Automatic Bangla Signboard and Region of Text Interests Detection from Natural Scene Image

No Author Given

No Institute Given

Abstract. Accurate signboard localization in natural scene photos is a critical task for retrieving information regarding a shop or organization. This information can be used for several purposes, including translators, assisting foreigners and visually impaired people, company listings etc. However, detecting a signboard and the text regions of the signboard from a natural scenario is very difficult because of variable dynamic factors. In this research, we propose a Bangla Signboard and Region of Text Interests (RTI) detection system. The suggested system is divided into two stages, with YOLOv4 being employed for detection in both stages. The first and second stage's model achieved an accuracy of 98.6% and 97.4% , respectively. The suggested model is capable of reliably detecting signboards and RTI, even in adverse environmental conditions. In addition, to our knowledge we have created the largest Bangla signboard image dataset, consisting of 17,308 images from all over Bangladesh.

Keywords: Yolov4 · Computer vision · Signboard detection · RTI detection.

1 Introduction

Detecting signboards and RTI is a challenging task in the field of computer vision. In Bangladesh, the majority of signboards are in Bangla. For roadside businesses, signboards are critical sources of information. Signboard and RTI detection is the first step towards developing text-to-speech translation tools to assist non-Bangla speakers in comprehending the contents of Bangla signboards, as well as tools to assist the blind and other visually impaired individuals. Information gained from signboard and RTI detection can assist location-based online applications such as Google Maps, Uber, and foodpanda, as well as some company listing businesses, such as Bdbusinessfinder.com, Address Bazar, and Yelp. Signboard detection from an image remains a daunting process due to degraded signboards and difficult environmental scenes, such as the presence of pillars, trees, and wires that cross the signboards, etc. Another challenge is detecting multiple signboards in a single image. Additionally, robust and accurate RTI detection from signboards in natural scenes remains a challenge due to text font, size, colour, alignment, and orientation variations. Furthermore, it is frequently influenced by complex backgrounds, lighting condition, image distortion, and the degradation of signboards. Before doing any optical character recognition, it is very important to be very precise about detecting the RTI in a signboard image. In recent years, signboard and RTI detection have gotten a lot of attention. Numerous studies have employed a variety of techniques, but the following are notable: manual image binarization, deep neural network image processing [7], gradient-based Hough transform [15], neural networks [8], Convolutional Neural Network (CNN) [19], and Faster RCNN [4], [18].

In this paper, we propose a robust Bangla Signboard and RTI Detection System. Our proposed system is divided into two subsystems. The first is the detection of signboards from natural scene images, and in the second subsystem RTI which are shop name, and shop information consisting of shop address and contact number, are detected from the output of the first subsystem which is the detected signboard. The proposed system is capable of detecting signboards in natural scene images and localizing their RTI. Additionally, the proposed system is intelligent enough to distinguish legitimate from illegitimate Bangla signboards, based on the motivation of this research. In both subsystems, a real-time detection model called ”You Only Look Once” (Yolov4 [3]) is used for detection. Prior work with Yolov4 has been conducted for the ALPDR system [12], activity detection system [10], and a variety of other applications. However, research on signboard detection using YOLOv4 is limited. Additionally, we have created the largest diverse dataset for signboard and RTI detection, consisting of a total of 17,308 images captured across all of Bangladesh.

Two significant contributions of this research are as follows:

- A very diversified dataset of 17,308 photos, including exclusively Bangla signboard images, has been created in the context of this research.
- A real-time Automated Bangla Signboard Detection system that can detect signboards from the natural scenes and localize RTI e.g., shop name and shop information from the detected signboard image.

The remainder of the paper is structured as follows: Section 2 discusses the literature review. Section 3 provides an overview of our vast and diversified dataset. Section 4 contains a detailed description of the proposed system. The findings and comparisons from this research can be found in section 5. Section 6 contains the conclusion. The remainder is a list of references.

2 Literature Review

Numerous attempts have been made to improve CNNs by using various methods, that results in improvement of detection and recognition systems [21], [20]. In this section we discuss some mentionable deep learning and machine learning approaches towards signboard detection.

2.1 Machine Learning Approach

Shen and X. Tang [15] devised a method for identifying generic signboards. To begin, they preprocessed their images with noise reduction and a Canny edge detection algorithm. They then detected the boundary lines using the Gradient Hough transformation. Finally, they were able to detect the signboard by deleting superfluous rectangular boxes. They had a success rate of more than 93% based on 104 images. Their system, however, became perplexed in noisy, overlapping signboard environments.

Tam, Angela [16] proposed an automated technique for detecting and extracting text from signboards based on the detection of quadrilateral signboards in images. To locate the signboard on the image, the Hough transform, edge density checking, and quadrilateral finding was used. The system then converted the signboard from an arbitrary quadrilateral to a rectangular shape suitable for text extraction using a geometric transformation. They achieved an accuracy of 85.04% for detection and 80% for segmentation.

2.2 Deep Learning Approach

Numerous studies involving the detection of signboards using deep learning have been conducted in recent years. Pin-Xu Chen [4] demonstrated a Faster R-CNN model for detecting street-view signboards. The Model's mean average precision (mAP) was approximately 94.87%. Even though the dataset contained only 574 images. For mobile applications, J.Park [9] proposed a technique for detecting and recognizing Korean text or store names in images of outdoor signboards. The candidate text region was detected using an edge-based technique.

