Parsing Digitized Vietnamese Paper Documents

Linh Truong Dieu[1,2], Thuan Trong Nguyen[1,2], Nguyen D. Vo[1,2],
Tam V. Nguyen[3], and Khang Nguyen[1,2]

[1] University of Information Technology, Ho Chi Minh, Vietnam
[2] Vietnam National University, Ho Chi Minh, Vietnam
[3] University of Dayton, Dayton, USA.
{17520691,18521471}@gm.uit.edu.vn,
{nguyenvd,khangnttm}@uit.edu.vn,
tamnguyen@udayton.edu

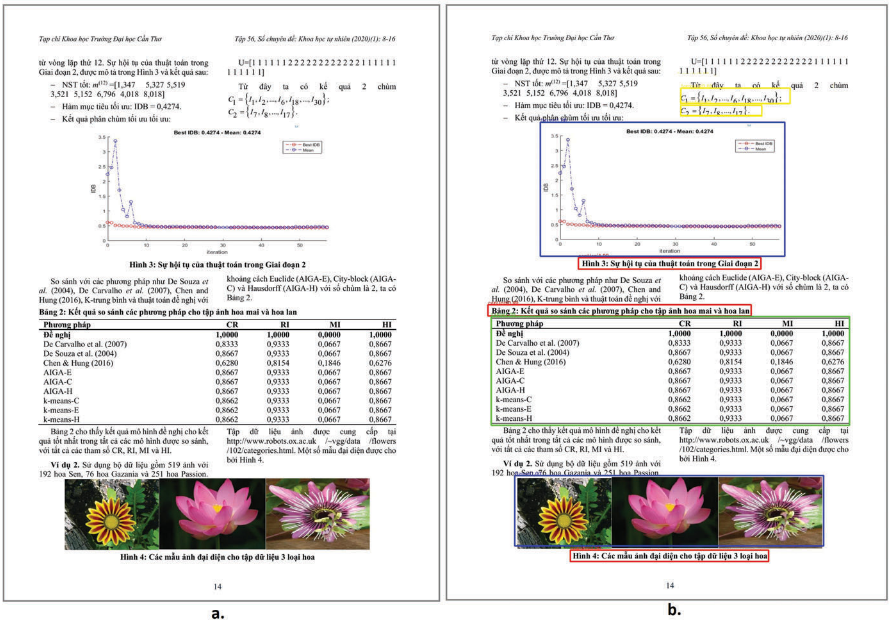
Abstract. In recent years, the need to exploit digitized document data has been increasing. In this paper, we address the problem of parsing digitized Vietnamese paper documents. The digitized Vietnamese documents are mainly in the form of scanned images with diverse layouts and special characters introducing many challenges. To this end, we first collect the UIT-DODV dataset, a novel Vietnamese document image dataset that includes scientific papers in Vietnamese derived from different scientific conferences. We compile both images that were converted from PDF and scanned by a smartphone in addition a physical scanner that poses many new challenges. Additionally, we further leverage the state-of-the-art object detector along with the fused loss function to efficiently parse the Vietnamese paper documents. Extensive experiments conducted on the UIT-DODV dataset provide a comprehensive evaluation and insightful analysis.

Keywords: Object detection · Page object detection · Deep learning · Convolutional neural network.

1 Introduction

The COVID19 pandemic has been changing our lives, which requires us to have a proactive approach toward accessing future technologies for manufacturing processes. With digital transformation, paper documents are also gradually converted and replaced by electronic documents for storage on the Cloud Storage, convenient for accessing and searching. The paper documents are stored in images or PDF files format depending on each organization, which leads to many challenges to extract necessary information. This requires a good enough detector model as the foundation for extracting information tasks. The problem's input is a document image with objects on a possible page: Caption, Table, Figure, and Formula. The output is an image containing the position of the objects expressed by bounding boxes and their labels (as shown in Fig. 1).

Almost of datasets for object detection on the document page are only converted from PDF, Latex, Word documents. In recent years, with smartphone development, documents are stored in image format using scan apps. Therefore,



Tạp chí Khoa học Trường Đại học Cần Thơ *Tập 56, Số chuyên đề: Khoa học tự nhiên (2020)(1): 8-16*

từ vòng lặp thứ 12. Sự hội tụ của thuật toán trong Giai đoạn 2, được mô tả trong Hình 3 và kết quả sau:

- NST tốt: $m^{(12)}$ =[1,347 5,327 5,519 3,521 5,152 6,796 4,018 8,018]
- Hàm mục tiêu tối ưu: IDB = 0,4274.
- Kết quả phân chùm tối ưu tối ưu:

U=[1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 1 1 1 1 1 1 1 1 1 1 1]

Từ đây ta có kế quả 2 chùm $C_1 = \{I_1, I_2, ..., I_6, I_{18}, ..., I_{30}\}$; $C_2 = \{I_7, I_8, ..., I_{17}\}$.

Hình 3: Sự hội tụ của thuật toán trong Giai đoạn 2

So sánh với các phương pháp như De Souza *et al.* (2004), De Carvalho *et al.* (2007), Chen and Hung (2016), K-trung bình và thuật toán đề nghị với khoảng cách Euclide (AIGA-E), City-block (AIGA-C) và Hausdorff (AIGA-H) với số chùm là 2, ta có Bảng 2.

