Book Segmentation and Identification using Segment Anything Model

Rujeet Jahagirdar
Dept. of Computer Science
University of Texas at Arlington
Arlington, USA

***Abstract*—The aim of this project is to automate the process of segmenting individual books from an image of a bookshelf and detecting the title of each book. Additionally, we aim to detect the orientation of each book, which can be useful in identifying unorganized books. To achieve this, we divided the problem into three individual sub-problems: book segmentation, orientation detection, and title detection. In this project, we propose a neural network-based approach for book segmentation and use Hough Transform for detecting the orientation of the books. For title detection, we used an optical character recognition (OCR) technique. Our experimental results demonstrate that our proposed approach achieved high accuracy in book segmentation and orientation detection, with some limitations in title detection. This work can potentially contribute to improving the efficiency and accuracy of book organization in libraries and bookstores.**

I. Introduction

Book shelf management is the crucial task in library and bookstore management. The traditional method of manually managing and cataloging the books is very time consuming and requires a lot of manual efforts. With the advancement of deep learning techniques, this task can be automated by using neural networks.

In this paper, we experimented with the recently developed foundation model "Segment Anything Model" for book segmentation task and used a pre-trained model "TrOCR" for Optical Character Recognition to identify books by their titles.

The proposed method has the potential to significantly reduce the manual effort required in cataloging and organizing books in libraries and bookstores. For future enhancements this method can be improvised to work in real time and can be combined with the automated robots such as drones to scan and catalog book shelves without human interventions.

A. Segment Anything Model

The Segment Anything Model (SAM) is a segmentation model that aims to segment an object of interest based on prompts given in various forms such as a point, a set of points, a bounding box, or text. SAM is designed to generate a valid segmentation mask even when there is ambiguity in the prompt. The model learns the concept of an object and can segment any object that is pointed out. This feature enables the model to perform well in the zero-shot learning regime, i.e., high performance without additional training on new types of objects. SAM's unique architecture and a large dataset were used to develop this capability. The model was trained progressively alongside the development of the dataset, which was annotated in three stages. Initially, human annotators annotated a set of images by clicking on objects and manually refining the masks generated by SAM. Next, annotators were asked to segment masks that SAM did not confidently generate, thereby increasing the diversity of objects. The final set of masks was generated automatically by prompting the SAM model with a set of points distributed in a grid across the image and selecting confident and stable masks.

B. Hough Line Transform

Hough Line Transform is a computer vision technique which is widely used for line detection in images. It transforms the lines in the image space to the parameter space and detects the lines using accumulator matrix by using voting.
The Hough Line Transform works by representing a line in an image as a point in a parameter space. Each point in the parameter space corresponds to a unique line in the image. The parameter space is defined by two parameters: the slope of the line and the y-intercept of the line.

To detect lines in an image using the Hough Line Transform, the algorithm first applies an edge detection algorithm, such as the Canny edge detector, to the image. The edge points are then converted to the Hough parameter space using the equation of a line in slope-intercept form. Each edge point corresponds to a sinusoidal curve in the parameter space. The intersection of the sinusoidal curves represents the parameters of the lines in the image.

II. RELATED WORK

The problem of book segmentation and identification has been widely studied in the field of computer vision and image processing. Various techniques have been proposed to address this problem. In this section, we will review some of the recent works related to this project.

One of the works related to book segmentation is "Detecting Book Pages from a Library Shelf Using Image Processing" by Wang et al. In this work, the authors propose a method to detect book pages from a library shelf. They use a combination of image processing techniques such as edge detection, thresholding, and morphological operations to detect the book pages.

In another work, ”Book Spine Detection and Recognition for Shelf Inventory Management” by Marai et al., the authors propose a method to detect book spines and recognize their titles. The method is based on image processing techniques such as color segmentation, edge detection, and template matching.

More recently, deep learning-based methods have been proposed for book segmentation and identification. In ”Book-CoverNet: A Deep Learning Network for Recognizing Books,” Zhang et al. propose a deep learning model for recognizing book covers. The model is based on a convolutional neural network (CNN) architecture.

Another deep learning-based method is proposed in ”Reading Books with Convolutional Neural Networks” by Gomez-Bigorda et al. In this work, the authors propose a CNN-based method to detect and recognize text in book pages.

In our work, we propose a method for book segmentation and identification using neural networks. Our method is based on a Segment Anything model and involves finding the masks of the segmented images. We also explore the use of transfer learning to identify the text printed on the book spine.

III. PROBLEM STATEMENT

The problem addressed in this project is to automate the segmentation of individual books from the image of a book-shelf and detect the title of each book. Additionally, we aim to detect the orientation of each book, which can be used to identify unorganized books. This problem has been divided into three individual sub-problems: firstly, segmenting each book from the image of the book-shelf; secondly, detecting the orientation of each book; and thirdly, detecting the title of each book. The objective is to develop an accurate and efficient solution that can handle variations in book sizes, orientations, and lighting conditions. This project aims to contribute to the field of computer vision and improve the process of book inventory management in libraries and bookstores.