Syed Yasser Arafat [1] proposed a method for detecting Urdu text on store signboards and recognizing Urdu ligatures in the image of the detected signboard. They detected the signboard using Fast-RCNN. Following that, the detected image is subjected to some preprocessing techniques in order to segment the words. The segmented words are then recognizing using a CNN model. Numerous word and character segmentation techniques are used for a variety of purposes [14, 2, 13, 11], but the author of this paper proposed a novel technique for segmenting the words of signboards. Overall, they achieved an accuracy rate of 89%. Susanta Malakar [6] proposed a system for detecting signboards in which they used a bounding box algorithm to locate the signboard and then developed a CNN model to recognize its meaning. Their dataset contains a total of 2400 images classified into six categories. They achieved a 98% overall accuracy.

Prior work's experimental results had a number of shortcomings. Due to blurry boundaries, the proposed system in [15] fails at the edge detection phase. In [7], it may occasionally be unable to correctly recognize the combined characters. Additionally, texts can be mistaken for objects or vice versa. Inadequate dataset is yet another issue.

3 Dataset

In Bangladesh, most signboards are in Bangla. A signboard has two aspects. It begins with the shop or organization's name and ends with its RTI which contains address and contact information. There is a lack of sufficient datasets for the detection of Bangla signboards and RTI. The available dataset also lacks natural characteristics such as fuzzy photos, multiple angles, a noisy background, low resolution, a deteriorated signboard, and multiple orientations. So, in this work, we present the largest Bangla signboard dataset to date, with 17,308 photos from all over Bangladesh. The suggested dataset includes images of insufficiently illuminated signboards, nighttime images, dashcam images, hazy images, degraded signboards, and even signboards with wires crossing them. This dataset's images are of varying quality, including zoomed-in and low-resolution images. Signboards were manually annotated.

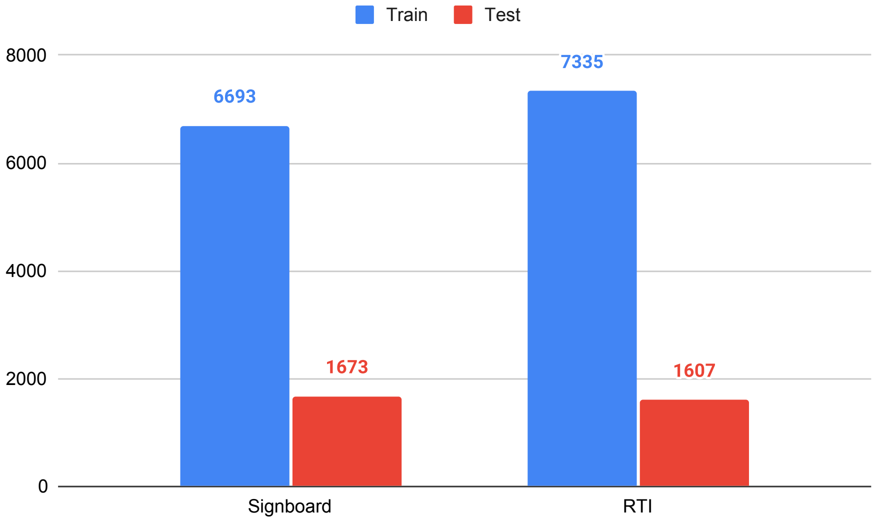

Fig. 1. Dataset distribution of both phases

This dataset is divided into two groups. Signboard images from the first group are used to train the signboard detection model, while images from the second group train the RTI detection model. There are 6693 training and 1673 testing images in the first group, while 7335 images for training and 1607 for testing in the second group. The data distribution of these two groups is illustrated in Fig 1. Fig 2 and Fig 3 show some sample images of the first and second groups of the dataset respectively.

4 Methodology

This study proposes an automatic real-time detection system for Bangla signboards and RTI. The suggested system utilizes the You Only Look Once (Yolov4) [3] object detection model, which is a real-time state-of-the-art object detection

