and Al-Omari, K. (2020). Book Title Extraction from Book Cover Images Using Deep Learning. IEEE Transactions on Image Processing, 29(5), 2540-2551.

[9]. Akar, H. A., Al-Omari, A. A., and Al-Omari, K. (2022). Book Title Extraction from Book Cover Images Using Image Matching. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(10), 2726-2739.

[10]. https://www.youtube.com

[11]. Li, Y., Fu, X., and Wang, J. (2019). A Survey on Object Detection and Tracking in Videos. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(11), 2789-2812.

[12]. Li, Z., Zhang, K., Mei, T., and Rui, Y. (2021). A Survey on Video Summarization Techniques. IEEE Transactions on Multimedia, 23(1), 281-304.

[13]. Li, H., Wang, Y., and Wang, W. (2022). A Survey on Multimodal Machine Learning for Video Understanding. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(10), 2711-2725.

[14]. He, J., Zhang, Z., Yang, W., Zhang, J., Zhang, Y., and Zhang, H. (2022). Deep Learning for Video Text Extraction: A Survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(10), 2699-2710.

[15]. He, J., Zhang, Z., Yang, W., Zhang, J., Zhang, Y., and Zhang, H. (2022). Deep Learning for Text Extraction from Images: A Survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(10), 2684-2698.