I.T.C

Content list available at ITC

Techno-Science Research Journal

Journal Homepage: http://techno-srj.itc.edu.kh/

Techno-Science Research Journal

Word Spotting on Khmer Palm Leaf Manuscript Documents

Vannkinh Nom[1*], Dona Valy[2], Sokkhey Phauk[3], Seng Hak Leng[3]

[1] *Department of Information and Communication Engineering, Institute of Technology of Cambodia, Russian Federation Blvd., P.O. Box 86, Phnom Penh, Cambodia*

[2] *Research and Innovation Center, Institute of Technology of Cambodia, Russian Federation Blvd., P.O. Box 86, Phnom Penh, Cambodia*

[3] *Department of Applied Mathematics and Statistics, Institute of Technology of Cambodia, Russian Federation Blvd., P.O. Box 86, Phnom Penh, Cambodia*

Received: 17 July 2023; Accepted: 07 August 2023; Available online: June 2024

Abstract: *Word spotting plays a crucial role in document analysis, particularly for ancient palm leaf manuscripts. Khmer palm leaf manuscripts, which are written on rectangularly cut and dried palm leaf sheets, hold significant cultural value in Cambodia. These manuscripts contain valuable historical, religious, and linguistic information, making their preservation essential. However, extracting information from them is challenging due to their fragility, age, and the complexity of Khmer writing and word formation. This study focuses on word spotting and investigates the construction of a Region Proposal Network (RPN) using the You Only Look Once (YOLO) technique and Convolutional Neural Network (CNN) for the accurate and efficient identification of specific words or phrases within the documents. The proposed method is evaluated using the SleukRith dataset, which consists of 1,971 images of Khmer palm leaf manuscripts. Among these, 1,379 images are allocated to the training set, 395 to the test set, and approximately 197 to the validation set. Parameter tuning is conducted on two variables: the number of layers and the number of filters. The results demonstrate that the optimal model comprises 3 layers and 24 filters, with a threshold of 0.4. The achieved detection performance accuracy is approximately 80.86%, while the classification performance reaches 69.29% for the 33 classes of Khmer characters.*

Keywords: Text detection; Text classification; You Only Look Once (YOLO); Convolutional Neural Network (CNN)

1. INTRODUCTION

The palm leaf manuscripts written in the Khmer language constitute a crucial element of Cambodia's cultural heritage, containing valuable information for researchers and historians. However, accessing this information can be challenging for the general public due to factors such as the manuscripts' distinct writing system, instances of damage, and the dense layout of text, all of which can hinder comprehension and the identification of the document's contents. Valy et al. [10] mentioned that one of the major difficulties in working with the Khmer language is the complexity of its characters and word forms. Khmer stands out from other languages in several aspects, including the positioning of vowels, which can be placed to the left, right, above, or below consonants. Additionally, it features the merging of two or more consonants into different shapes known as low-consonants or subscripts, which are located beneath the main consonant.

Valy et al. [11] reported that the Khmer language is widely regarded as one of the most complex languages globally, primarily due to its alphabet's extensive use of visually similar symbols. This complexity poses a significant challenge for researchers and historians eager to delve into this invaluable cultural heritage. Consequently, recognizing and extracting text from Khmer palm leaf manuscript documents requires specialized methods and equipment, making it a complex endeavor.

Several previous investigations have explored word spotting models; however, some of them have employed varying methodologies and datasets. In a recent study by Ronen et al. [8], an innovative deep learning-based technique was introduced to address the complexities associated with detecting irregular and multi-oriented text in cluttered backgrounds. The effectiveness of this approach was evaluated on four widely-used benchmark datasets: ICDAR 2013, ICDAR 2015, SCUT-CTW1500, and Total-Text. Almazan et

[*] Corresponding author: Vannkinh Nom
E-mail: vannkinhnom123@gmail; Tel: +855-92 737 118