Bảng 2: Kết quả so sánh các phương pháp cho tập ảnh hoa mai và hoa lan

Phương pháp	CR	RI	MI	HI
Đề nghị	**1,0000**	**1,0000**	**0,0000**	**1,0000**
De Carvalho et al. (2007)	0,8333	0,9333	0,0667	0,8667
De Souza et al. (2004)	0,8667	0,9333	0,0667	0,8667
Chen & Hung (2016)	0,6280	0,8154	0,1846	0,6276
AIGA-E	0,8667	0,9333	0,0667	0,8667
AIGA-C	0,8667	0,9333	0,0667	0,8667
AIGA-H	0,8667	0,9333	0,0667	0,8667
k-means-C	0,8662	0,9333	0,0667	0,8667
k-means-E	0,8662	0,9333	0,0667	0,8667
k-means-H	0,8662	0,9333	0,0667	0,8667

Bảng 2 cho thấy kết quả mô hình đề nghị cho kết quả tốt nhất trong tất cả các mô hình được so sánh, với tất cả các tham số CR, RI, MI và HI.

Ví dụ 2. Sử dụng bộ dữ liệu gồm 519 ảnh với 192 hoa Sen, 76 hoa Gazania và 251 hoa Passion. Tập dữ liệu ảnh được cung cấp tại http://www.robots.ox.ac.uk /~vgg/data /flowers /102/categories.html. Một số mẫu đại diện được cho bởi Hình 4.

Hình 4: Các mẫu ảnh đại diện cho tập dữ liệu 3 loại hoa

14

Tạp chí Khoa học Trường Đại học Cần Thơ *Tập 56, Số chuyên đề: Khoa học tự nhiên (2020)(1): 8-16*

từ vòng lặp thứ 12. Sự hội tụ của thuật toán trong Giai đoạn 2, được mô tả trong Hình 3 và kết quả sau:

- NST tốt: $m^{(12)}$ =[1,347 5,327 5,519 3,521 5,152 6,796 4,018 8,018]
- Hàm mục tiêu tối ưu: IDB = 0,4274.
- Kết quả phân chùm tối ưu tối ưu:

U=[1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 1 1 1 1 1 1 1 1 1 1 1]

Từ đây ta có kế quả 2 chùm $C_1 = \{I_1, I_2, ..., I_6, I_{18}, ..., I_{30}\}$; $C_2 = \{I_7, I_8, ..., I_{17}\}$.

Hình 3: Sự hội tụ của thuật toán trong Giai đoạn 2

So sánh với các phương pháp như De Souza *et al.* (2004), De Carvalho *et al.* (2007), Chen and Hung (2016), K-trung bình và thuật toán đề nghị với khoảng cách Euclide (AIGA-E), City-block (AIGA-C) và Hausdorff (AIGA-H) với số chùm là 2, ta có Bảng 2.

Bảng 2: Kết quả so sánh các phương pháp cho tập ảnh hoa mai và hoa lan

Phương pháp	CR	RI	MI	HI
Đề nghị	**1,0000**	**1,0000**	**0,0000**	**1,0000**
De Carvalho et al. (2007)	0,8333	0,9333	0,0667	0,8667
De Souza et al. (2004)	0,8667	0,9333	0,0667	0,8667
Chen & Hung (2016)	0,6280	0,8154	0,1846	0,6276
AIGA-E	0,8667	0,9333	0,0667	0,8667
AIGA-C	0,8667	0,9333	0,0667	0,8667
AIGA-H	0,8667	0,9333	0,0667	0,8667
k-means-C	0,8662	0,9333	0,0667	0,8667
k-means-E	0,8662	0,9333	0,0667	0,8667
k-means-H	0,8662	0,9333	0,0667	0,8667

Bảng 2 cho thấy kết quả mô hình đề nghị cho kết quả tốt nhất trong tất cả các mô hình được so sánh, với tất cả các tham số CR, RI, MI và HI.

Ví dụ 2. Sử dụng bộ dữ liệu gồm 519 ảnh với 192 hoa Sen, 76 hoa Gazania và 251 hoa Passion. Tập dữ liệu ảnh được cung cấp tại http://www.robots.ox.ac.uk /~vgg/data /flowers /102/categories.html. Một số mẫu đại diện được cho bởi Hình 4.

Hình 4: Các mẫu ảnh đại diện cho tập dữ liệu 3 loại hoa

14

Fig. 1: *The problem of detecting objects in Vietnamese document images. a. The input is the document image (left), b. The output is the formula position (yellow), figure (blue), caption (red), table (green).* [View better in colored version]

in this paper, we have introduced the UIT-DODV dataset, including images from PDF and scanned from both scanner and smartphone with more challenges and more practical, suitable for present and future trends. The first reason for choosing the Vietnamese scientific documents is the lack of current datasets. Secondly, the semantic complexity is also a more challenging part of the dataset, promising cause complex problem for the "state of the art" (SOTA) method. We hope that our dataset will play an important role in future works such as OCR (Optical Character Recognition), VQA (Visual Question Answering) on Vietnamese documents. The main contributions to this paper include:

- To the best of our knowledge, UIT-DODV is the first Vietnamese scientific documents dataset.
- We explored the performance of four SOTA models: CascadeTabNet, Faster-RCNN, YOLOv4, YOLOv4x-mish to evaluate challenges encountering from the dataset.
- We proposed a fused loss function for this task. We believe this work is a cornerstone for the development of future algorithms for the given problem.