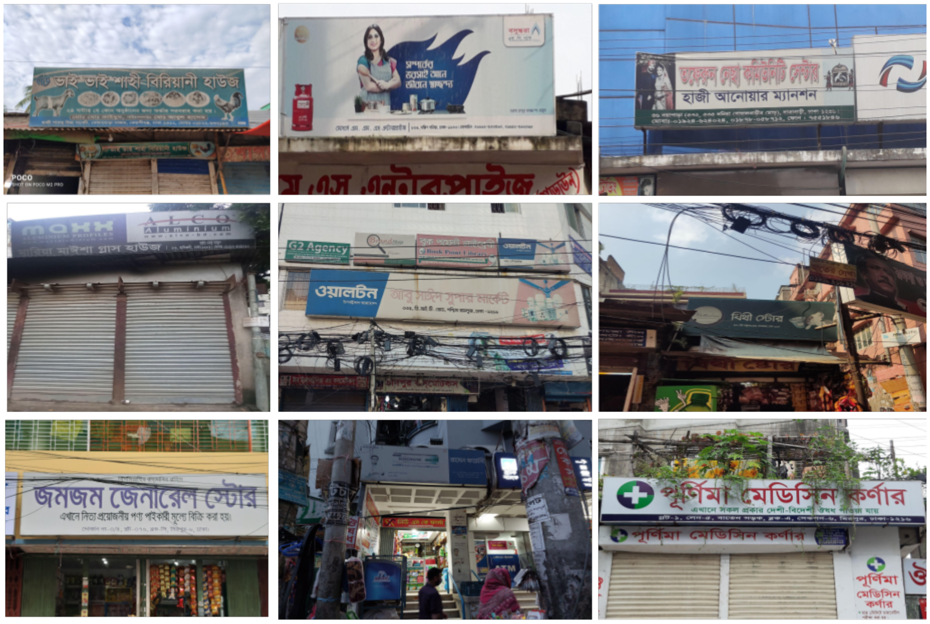

Fig. 2. Dataset sample of first phase.

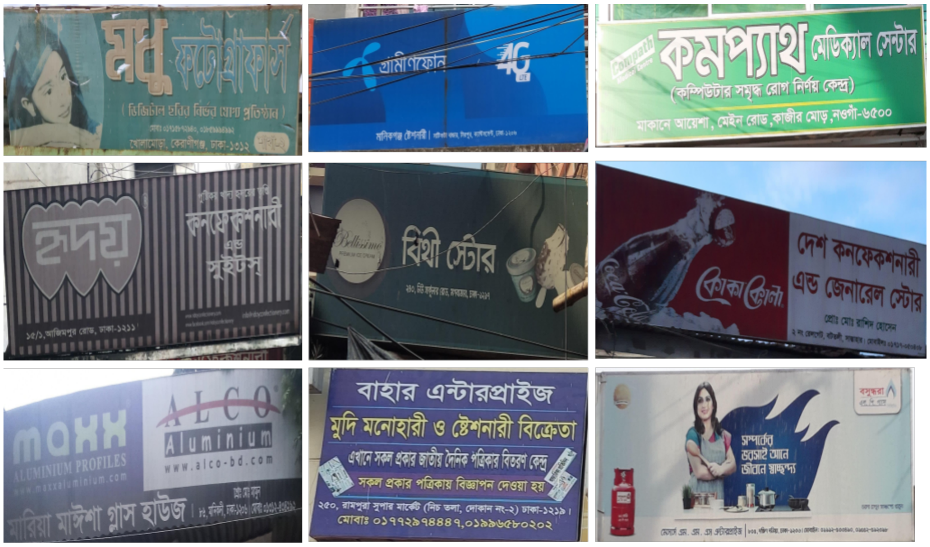
মধু ফটোগ্রাফার্স
গ্রামীণফোন
4G
কমপ্যাথ মেডিক্যাল সেন্টার
(কম্পিউটার সমৃদ্ধ রোগ নির্ণয় কেন্দ্র)
হৃদয় কনফেকশনারী এন্ড সুইটস
বিথী স্টোর
কোকা-কোলা
দেশ কনফেকশনারী এন্ড জেনারেল স্টোর
maxx
ALUMINIUM PROFILES
ALCO Aluminium
www.alco-bd.com
বাহার এন্টারপ্রাইজ
মুদি মনোহারী ও ষ্টেশনারী বিক্রেতা
এখানে সকল প্রকার জাতীয় দৈনিক পত্রিকার বিতরণ কেন্দ্র
সকল প্রকার পত্রিকায় বিজ্ঞাপন দেওয়া হয়

Fig. 3. Dataset sample of second phase.

model. YOLOv4 is a sophisticated neural network that employs a revolutionary technique for real-time item recognition that requires only one look at the train image. Our system is divided into two distinct phases. To begin, the first phase detection model is fed an entire image or video frame. Following that, the output of the first model (detected signboard picture) is transferred to the second phase detection model, which detects RTI such as store name, address, and phone number. Fig 4 illustrates the whole system's flow diagram. Utilized YOLOv4 architecture and proposed hyperparameters are provided in the following sub-sections.