al. [1] proposed a framework for word spotting and identification in scene images, aiming to address similar challenges. Their framework offers valuable insights into the field of word spotting and identification. The approach includes attribute embedding, word spotting, and recognition modules, which were evaluated on benchmark datasets. The attribute embedding module trains a neural network to embed attributes into feature vectors used for word spotting and recognition. Dey et al. [3] demonstrated the effectiveness of Local Binary Pattern (LBP) for word spotting in handwritten historical documents. In some studies, LBP has been combined with other computer vision techniques, such as support vector machines (SVM) or k-nearest neighbors (KNN) classifiers, to enhance the accuracy of word spotting. LBP shows promise as an approach for word spotting in handwritten historical documents. The performance of this approach was evaluated using three commonly used benchmark datasets: The George Washington's Correspondence dataset, the Châteauroux Historical Document Image dataset, and the IAM Handwriting Database. Sudholt & Fink [7] demonstrated that Attribute CNNs can accurately locate text regions in handwritten documents. They emphasized the importance of combining multiple Attribute CNNs and employing additional techniques, such as Multiple Instance Learning (MIL) or graph-based inference, to enhance system performance using the IAM Handwriting Database. In a study by Daraee et al. [2], they proposed the development of an automated system for recognizing keywords in handwritten documents. The authors aim to enhance the accuracy of keyword spotting by combining the outputs of multiple deep neural networks (DNNs) and utilizing certainty prediction to assess the confidence of the DNNs in their predictions. They evaluate the proposed word spotting method on three datasets: IAM offline database, the George Washington database, and the Botany and Konzilsprotokolle database. In a study by Frinken et al. [4], a new approach to word spotting in handwritten documents is presented. The authors propose a solution that utilizes recurrent neural networks (RNNs) to overcome the limitations of traditional word spotting methods, which are often dependent on feature extraction and have high computational complexity. They utilize three datasets, namely the IAM offline dataset, George Washington dataset, and Medieval manuscripts. Rath & Manmatha [5] demonstrate that Dynamic Time Warping (DTW) effectively aligns and matches word images with different shapes, sizes, and orientations. The study compares DTW with traditional feature-based matching methods and finds that DTW outperforms the traditional methods in terms of accuracy and robustness. This work has laid the foundation for the widespread use of DTW in word image matching, establishing it as a powerful tool in the field of document analysis. Zhao et al. [12] introduce an effective end-to-end trainable technique for segmentation-free query-by-string word spotting in handwritten historical document images. They utilized three public benchmarks: the George Washington Dataset (GW), Konzilsprotokolle Dataset, and Barcelona Historical Handwritten Marriages Dataset (BH2M). According to a study by Redmon et al. [6], You Only Look Once (YOLO) is an approach to object detection that has the ability to predict the presence and location of objects at a glance. YOLO achieves more than twice the mean average precision of other real-time systems and considers the entire image during training and test time, implicitly encoding contextual information about classes and their appearance.

In this research work, we will develop a model for word spotting that can recognize text in Khmer palm leaf manuscript documents by utilizing the YOLO architecture with a Convolutional Neural Network (CNN). The outcome of this work will represent a significant advancement in preserving cultural heritage and creating new opportunities for further research and exploration in this field.

The paper is structured into four sections. The first section provides an overview of the research, followed by the second section which details the methodology used. Meanwhile, the third section showcases the results and includes a discussion of the findings. Finally, Section four offers a conclusion.

2. METHODOLOGY

2.1 Dataset

In this research study, we will make use of the SleukRith set introduced by Valy et al. [9], which consists of 657 pages of palm leaf manuscripts that have been digitized from the palm leaf image corpus, as shown in Fig. 1. To simplify the process of locating words within the long images, we propose dividing each original image into three parts for both jpg and xml formats. This will result in a total of 1971 images in both jpg and xml formats. As part of our research, we categorized the data into three sets: a validation set comprising 10% of the data, a testing set comprising 20% of the data, and a training set comprising 70% of the data.

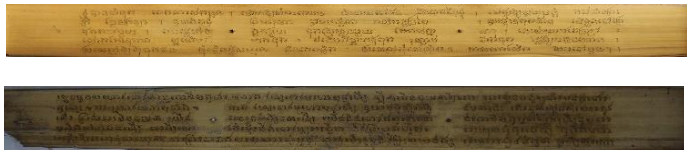

Fig. 1. SleukRith set

In order to address the challenge posed by the large size of the initial dataset, we propose dividing the image into three equal sections through cropping, which can facilitate the learning process of our model, as presented in Fig. 2. The dataset comprises pairs of images and XML files, with the XML files containing vital information such as characters, words, lines, and coordinates that describe the image's content and the location of each character. This information is crucial for identifying the characters and their respective positions within the image. The XML file contains multiple points in the form of polygons that represent characters or words. Our task is to analyze the polygon data and determine four parameters: x_min, y_min, x_max, and y_max, which will be utilized to reestablish the rectangular bounding box for each character. After cropping the image and XML components, it is necessary to create a new dataset that links each image to its corresponding XML file. This ensures that the essential information is available for the accurate identification and localization of characters within the dataset.