IV. PROBLEM SOLUTION

I have experimented to solve this problem by dividing it into three smaller tasks: segmentation task, orientation detection task and text recognition task.

A. Segmentation Task

What is segmentation task, why it is required ?

Segmentation is the task to partition different objects/regions of the given input image. It is used for a wide range of tasks, such as object recognition, tracking, and classification, as well as image compression, enhancement, and restoration. In this project I have used segmentation technique to detect individual book from the collection of books.

Input of this project is the image of book shelf. This image will have number of books stacked onto the shelf in different orientations. Along with books, these images will also have some noise such as shelf seperator, background. So my first task was to try to determine and seperate out the each individual book from the book shelf. I have tried to use recently released ”Segment Anything Model” for this task. This functional model was recently released by Meta (Facebook) and is specifically designed to segment out different objects in the image. It was trained on dataset called ”SA-1B” which consists of 1 billion images.

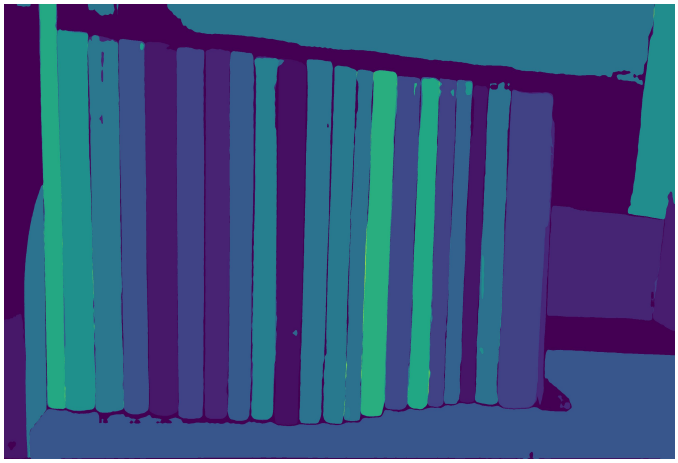

Fig. 1. Segmentation Output 1

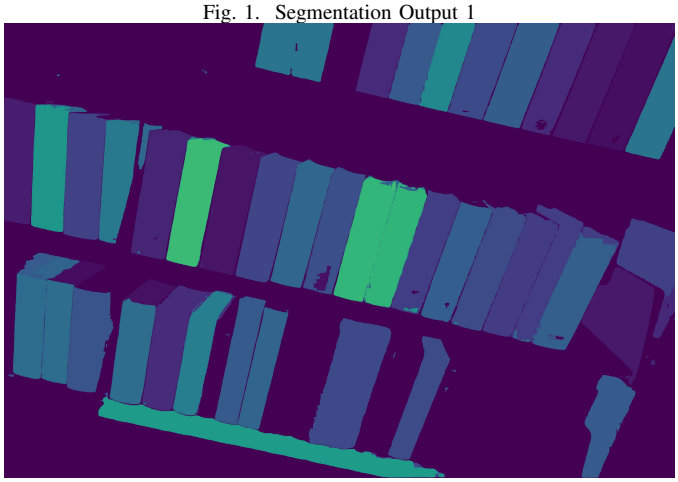

Fig. 2. Segmentation Output 2

Output and Future work

As seen in the below output images, out-of-the-box SAM model can accurately detect and segment the individual book from the book shelf. But it is also segmenting the backgrounds. This is because the SAM model was trained to detect any general object from the image. To improve this model’s accuracy to segment only books, we need to fine tune this model for our specific task. In future work, I will be fine tuning this model for annotated books dataset.

B. Orientation Detection Task

What is orientation detection task, why is it required ?

In a neatly organized book shelf, all the books are alined

in perfectly vertcial orientation. Whereas in our dataset there are some book images which are not perfectly vertical, I will use this measure as a deciding factor if the book shelf needs reorganization or not.

How have you done this task ?

So to determine the orientation of the book, we can use line detection techniques which will helps us to get the slope of the book. If this slope is too much deviating from 90degrees which is perfectly vertical line, then we can conclude that the book needs to be reorganized.

There are many different techniques to detect lines like Hough Transform, Canny Edge detection, Soble Operator for edge detection. In this project I have used Hough Transform technique.

What is hough transform and How hough transform works ?

In Hough Transform technique, the points in the image space are converted into the parameter space. Parameter space is the two dimensional space with parameters of the image space as its co-ordinates. For line detection, the two parameters, slope and y-intercept of the line are used to repesent it in the parameter space as a point. For all the edges detected by the edge detection, a corresponding point is plotted in the paramter space. For all such points in parameter space, a set of possible lines is determined and an accumulator matrix is calculated. This matrix is used to find the maximum voted parameters for the corresponding line in the image space.

Results and Future Work

Using Hough Transform I got the following results as shwon in Fig.1 and Fig.2. As seen in the Fig. 2, the results are not consistent. There could be multiple reasons like text book color and background color, brightness and contrast.

To improve the results of the line detection, I'm planning to use different image preprocessing techniques and also experiment with different line detection techniques.

C. Text Recongnition Task

What is Text Recongnition ?