[4] https://www.icst.pku.edu.cn/cpdp/sjzy/index.htm

Table 1: *The statistic of publicly available datasets.*

Dataset	Images	Categories	Coverage	Source	Year
TableBank [13]	417,234	Table	English, Chinese, Japanese, Arabic	Miscellaneous	2019
PublayNet [26]	358,353	Title, Text, List, Table, Figure	Medical	PubMed Central	2019
cTDaR2019 Modern Dataset [5]	840	Table	Printed documents	Miscellaneous	2019
Marmot[4]	2,000	Table	Chinese, English	Founder Apabi Library and Citeseer website	2020
UIT-DODV (Ours)	**2,394**	**Table, Figure, Formula, Caption**	**Vietnames research papers**	**National Conference, Can Tho University**	**2021**

2 Related Work

In this section, we briefly review the available datasets and the current approaches to tackle the problem.

2.1 Existing Datasets

TableBank [13] contains more than 278,000 images with more than 47,000 table objects. In which 200,000 images are made of Latex are scientific papers collected from ArXiv.org. **DocBank** [12] is an extension of TableBank. **PubLayNet** [26] is the largest documentary image dataset ever, including 358,353 images from research papers and scientific papers in the medical fields with five classes: Title, Text, Figure, Table, and List. **cTDaR 2019** [5] is a dataset used in the ICDAR2019 competition with two new versions, including modern printed documents and archival documents. **Marmot** consists of 2,000 pages in PDF format used for the table detection task. Besides, another dataset for formula detection is taken from 400 pages of research papers with 1,575 formulas and 7,907 embedded formulas from 194 PDF documents. The details of the datasets mentioned above are described in Table 1.

2.2 Parsing Digitized Paper Documents Problem

Traditional Approaches: Most traditional methods mainly consist of shape-based methods [3, 24, 7] and texture-based [21, 4]. In 2002, a new table detection method was proposed by Cesarini et al. [2], which detecting horizontal and vertical lines, then defining the area surrounded by these lines. To reduce candidate region detection errors, Gatos et al. [6] improved it by adding intersection detection in 2005. Overall, the traditional table detection method made significant progress, but many problems still exist, such as mistaken or omitted objects.

Deep Learning Approaches: In 2016, Hao et al. [8] proposed a table detection method based on CNN to determine if each proposed region contains a table or

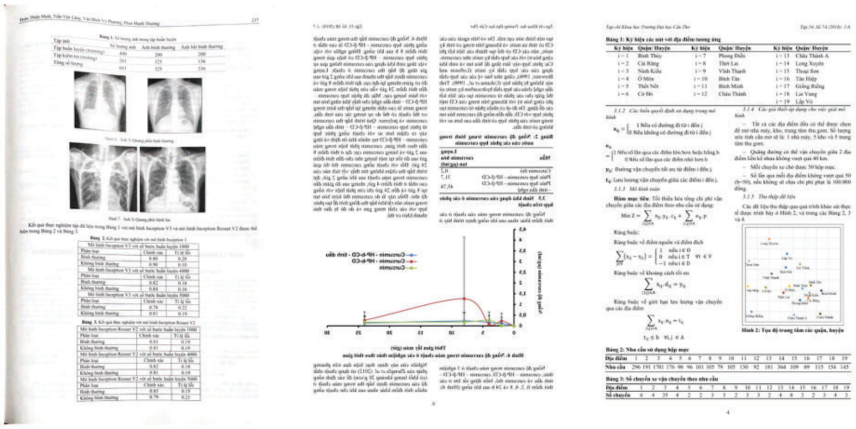


Fig. 2: *Exemplary samples in UIT-DODV dataset.*

not. In 2017, Yang et al. [25] proposed the FCN network for page segmentation to detect tables, figures, and other page objects. He et al. [9] proposed the multi-scale multi-task FCN to detect table areas and borders and used contour detection results to help improve detection tasks. Rashid et al. [16] proposed a table recognition method based on the AutoMLP algorithm. In 2018, Li et al. [14] used the layout analysis methods to identify some candidate tablespaces, then applied a conditional random field (CRF) and CNN to classify it as formula, tables, and images or graphs. In 2018, Vo et al. [23] combined the region proposals from Fast-RCNN and Faster-RCNN before applying bounding box regression to boost performance. In 2019, Huang et al. [10] proposed a method of object detection table based on YOLOv3. Sun et al. [22] proposed a method by combining Faster R-CNN [20] and corner locating. In 2020, Prasad et al. [15] proposed CascadeTabNet that uses Cascade Mask R-CNN HRNet to recognize table structures on image documents.