Fig. 2. New SleukRith set

2.2 *YOLO architecture*

Within the architecture of YOLO, a single neural network is utilized to make predictions for both object class probabilities and bounding boxes, as introduced by Redmon et al. [6]. The YOLO network takes an image of a specific size as input, and the output is a grid containing class probabilities and bounding boxes for each cell. YOLO is designed to maintain high precision while supporting end-to-end training and real-time processing speeds. The input image is divided into a grid of size S x S, and if the center of an object falls within a grid cell, that cell is responsible for detecting the object. For each grid cell, the model predicts B bounding boxes and their respective confidence scores, which represent the confidence level of object presence within the box and the accuracy of the predicted box.

2.3 *Convolutional Neural Network*

We designed the model as a convolutional neural network and evaluated it using our SleukRith dataset, which includes 24 layers. The design of the model was inspired by Redmon et al. [6], who focused on object detection in the PASCAL VOC 2007 dataset, consisting of twenty object classes. The constructed model takes 3-channel input images and applies a sequence of convolutional layers, followed by the LeakyReLU activation function and max pooling operations for downsampling. It gradually increases the number of filters in the layers. The input images have a size of 416 x 1408 with 3 channels, and the model produces a tensor of predictions with a size of 38 x 13 x 88 as the final output. The model's architecture is illustrated in Fig. 3.

2.4 *Evaluation Metrics*

In this study, we will evaluate the effectiveness of the word spotting model using the F-measure score. We calculate the Intersection over Union (IOU) to measure the degree of overlap between the predicted bounding boxes and the actual boxes, based on the area of their union. The IOU score can be computed using Eq. 1. The score can range from 0 to 1, where a score of 1 indicates a complete match between the predicted and actual bounding boxes. A higher IOU score indicates precise object identification by the object detection algorithm, while a lower score suggests potential inaccuracies in the predictions.

$$IOU = \frac{\text{Area of Intersection}}{\text{Area of Union}} \tag{Eq. 1}$$

$$DR = \frac{o2o}{N} \tag{Eq. 2}$$

$$RA = \frac{o2o}{M} \tag{Eq. 3}$$

$$FM = \frac{2.DR.RA}{DR+RA} \tag{Eq. 4}$$

To determine the count of one-to-one matches, we only consider region pairs with an IOU score above a defined threshold of 0.5. Assuming N represents the number of bounding boxes in the ground truth and M indicates the number of bounding boxes detected by the approach, o2o signifies the count of one-to-one match pairs. Consequently, we can calculate the detection rate (DR) using Eq. 2 and the recognition accuracy (RA) using Eq. 3. The F-measure (FM) is determined by combining DR and RA according to Eq. 4.

3. RESULTS AND DISCUSSION

In the first experiment, we evaluated our model on Khmer characters, which consist of 33 classes. Our dataset comprised 1,971 images, with 1,379 assigned to the training set, 395 to the test set, and approximately 197 to the validation set. We conducted parameter tuning on two variables: the number of layers and the number of filters. The goal was to identify the optimal model that achieved the highest accuracy on the validation set. To accomplish this, we created a module list with different combinations of layer numbers (1, 2, 3) and filter amounts (8, 16, 24) randomly. Additionally, we experimented with threshold values of 0.4, 0.5, and 0.6. Subsequent to thorough experimentation, we determined that the best-performing model is the one with a higher number of filters and layers, specifically 3 layers and 24 filters, with a threshold of 0.4. This configuration resulted in an 80.89% detection performance and approximately 68.27% classification accuracy on the validation set. The results and supporting data are presented in Table 1.