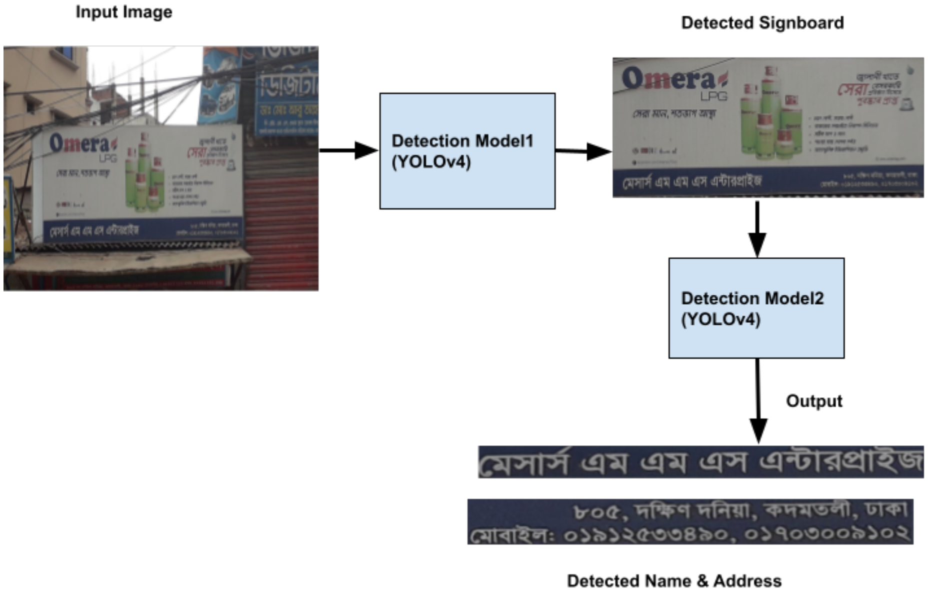

Fig. 4. Overview of proposed system.

4.1 Cross-Stage-Partial-connections (CSP) Block

Each convolution layer in a Dense Block comprises batch normalization, ReLU, and convolution. CSPNet distributes the input feature maps of the DenseBlock into two portions. The first portion omits the DenseBlock and serves as an input to the subsequent transition layer. The second segment will traverse the Dense block in the manner seen in Fig 5. By separating the input into two halves and passing only one via the Dense Block, this unique design reduces computations.

4.2 Spatial Pyramid Pooling (SPP) Block

SPP employs a somewhat different technique when detecting objects of diverse sizes. It replaces the previous pooling layer with a spatial pyramid pooling layer.

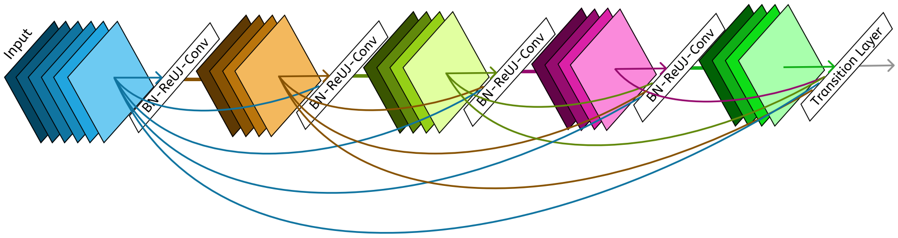

Fig. 5. CSP DensNet block.

Spatially, the feature maps are divided into m * m bins. YOLO changes the SPP in order to preserve the spatial dimension of the output. A sliding kernel with a maximum pool is used. The output is then concatenated with the feature maps obtained from the various kernel sizes, as illustrated in Fig 6. Integration of SPP and CSP block in YOLOv4 architecture is given in Fig 7.

4.3 Hyper-parameter Tuning

Manual tuning of the hyper-parameters ensures the best possible result. Table 1 contains information on these parameters. For both models, the values shown in Table 1 are proven to provide the best results. While the majority of the hyper-parameters are identical for both models, there are a few that are mentioned explicitly. The height and width of the input image are set to 416 and 416, respectively. Due to the fact that we are training on RGB images, the channel has been set to 3. Momentum is adjusted to 0.95 to increase the stability of the gradient. To further stabilize the model, we set decay to a very tiny value of 0.0005. burn_in is set to 1000 because the learning rate will gradually climb to our specified value after the first 1000 photographs. The steps are set to (4800, 5400) for the first model and (8000,9000) for the second model because we intended to tweak the learning rate after every (4800, 5400), (8000,9000) batches respectively to produce a near-perfect model. As the first model detects only one type of object, namely signboards with proper RTI information and the second model detect RTI which is shop name and shop information thus the classes parameter is set to 1 and 2 respectively. The filter parameter of the final convolutional layer before the Yolo layer is set to 18 and 21 for first and second detection model respectively.

Cut_Mix is set to 0 because we are performing a detection task, not classifying. Mosaic, random is set to 1, the mosaic feature creates a mosaic of four images and the random feature randomly resizes the image between 1/1.4 and 1.4 for every ten iterations, resulting in a highly generalized model. Both mosaic and random data augmentation functions are employed exclusively on the final layer. The batch size is set to 64 and subdivision is set to 16,32 for the first and second model respectively, as a result were able to pass 64/16 = 4 pictures in the