3 UIT-DODV Dataset

3.1 Dataset Collection

We collected the data from the scientific paper available on the website of Can Tho University (CTU). In addition, we used the physical scanner and the scanning app on smartphone to scan the hard-copy of National Conference "Some Selected Issues of Information Technology and Communication" from the following versions:

- The XXI edition with the topic "Internet of Things" was held from July 27-28, 2018, at Hong Duc University, Thanh Hoa Province.
- The XXII edition with the topic "Transforming the number of socio-economic operating in the Industry 4.0" was held from June 28 to 29, 2019, at Thai Binh University, Thai Binh Province.

The images in UIT-DODV are created by converting paper documents to digital images, using document conversion program, physical scanners and scanning app on smartphone. The purpose of collecting data from multiple sources is to create diversity for data in terms of layout, presentation form and data domain is also expanded with scanned images instead of just using the transferred image convert from PDF. After collecting the desired data, we proceeded to label attachments. Instead of tagging the data from where, we used the pretrain model from the PAA [11] method to predict the bounding box for the object before manually editing those objects. Fig. 2 visualizes some exemplary samples in our dataset.

3.2 Category Selection

We followed other image-based document datasets to select the categories. In particular, the first selected object is **Table -** appearing in most published datasets. **Figure** is the simplest and most effective way to turn complex ideas into a concise form, which can be a statistical graph that helps visualize the results of research. Shapes include natural sceneries, graphs, charts, layout designs, block diagrams, or maps. Likewise, **Formula** is equally important to describe relationships between concepts and objects concretely and efficiently. The formulas are usually numbered and may occupy several text lines. In addition, the formula object that contains equations and non-math text in a math region leads to the challenge of this object. Besides, with the desire to build a dataset that can be used for many different tasks such as OCR or VQA, we chose to add a new label, **Caption -** presenting a brief and yet complete explanation of the figure or table. The caption for a figure usually appears below the graphic; for a table, above.

3.3 Dataset Description

UIT-DODV is the first Vietnamese document image dataset, including 2,394 images with four classes: Table, Figure, Caption, Formula. UIT-DODV converted 1,696 images from PDF with size 1,654 x 2,338, 247 images scanned from the physical scanner and expanded with 451 images scanned from the smartphone.

UIT-DODV has the following highlights: **(1) Variety of images:** images in our dataset are of two types, with images converted from PDF as complete documents and images. Scan images often have lower resolutions depending on the scanning angle as well as the lighting conditions that can cause the document page to be blurred, distorted, skewed, or obscured. **(2) Variety of layout:** data collected from other scientific conferences/journals, a common feature of these conferences/journals is that they often use their templates (typically document pages can represent document pages in the form of one column or two columns). **(3) The challenge comes from data classes:** with the simultaneous use of two formula objects (Formula) and Caption creates a challenge for our dataset as well. As in building detection models for these objects. The vast majority of a document page is represented as text, so spotting these objects quickly is very difficult.

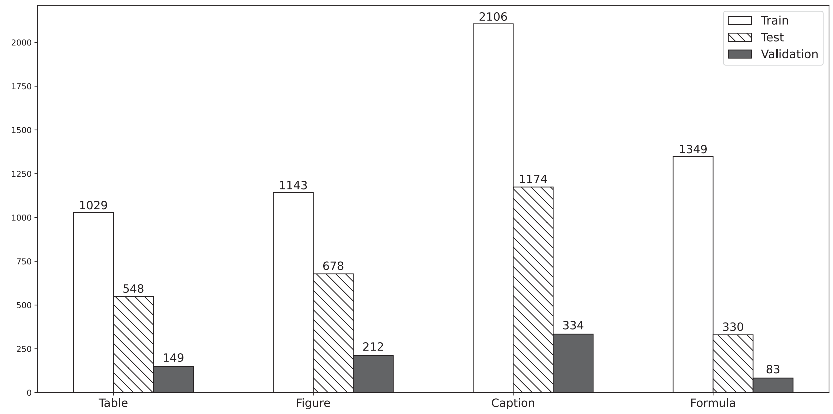


Fig. 3: Statistical of experimental data.

4 Computational Model

4.1 Object Detector

In this work, we leveraged the SOTA object detector for the problem. In particular, we took four object detection methods into consideration. The details are listed as follows.

CascadeTabNet [15] is a new approach that uses a single CNN to recognize table structures on document image. CascadeTabNet is a three-phase Cascade mask R-CNN HRNet model, with each input image, the CNN HRNetV2p W32 backbone is responsible for converting this image into a feature map. “RPN Head” (Dense Head) predicts preliminary object proposals for this feature map. “Bbox Heads” takes RoI features as input and delivers RoI-wise predictions. The output for each section includes two predictions, classification score and box regression points. “Mask Head” predicts masks for objects. CascadeTabNet uses Cascade R-CNN’s late-stage segmentation branching strategy, object detection performed by “Bbox Heads” is complemented with segmentation masks implemented by “Mask Head”, for all. detected objects.

Faster-RCNN [20] is a revamp of its precursors R-CNN and Fast R-CNN and achieves near real-time speeds when skipping the time spent on regional suggestions when using the search algorithm. Selective Search, instead using pre-trained models to create feature maps, the feature map is then fed through the CNN Region proposal network (RPN) to find the proposed regions, from there generating anchor boxes, anchor boxes continue to be layered. Finally, Non maximum suppression (NMS) algorithm filters out overlapping anchors.