After performing parameter tuning, we assessed the results by testing the accuracy on the test set. We observed that the model’s performance, which was favorable on the validation

Table 1. Result of performance on validation set

Parameters		Detection Performance					Classification Performance				
Number Layer	Number Filter	M	o2o	DR(%)	RA(%)	FM(%)	M	o2o	DR(%)	RA(%)	FM(%)
1	8	11970	10422	74.21	87.06	80.12	11970	8713	62.04	72.79	66.98
1	16	12067	10515	74.87	87.13	80.54	12067	8884	63.26	73.62	68.05
1	24	12273	10555	75.16	86.00	80.21	12273	8897	63.35	72.49	67.61
2	8	11971	10396	74.02	86.84	79.92	11971	8623	61.40	72.03	66.29
2	16	11855	10334	73.58	87.16	79.80	11855	8721	62.10	73.56	67.34
2	24	12124	10563	75.21	87.12	**80.73**	12124	8996	64.06	74.19	**68.75**
3	8	12168	10537	75.03	86.59	80.40	12168	8861	63.09	72.82	67.61
3	16	12101	10565	75.23	87.30	**80.82**	12101	8937	63.64	73.85	**68.36**
3	24	12250	10635	75.73	86.81	**80.89**	12250	8976	63.91	73.27	**68.27**

Table 2. Result of performance on test set

Parameters		Detection Performance					Classification Performance				
Number Layer	Number Filter	M	o2o	DR(%)	RA(%)	FM(%)	M	o2o	DR(%)	RA(%)	FM(%)
2	24	23977	20782	75.02	86.67	80.43	23977	17940	64.76	74.82	69.43
3	16	23936	20774	74.99	86.78	80.46	23936	17683	63.83	73.87	68.49
3	24	24130	20908	75.48	86.64	**80.68**	24130	17957	64.82	74.41	**69.29**

set, also translated to good results on the test set. The accuracy achieved was approximately 80.86% for detection performance and 69.29% for classification performance. These results indicate that the model demonstrates reliable detection and classification capabilities, as displayed in Table 2. As a result, we obtained the bounding boxes for every character on Palm leaf manuscript documents, as illustrated in Fig. 4.

4. CONCLUSIONS

This research work introduces a word spotting method specifically designed for Khmer palm leaf manuscript documents. Our approach involves utilizing a Region Proposal Network (RPN) built on the You Only Look Once (YOLO) technique and Convolutional Neural Network (CNN) to effectively and accurately identify targeted words or phrases within these ancient manuscripts. We conducted extensive experiments on a dataset comprising 1,971 images, with 70% allocated to the training set, 20% to the test set, and approximately 10% to the validation set. Through parameter tuning, we explored the impact of varying the number of layers (1, 2, 3) and the number of filters (8, 16, 24). The results demonstrate that the optimal model consists of 3 layers and 24 filters, with a threshold set at 0.4. The obtained accuracy for detecting the desired targets was around 80.86%, while the classification performance on the 33 Khmer characters reached 69.29%.

In future work, it would be valuable to expand the dataset and increase the number of classes to further enhance the performance of the word spotting method. Increasing the number of classes beyond the existing 33 categories would allow for the identification of a broader range of specific words and phrases within the documents.

ACKNOWLEDGMENTS

This research study is supported by Cambodia Higher Education Improvement Project (Credit No. 6221-KH).

REFERENCES

[1] Almazan, J., Gordo, A., Fornes, A., & Valveny, E. (2014, December 1). Word Spotting and Recognition with Embedded Attributes. IEEE Transactions on Pattern Analysis and Machine Intelligence, 36(12), 2552–2566. https://doi.org/10.1109/tpami.2014.2339814.

[2] Daraee, F., Mozaffari, S., & Razavi, S. M. (2021, June). Handwritten keyword spotting using deep neural networks and certainty prediction. Computers & Electrical Engineering, 92, 107111. https://doi.org/10.1016/j.compeleceng.2021.107111.

[3] Dey, S., Nicolaou, A., Llados, J., & Pal, U. (2016). Local Binary Pattern for Word Spotting in Handwritten

Historical Document. Lecture Notes in Computer Science, 574–583. https://doi.org/10.1007/978-3-319-49055-7_51.

[4] Frinken, V., Fischer, A., Manmatha, R., & Bunke, H. (2012, February). A Novel Word Spotting Method Based on Recurrent Neural Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 34(2), 211–224. https://doi.org/10.1109/tpami.2011.113.

[5] Manmatha, R. (2003). Word image matching using dynamic time warping. 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings. https://doi.org/10.1109/cvpr.2003.1211511.

[6] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016, June). You Only Look Once: Unified, Real-Time Object Detection. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/cvpr.2016.91.

[7] Sudholt, S., & Fink, G. A. (2018, February 14). Attribute CNNs for word spotting in handwritten documents. International Journal on Document Analysis and Recognition (IJDAR), 21(3), 199–218. https://doi.org/10.1007/s10032-018-0295-0.