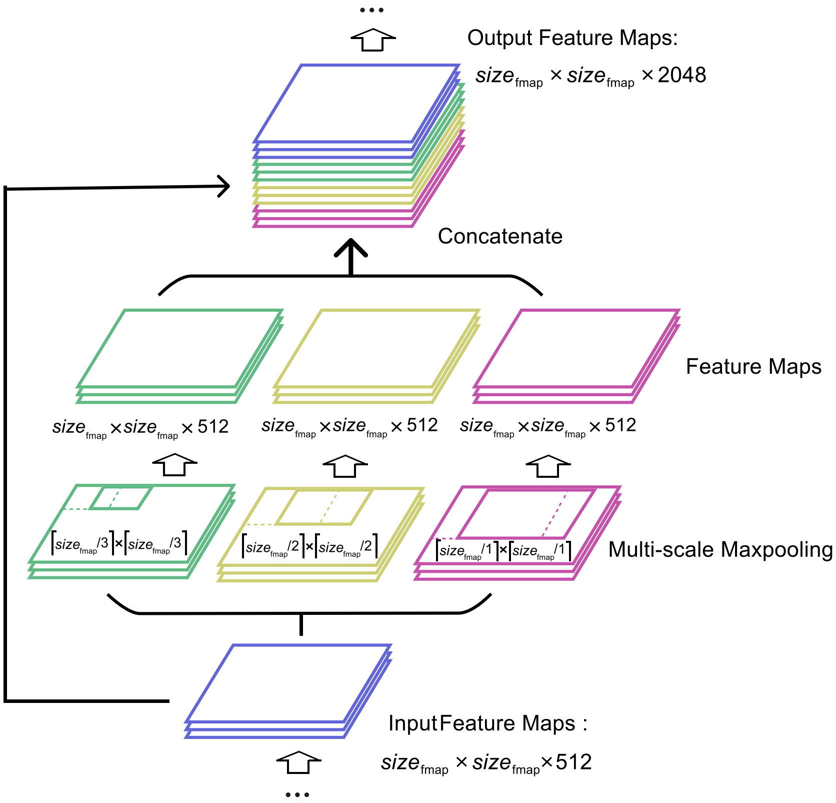
Output Feature Maps:
size_fmap × size_fmap × 2048
Concatenate
Feature Maps
size_fmap × size_fmap × 512
size_fmap × size_fmap × 512
size_fmap × size_fmap × 512
⌈size_fmap/3⌉×⌈size_fmap/3⌉
⌈size_fmap/2⌉×⌈size_fmap/2⌉
⌈size_fmap/1⌉×⌈size_fmap/1⌉
Multi-scale Maxpooling
Input Feature Maps :
size_fmap × size_fmap ×512

Fig. 6. SPP block.

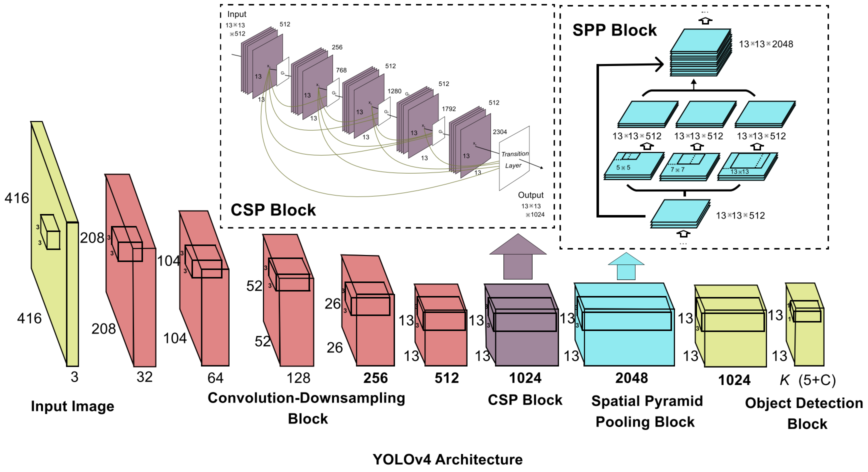

Fig. 7. YOLOv4 Architecture.

first model and 64/32 = 2 in the second model concurrently in each mini batch for faster convergence. Our model was trained over a period of 6000, 10000 iterations for first and second model respectively. After modifying the YOLOv4's defalt hyper-parameter, we trained and tested it on our dataset.

Table 1. Final Hyperparameters

Hyper-parameter	**Value**	**Hyper-parameter**	**Value**
Momentum	0.95	classes	1,2
channel	3	filter	18,21
decay	0.0005	Cut_Mix	0
burn_in	1000	mosaic	1
steps	(4800,5400) , (8000,9000)	random	1
batch size	64	subdivision	16,32

5 Results

Our model was trained and tested on a Nvidia GTX 1070 8GB(VRAM) GPU along with a Intel i7 7700k processor and 16GB of RAM. When tested on the validation dataset, the Yolov4 model demonstrated very high accuracy. This method is very accurate at detecting Bangla signboards and RTI in hazy, distorted, zoomed-in, poor lighting condition, and low-resulation images or video frames. Our proposed model outperforms comparable models. The Table 2 shows the accuracy comparison of signboard detection among our and other's proposed models. The average detection time is 42ms and 44ms for the first and second model respectively when tested for 100 images. Thus the total average execution time of the whole proposed system is 86ms which makes the model perform in real-time detection from video frames. Additionally, the model is intelligent enough to determine which signboards are useful and which are not based on the motivation of this research and detects only the useful ones. Here, useful signboards include those that display the shop's name and contact information (address and phone number) in Bangla. If neither of these two RTI are present on the signboard, the model will discard it. To summarise, we successfully transferred human expertise and knowledge in selecting useful signboards to the model. This is because of the diverse dataset we created, the meticulous annotation, and hyper-parameter tuning.