YOLOv4 is proposed by Bochkovskiy, Wang, and Liao which considered as the best version in terms of accuracy and calculation speed compared to the previous 3 versions [19, 17, 18] thanks to the combination of CSPNet architecture with Darknet-53 (YOLOv3) for the backbone and adding 2 SPP and PANet modules. At the same time, the BoF and BoS techniques are utilized to help YOLOv4 achieve an AP score of 43.5% on the COCO dataset with an execution speed of 65 FPS.

YOLOv4x-mish is an additional version with some changes from YOLOv4. Basically, the YOLOv4x-mish change first CSPDarknet stage of backbone to original Darknet-53 balances speed and accuracy. In the neck, instead of keeping the PAN architecture the same as YOLOv4, they made CSP-ize PAN block to decrease about 40% in computational volume. In this CSPPAN block, the SPP model stays the same as YOLOv4 to increase the receptive field. In Head section, YOLOv4x-mish change activation function Leaky ReLU to Mish - which is a non-monotonic activation function, which helps increase accuracy for predicted model.

4.2 Loss Function

In object detection, the loss function plays a vital role to the detection performance. In literature, the cross-entropy (CE) loss function is widely used. The idea behind CE loss is to penalize the wrong predictions more than to reward the right predictions. The cross-entropy loss function is defined as follows.

$$\mathcal{L}_{CE}(p_t) = -log(p_t) \tag{1}$$

where p_t is the probability for the class t. Recently, the focal loss (FL) was proposed as an version of cross-entropy loss that handles the class imbalance problem by assigning more weights to hard or easily misclassified examples. The focal loss function is defined as:

$$\mathcal{L}_{FL}(p_t) = -\alpha(1 - p_t)^{\gamma} log(p_t), \tag{2}$$

where α is a balanced form for Focal Loss, defaults to 0.25; the gamma (γ) for calculating the modulating factor, defaults to 2.0.

In this work, we argue that each loss function has its own advantages and drawbacks. In the given problem of parsing digitized Vietnamese paper documents, we propose combining different loss functions in order to further improve the performance. Here, the fused loss function is defined as below.

$$\mathcal{L}_{fused}(p_t) = \lambda\mathcal{L}_{CE}(p_t) + (1 - \lambda)\mathcal{L}_{FL}(p_t), \tag{3}$$

where the effect of each individual loss function is decided by the weight λ. In our implementation, we set λ as 0.6 to emphasize the cross entropy loss.

5 Experimental Results and Discussion

5.1 Experimental Setting

Our dataset is divided into 3 subsets: training (1,440), validation (234) and testing (720) sets as shown in Fig. 3. The entire experiment was conducted on a GeForce RTX 2080 Ti GPU with 11019MiB memory. We trained Faster R-CNN on the MMDetection framework *V2.10.0* using the default configuration

Table 2: *Experimental results of different object detection methods with default configuration. The best performance is marked in boldface.*

Architecture	**Table**	**Figure**	**Formula**	**Caption**	**AP$_{50}$**	**AP$_{75}$**	**AP**
Faster-RCNN	91.60	79.70	45.60	57.70	86.20	76.20	68.70
CascadeTabNet	**95.70**	**83.40**	**48.10**	**67.40**	**89.00**	**80.20**	**73.60**
YOLOv4	84.20	78.00	40.20	60.80	90.20	75.20	65.80
YOLOv4x_mish	82.00	75.70	45.20	61.30	90.70	77.70	66.10

Table 3: *Experimental results of different loss functions. The best performance is marked in boldface.*