[8] Ronen, R., TsipeFr, S., Anschel, O., Lavi, I., Markovitz, A., & Manmatha, R. (2022, August 5). GLASS: Global to Local Attention for Scene-Text Spotting. arXiv.org. https://arxiv.org/abs/2208.03364v1.

[9] Valy, D., Verleysen, M., Chhun, S., & Burie, J. C. (2017, November 10). A New Khmer Palm Leaf Manuscript Dataset for Document Analysis and Recognition. Proceedings of the 4th International Workshop on Historical Document Imaging and Processing. https://doi.org/10.1145/3151509.3151510.

[10] Valy, D., Verleysen, M., & Sok, K. (2016, October). Line Segmentation Approach for Ancient Palm Leaf Manuscripts Using Competitive Learning Algorithm. 2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR). https://doi.org/10.1109/icfhr.2016.0032.

[11] Valy, D., Verleysen, M., Chhun, S., & Burie, J. C. (2018, August). Character and Text Recognition of Khmer Historical Palm Leaf Manuscripts. 2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR). https://doi.org/10.1109/icfhr-2018.2018.00012.

[12] Zhao, P., Xue, W., Li, Q., & Cai, S. (2021). Query by Strings and Return Ranking Word Regions with Only One Look. Computer Vision – ACCV 2020, 3–18. https://doi.org/10.1007/978-3-030-69544-6_1.

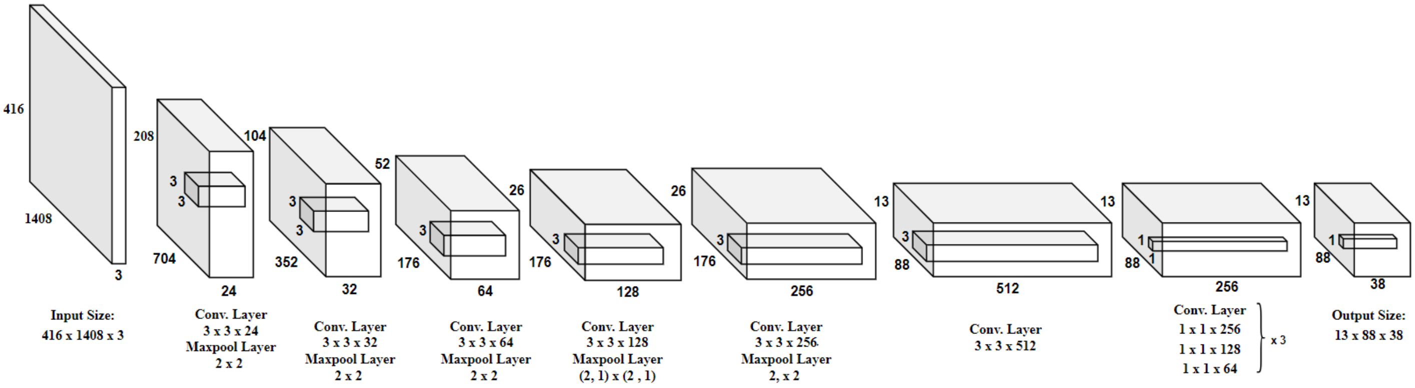

Fig. 3. Model architecture: Our model consists of 24 layers, including convolutional layers, activation functions, and max pooling. We have designed the model to accommodate an input size of 416 x 1408 with 3 image channels, resulting in an output size of 13 x 88 x 38 for efficient detection and classification of 33 classes.

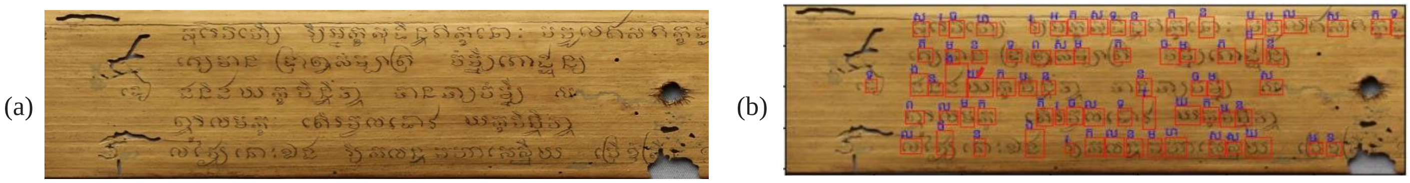

Fig. 4. (a) Input image, (b) Output image from approach