The AP is the weighted sum of precisions at each threshold, with the weight equal to the recall increment.

$$AP = \sum_{k=0}^{k=j-1} [r(k) - r(k+1)] * p(k) \tag{1}$$

The mAP is a weighted average of the AP values, which is itself a weighted average of all classes. The mAP is calculated by establishing a confidence level. We calculate the mAP across several IoU thresholds for each class j, and the final measure mAP across test data is produced by averaging all mAP values per class. The mAP is calculated using Equation 2.

$$mAP = \frac{1}{j} \sum_{i=0}^{i=j-1} AP_i \tag{2}$$

5.1 Signboard Detection Model Evaluation

The proposed system's first phase is to detect signboards in natural scenes. The first model is extremely accurate at detecting signboards in a variety of challenging environmental scenarios. After training for 6,000 epochs the model achieved an accuracy of 98.6%.

Fig 8 illustrates the detection of a signboard in images from a pristine environmental scenario. As seen in the first row, the model is capable of detecting multiple signboards within an image. In the first image of the second row, the model detects only the second signboard and ignores the first, because the model is intelligent enough to understand that the first signboard is missing the shop address, which is a required component of RTI (shop information). As illustrated in Fig 9, the model correctly detects signboards with a high degree of accuracy in challenging environmental scenarios involving wires crossing the signboards, images taken from varying angles, and varying lighting conditions. In addition, in the last image of the last row of Fig 9, there are two signboards, but the

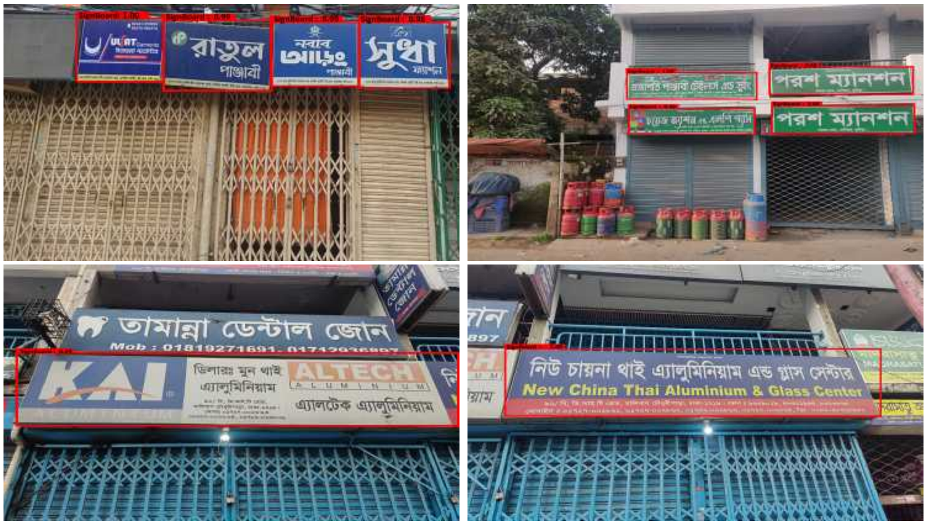

Fig. 8. Sample output of the first detection model.

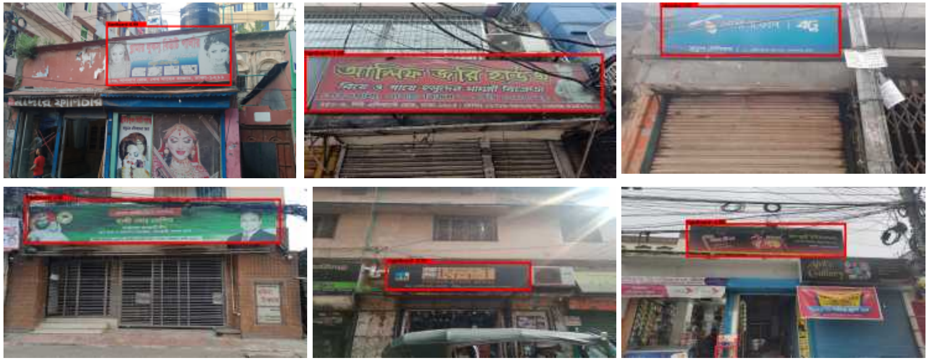

Fig. 9. Sample output of the first detection model.

model only detects one because the RTI of the undetected signboard is in English, which does not meet the requirements because this research is focused on signboards with Bangla RTI. As a result of being trained on a diverse dataset and having its hyperparameters fine-tuned, the proposed model has the ability to detect signboards, which is in line with the purpose of this research.

5.2 RTI Detection Model Evaluation

The second phase of the proposed system is to detect RTI using the output (detected signboards) of the first phase. The model was trained over a period of 10,000 epochs and it converged after 8000 epochs with a mAP of 97.4% and a loss of 0.89. Fig 10 illustrates that the RTI detection model can correctly detect RTI with high accuracy. Additionally, as illustrated in Fig 11, the model is capable of detecting RTI from challenging images, e.g., a degraded signboard, wires running through them, gloomy lighting condition. However the model is correctly able to detect RTI with high accuracy from these challenging environmental scenarios.