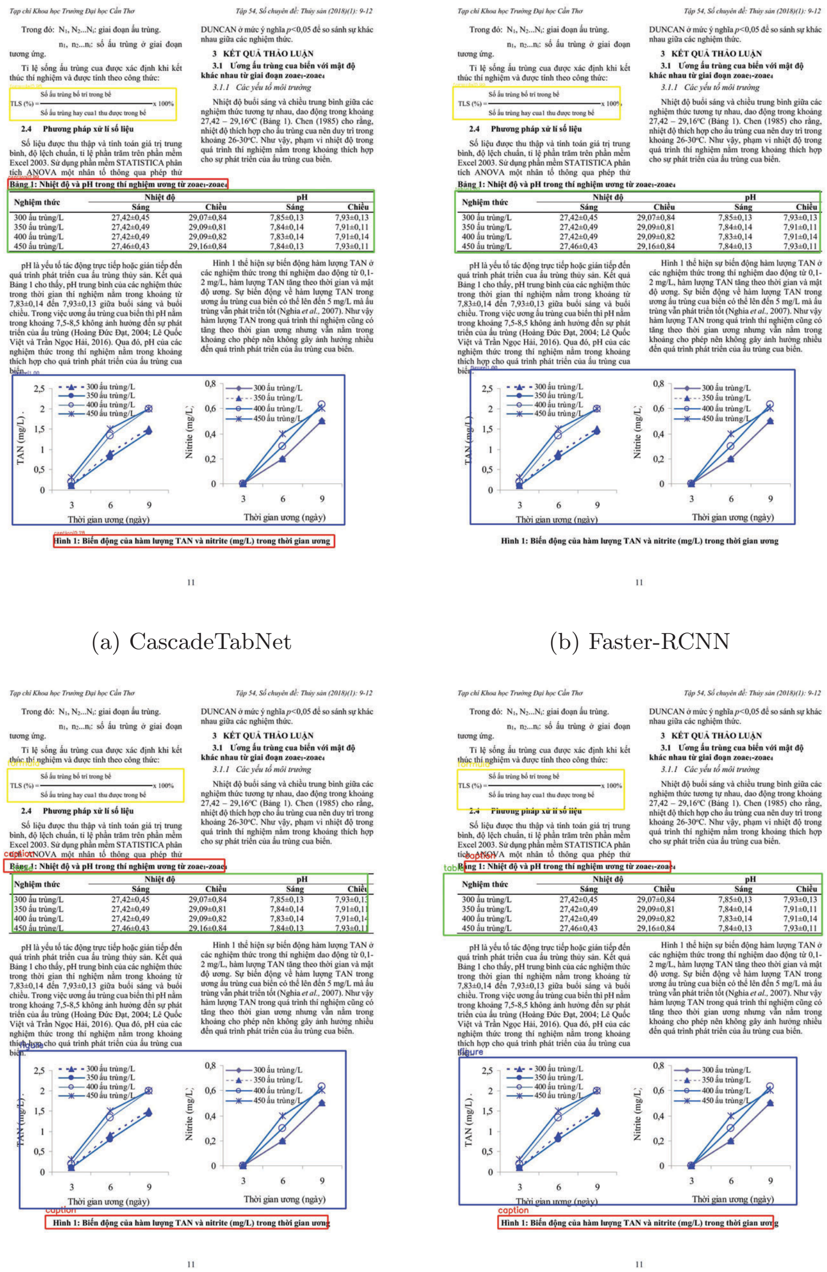
Method	**Cls Loss**	**Table**	**Figure**	**Formula**	**Caption**	**AP$_{50}$**	**AP$_{75}$**	**AP**
Faster R-CNN	Cross-entropy	91.60	79.70	45.60	57.70	86.20	76.20	68.70
Faster R-CNN	Focal	91.60	79.30	37.60	64.30	87.90	74.50	68.20
Faster R-CNN	Fused loss function	92.40	79.70	45.50	66.40	89.00	77.90	71.00
CascadeTabNet	Cross-entropy	**95.70**	**83.40**	**48.10**	67.40	89.00	80.20	73.60
CascadeTabNet	Focal	95.50	82.40	44.20	65.20	87.80	77.60	71.80
CascadeTabNet	Fused loss function	94.30	83.00	47.50	**73.30**	**89.10**	**81.60**	**74.50**

with the backbone *X-101-64x4d-FPN* trained within 24 epochs. For YOLOv4 and YOLOv4x-mish, we implemented it on the Darknet framework. To train the CascadeTabNet model, we ran it with the default configuration provided by Prasad et al. [5]. We used mAP to evaluate the effectiveness of the model as in object detection competition on MS COCO.

5.2 Analysis Results

We used the best weights on the validation set to evaluate and report the results in Table 2. We found that the two-stage methods gave high accuracy for object detection. CascadeTabNet gave the best results when achieved AP@0.75 and AP@[0.5: 0.5: 0.95] is 80.05%, 73.40% respectively. Faster-RCNN gives quite good results in object detection. However, it still missed Caption - the object that accounts for the largest distribution of the dataset with AP is 57.70%, that lower than the other three methods. The two one-stage methods gave the best results on AP@0.5. However, when increasing the IoU threshold to 0.75, the AP score of YOLOv4 (75.17%), YOLOv4x-mish (77.72%) was lower than that of the two-stage methods. After visualizing the result shown in Fig. 4, we notice that YOLOv4, YOLOv4x-mish have difficulty creating a perfect bounding box compared to the two two-stage methods. The Table and Figure objects also create a challenge to distinguish, and there are also many cases of overlapping bounding boxes. We further conducted the experiment on different loss functions for the top-2 methods, i.e., Faster-RCNN and CascadeTabNet. As shown in

[5] https://github.com/DevashishPrasad/CascadeTabNet



Tạp chí Khoa học Trường Đại học Cần Thơ *Tập 54, Số chuyên đề: Thủy sản (2018)(1): 9-12*

Trong đó: $N_1, N_2 \ldots N_i$: giai đoạn ấu trùng.

$n_1, n_2 \ldots n_i$: số ấu trùng ở giai đoạn tương ứng.

Tỉ lệ sống ấu trùng cua được xác định khi kết thúc thí nghiệm và được tính theo công thức:

$$\text{TLS (\%)} = \frac{\text{Số ấu trùng bố trí trong bể}}{\text{Số ấu trùng hay cua1 thu được trong bể}} \text{x } 100\%$$

2.4 Phương pháp xử lí số liệu

Số liệu được thu thập và tính toán giá trị trung bình, độ lệch chuẩn, tỉ lệ phần trăm trên phần mềm Excel 2003. Sử dụng phần mềm STATISTICA phân tích ANOVA một nhân tố thông qua phép thử DUNCAN ở mức ý nghĩa $p<0,05$ để so sánh sự khác nhau giữa các nghiệm thức.