Text recognition is the process of automatically recognizing text from an image or a document and converting it into a digital text format that can be edited, searched, and analyzed. This task involves the use of various machine learning techniques such as deep learning, computer vision, and natural language processing to identify and extract text from an image. The text recognition process typically involves several steps. First, the image is preprocessed to enhance the text and remove any noise or background interference. Then, the text is detected and localized in the image using techniques such as object detection or edge detection. Once the text regions are identified, the characters are recognized and converted into digital text using optical character recognition (OCR) algorithms.

In this project I have used "TrOCR: Transformer-based Optical Character Recognition with Pre-trained Models". This algorithms work by analyzing the shapes and patterns of

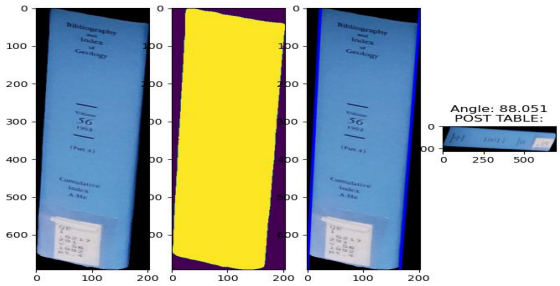

Fig. 3. Orientation Detection output 1

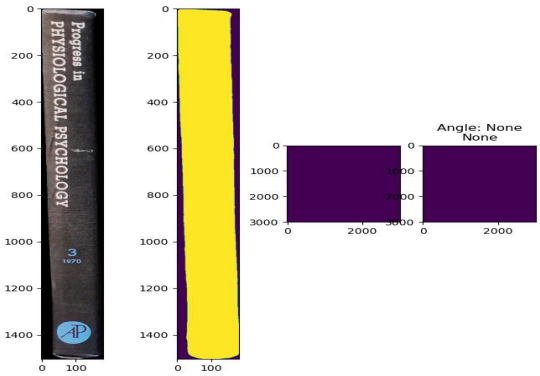

Fig. 4. Orientation Detection output 2 - No lines detected

individual characters within an image of text and using machine learning techniques to match them to known character sets. This allows for the automatic extraction of text from documents, which can then be further processed and analyzed by computers.

Results and Next steps ?

I have used the TrOCR model out-of-the-box which was trained on general dataset, so the output of this step is not as expected. Size of the image, contrast of the image, model training are some of the reasons due to which I got poor results.

For improving this results, I will be preprocessing the image to adjust brightness and quality of the image. Further, I will be training this OCR model for my specific dataset of book images which could help to improve these results.

V. CONCLUSION

In this project, I have implemented the book segmentation and identification technique using Segment Anything Model. I found that the accuracy of the SAM model for book segmentation is moderate and it could be improved by fine

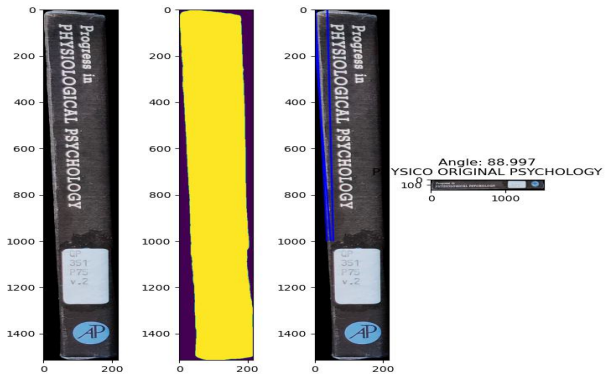

Fig. 5. Output of the Text Detection 1

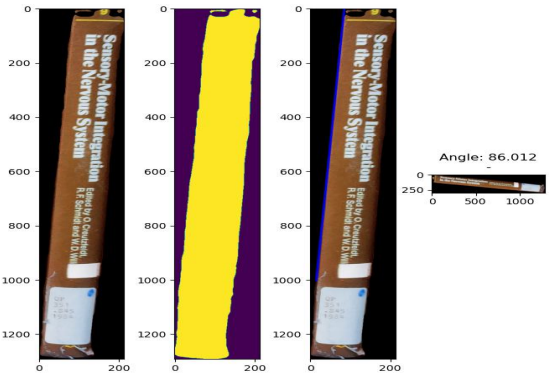

Fig. 6. Output of the Text Detection 2

tuning this model to specific dataset of the book.

REFERENCES

[1] Minghao Li, Tengchao Lv, Jingye Chen, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. TrOCR: Transformer-based Optical Character Recognition with Pre-trained Models.

[2] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. "Segment Anything." *arXiv preprint arXiv:2304.02643* (2023): arXiv:2304.02643 [cs.CV].

[3] Dongjie Cheng, Ziyuan Qin, Zekun Jiang, Shaoting Zhang, Qicheng Lao, and Kang Li. "SAM on Medical Images: A Comprehensive Study on Three Prompt Modes," *arXiv preprint arXiv:2305.00035*, 2023.

[4] Hough Line Transform (https://docs.opencv.org/3.4/d9/db0/tutorial$_hough_lines.html$)