Table 2. Detection Accuracy Comparison

Name	**Proposed method**	**Accuracy**
Yohannes [21]	Deep attention network	70.92%
Toaha [17]	Faster RCNN	91%
Toaha [18]	Faster R-CNN	80%
Chen-Ya [5]	YOLOv2	83.7%
Our proposed model	**YOLOv4**	**Signboard detection: 98.6%, RTI detection: 97.4%**

6 Conclusion

The purpose of this research is to develop a real-time Automatic Bangla Signboard and RTI Detection system. This paper discusses an overall framework of signboard detection from natural scenes and RTI detection from detected signboards. A dataset containing around 17,308 images that covers nearly all possible real-world scenarios, e.g., a variety of lighting conditions, angles, pixelated images, both low and high resolution photos, various environmental conditions, etc is proposed in this research. A real-time Yolov4 model is used for both of the detection models. Extraction of text information from the detected RTI should be a primary focus of future study.

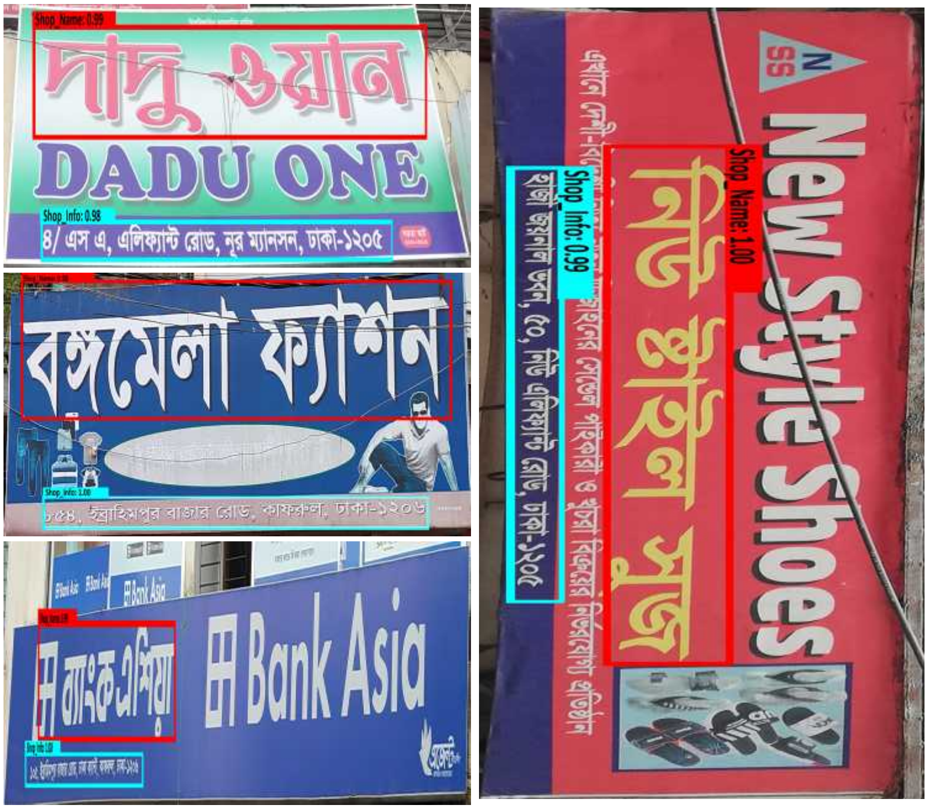

Fig. 10. Sample output of the second model.

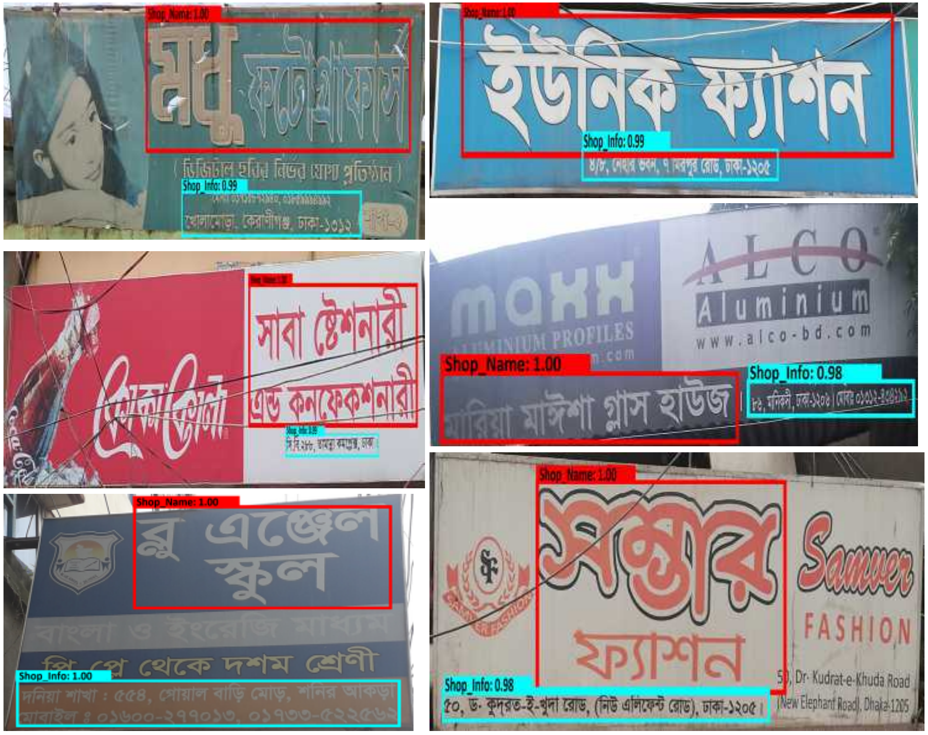

Fig. 11. Sample output of the second model.