3 KẾT QUẢ THẢO LUẬN

3.1 Ương ấu trùng cua biển với mật độ khác nhau từ giai đoạn zoae$_1$-zoae$_4$

3.1.1 Các yếu tố môi trường

Nhiệt độ buổi sáng và chiều trung bình giữa các nghiệm thức tương tự nhau, dao động trong khoảng 27,42 – 29,16°C (Bảng 1). Chen (1985) cho rằng, nhiệt độ thích hợp cho ấu trùng cua nên duy trì trong khoảng 26-30°C. Như vậy, phạm vi nhiệt độ trong quá trình thí nghiệm nằm trong khoảng thích hợp cho sự phát triển của ấu trùng cua biển.

Bảng 1: Nhiệt độ và pH trong thí nghiệm ương từ zoae$_1$-zoae$_4$

Nghiệm thức	Nhiệt độ		pH	
	Sáng	Chiều	Sáng	Chiều
300 ấu trùng/L	27,42±0,45	29,07±0,84	7,85±0,13	7,93±0,13
350 ấu trùng/L	27,42±0,49	29,09±0,81	7,84±0,14	7,91±0,11
400 ấu trùng/L	27,42±0,49	29,09±0,82	7,83±0,14	7,91±0,14
450 ấu trùng/L	27,46±0,43	29,16±0,84	7,84±0,13	7,93±0,11

pH là yếu tố tác động trực tiếp hoặc gián tiếp đến quá trình phát triển cua ấu trùng thủy sản. Kết quả Bảng 1 cho thấy, pH trung bình của các nghiệm thức trong thời gian thí nghiệm nằm trong khoảng từ 7,83±0,14 đến 7,93±0,13 giữa buổi sáng và buổi chiều. Trong việc ương ấu trùng cua biển thì pH nằm trong khoảng 7,5-8,5 không ảnh hưởng đến sự phát triển của ấu trùng (Hoàng Đức Đạt, 2004; Lê Quốc Việt và Trần Ngọc Hải, 2016). Qua đó, pH của các nghiệm thức trong thí nghiệm nằm trong khoảng thích hợp cho quá trình phát triển của ấu trùng cua biển.

Hình 1 thể hiện sự biến động hàm lượng TAN ở các nghiệm thức trong thí nghiệm dao động từ 0,1-2 mg/L, hàm lượng TAN tăng theo thời gian và mật độ ương. Sự biến động về hàm lượng TAN trong ương ấu trùng cua biển có thể lên đến 5 mg/L mà ấu trùng vẫn phát triển tốt (Nghia *et al.*, 2007). Như vậy hàm lượng TAN trong quá trình thí nghiệm cũng có tăng theo thời gian ương nhưng vẫn nằm trong khoảng cho phép nên không gây ảnh hưởng nhiều đến quá trình phát triển của ấu trùng cua biển.

Hình 1: Biến động của hàm lượng TAN và nitrite (mg/L) trong thời gian ương

11

(a) CascadeTabNet

(b) Faster-RCNN

(c) YOLOv4

(d) YOLOv4x-mish

Fig. 4: Visualize results for the four object detection methods with 4 classes: formula (yellow), caption (red), table (green), figure (blue) [View better in colored version]

Table 3, the fused loss function yields the best performance in terms of AP. The fused loss function also achieves the best performance in the "caption" semantic class. This clearly demonstrates the need of using the fused loss function in the problem of parsing digitized Vietnamese paper documents.

6 Conclusion and Future Work

In this paper, we have introduced the first Vietnamese scientific document image dataset - UIT-DODV - with 4 main objects that are the elements of a research paper, namely, Table, Figure, Caption and Formula with a total of $2,394$ images. We conducted experiments on SOTA object detection methods: Faster-RCNN, YOLOv4, YOLOv4x-mish and a method was applied on the table detection problem - CascadeTabNet on our dataset. In which, the CascadeTabNet method achieved the highest mAP result of 74.50%. In the future, we will build a mobile application that identifies elements in an image document page. Additionally, we continue to expand and develop the UIT-DODV dataset to a larger number along with the diversity of the structure of the documents. Besides, many other problems will be applied to this dataset such as OCR or VQA.

Acknowledgment

This project is partially funded under National Science Foundation (NSF) under Grant No. 2025234 and Vietnam National University Ho Chi Minh City (VNU-HCM) under grant number DSC2021-26-03.

References

[1] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. *YOLOv4: Optimal Speed and Accuracy of Object Detection.* 2020. arXiv: 2004.10934 [cs.CV].

[2] F. Cesarini et al. "Trainable table location in document images". In: *Object recognition supported by user interaction for service robots.* Vol. 3. 2002, 236–240 vol.3. DOI: 10.1109/ICPR.2002.1047838.

[3] Fu Chang, Shih-Yu Chu, and Chi-Yen Chen. "Chinese Document Layout Analysis Using An Adaptive Regrouping Strategy". In: *Pattern Recognition* 38 (Feb. 2005), pp. 261–271. DOI: 10.1016/j.patcog.2004.05.010.

[4] K. Etemad, D. Doermann, and R. Chellappa. "Multiscale segmentation of unstructured document pages using soft decision integration". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 19.1 (1997), pp. 92–96. DOI: 10.1109/34.566817.

[5] L. Gao et al. "ICDAR 2019 Competition on Table Detection and Recognition (cTDaR)". In: *2019 International Conference on Document Analysis and Recognition (ICDAR).* 2019, pp. 1510–1515. DOI: 10.1109/ICDAR.2019.00243.

[6] Basilios Gatos et al. "Automatic Table Detection in Document Images". In: vol. 3686. Aug. 2005, pp. 609–618. ISBN: 978-3-540-28757-5. DOI: 10.1007/11551188_67.

[7] Jaekyu Ha, Ihsin Phillips, and Robert Haralick. "Document page decomposition using bounding boxes of connected components of black pixels". In: *Proceedings of SPIE - The International Society for Optical Engineering* (Mar. 1995). DOI: 10.1117/12.205816.

[8] Leipeng Hao et al. "A Table Detection Method for PDF Documents Based on Convolutional Neural Networks". In: Apr. 2016, pp. 287–292. DOI: 10.1109/DAS.2016.23.

[9] D. He et al. "Multi-Scale Multi-Task FCN for Semantic Page Segmentation and Table Detection". In: *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*. Vol. 01. 2017, pp. 254–261. DOI: 10.1109/ICDAR.2017.50.

[10] Yilun Huang et al. "A YOLO-based table detection method". In: *2019 International Conference on Document Analysis and Recognition (ICDAR)*. IEEE. 2019, pp. 813–818.

[11] Kang Kim and Hee Seok Lee. *Probabilistic Anchor Assignment with IoU Prediction for Object Detection*. 2020. arXiv: 2007.08103 [cs.CV].

[12] Minghao Li et al. *DocBank: A Benchmark Dataset for Document Layout Analysis*. 2020. arXiv: 2006.01038 [cs.CL].

[13] Minghao Li et al. "Tablebank: Table benchmark for image-based table detection and recognition". In: *Proceedings of The 12th Language Resources and Evaluation Conference*. 2020, pp. 1918–1925.

[14] X. Li et al. "Page Object Detection from PDF Document Images by Deep Structured Prediction and Supervised Clustering". In: *2018 24th International Conference on Pattern Recognition (ICPR)*. 2018, pp. 3627–3632. DOI: 10.1109/ICPR.2018.8546073.

[15] Devashish Prasad et al. *CascadeTabNet: An approach for end to end table detection and structure recognition from image-based documents*. 2020. arXiv: 2004.12629 [cs.CV].

[16] S. F. Rashid et al. "Table Recognition in Heterogeneous Documents Using Machine Learning". In: *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*. Vol. 01. 2017, pp. 777–782. DOI: 10.1109/ICDAR.2017.132.

[17] Joseph Redmon and Ali Farhadi. *YOLO9000: Better, Faster, Stronger*. 2016. arXiv: 1612.08242 [cs.CV].

[18] Joseph Redmon and Ali Farhadi. *YOLOv3: An Incremental Improvement*. 2018. arXiv: 1804.02767 [cs.CV].

[19] Joseph Redmon et al. *You Only Look Once: Unified, Real-Time Object Detection*. 2016. arXiv: 1506.02640 [cs.CV].

[20] Shaoqing Ren et al. *Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks*. 2016. arXiv: 1506.01497 [cs.CV].

[21] Jaakko Sauvola and Matti Pietikäinen. "Page segmentation and classification using fast feature extraction and connectivity analysis". In: Sept. 1995, 1127–1131 vol.2. ISBN: 0-8186-7128-9. DOI: 10.1109/ICDAR.1995.602118.

[22] Sun et al. “Table Detection Using Boundary Refining via Corner Locating”. In: *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*. Springer. 2019, pp. 135–146.

[23] Nguyen D Vo et al. “Ensemble of deep object detectors for page object detection”. In: *Proceedings of the 12th International Conference on Ubiquitous Information Management and Communication*. 2018, pp. 1–6.

[24] Jie Xi, Jianming Hu, and Lide Wu. “Page segmentation of Chinese newspapers”. In: *Pattern Recognition* 35.12 (2002). Pattern Recognition in Information Systems, pp. 2695–2704. ISSN: 0031-3203. DOI: `https://doi.org/10.1016/S0031-3203(01)00248-5`. URL: `https://www.sciencedirect.com/science/article/pii/S0031320301002485`.

[25] X. Yang et al. “Learning to Extract Semantic Structure from Documents Using Multimodal Fully Convolutional Neural Networks”. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 4342–4351. DOI: `10.1109/CVPR.2017.462`.

[26] X. Zhong, J. Tang, and A. Jimeno Yepes. “PubLayNet: Largest Dataset Ever for Document Layout Analysis”. In: *2019 International Conference on Document Analysis and Recognition (ICDAR)*. 2019, pp. 1015–1022. DOI: `10.1109/ICDAR.2019.00166`.