7 Acknowledgment

We would like to express our appreciation and gratitude to the data curation team(link will be added in camera-ready submission) for their contributions to data curation.

References

1. Arafat, S.Y., Ashraf, N., Iqbal, M.J., Ahmad, I., Khan, S., Rodrigues, J.J.: Urdu signboard detection and recognition using deep learning. Multimedia Tools and Applications pp. 1–23 (2021)
2. Berriche, L., Al-Mutairy, A.: Seam carving-based arabic handwritten sub-word segmentation. Cogent Engineering **7**(1), 1769315 (2020)
3. Bochkovskiy, A., Wang, C.Y., Liao, H.Y.M.: Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934 (2020)
4. Chen, P.X., Rau, J.Y.: Using convolutional neural network for signboard detection on street view images. In: 39th Asian Conference on Remote Sensing: Remote Sensing Enabling Prosperity, ACRS 2018 (2018)
5. Hong, C.Y., Lin, C.Y., Shih, T.K.: Automatic signboard detection and semi-automatic ground truth generation. In: 2019 Twelfth International Conference on Ubi-Media Computing (Ubi-Media). pp. 256–261. IEEE (2019)
6. Malakar, S., Chiracharit, W.: Thai text detection and classification using convolutional neural network. In: 2020 59th Annual Conference of the Society of Instrument and Control Engineers of Japan (SICE). pp. 99–102. IEEE (2020)
7. Moyeen, M., Alam, K.M.R., Awal, M.A.: Bangla text extraction from natural scene images for mobile applications. J. Electr. Eng. Inst. Eng. EE **39** (2013)
8. Panhwar, M.A., Memon, K.A., Abro, A., Zhongliang, D., Khuhro, S.A., Memon, S.: Signboard detection and text recognition using artificial neural networks. In: 2019 IEEE 9th International Conference on Electronics Information and Emergency Communication (ICEIEC). pp. 16–19. IEEE (2019)
9. Park, J., Lee, G., Lai, A.N., Kim, E., Lim, J., Kim, S., Yang, H., Oh, S.: Automatic detection and recognition of shop name in outdoor signboard images. In: 2008 IEEE International Symposium on Signal Processing and Information Technology. pp. 111–116. IEEE (2008)
10. Rahman, M., Rahman, R., Supty, K.A., Sabah, R.T., Islam, M.R., Islam, M.R., Ahmed, N.: A real time abysmal activity detection system towards the enhancement of road safety. In: 2022 2nd International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET). pp. 1–5. IEEE (2022)
11. Rahman, R., Pias, T.S., Helaly, T.: Ggcs: A greedy graph-based character segmentation system for bangladeshi license plate. In: 2020 4th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT). pp. 1–7. IEEE (2020)
12. Rahman, R., Rakib, A.F., Rahman, M., Helaly, T., Pias, T.S.: A real-time end-to-end bangladeshi license plate detection and recognition system for all situations including challenging environmental scenarios. In: 2021 5th International Conference on Electrical Engineering and Information & Communication Technology (ICEEICT). pp. 1–6. IEEE (2021)
13. Rajesh, B., Javed, M., Nagabhushan, P.: Automatic tracing and extraction of text-line and word segments directly in jpeg compressed document images. IET Image Processing **14**(9), 1909–1919 (2020)

14. Sanasam, I., Choudhary, P., Singh, K.M.: Line and word segmentation of handwritten text document by mid-point detection and gap trailing. Multimedia Tools and Applications **79**(41), 30135–30150 (2020)
15. Shen, H., Tang, X.: Generic sign board detection in images. In: Proceedings of the 5th ACM SIGMM international workshop on Multimedia information retrieval. pp. 144–149 (2003)
16. Tam, A., Shen, H., Liu, J., Tang, X.: Quadrilateral signboard detection and text extraction. In: CISST. pp. 708–713 (2003)
17. Toaha, M., Islam, S., Asad, S.B., Rahman, C.R., Haque, S., Proma, M.A., Shuvo, M., Habib, A., Ahmed, T., Basher, M., et al.: Automatic signboard detection and localization in densely populated developing cities. arXiv preprint arXiv:2003.01936 (2020)
18. Toaha, M.S.I., Rahman, C.R., BinAsad, S., Ahmed, T., Proma, M.A., Haque, S.S.: Automatic signboard detection from natural scene image in context of bangladesh google street view. arXiv: 2003.01936 v2 (2020)
19. Wojna, Z., Gorban, A.N., Lee, D.S., Murphy, K., Yu, Q., Li, Y., Ibarz, J.: Attention-based extraction of structured information from street view imagery. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). vol. 1, pp. 844–850. IEEE (2017)
20. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: Cbam: Convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV). pp. 3–19 (2018)
21. Yohannes, E., Lin, C.Y., Shih, T.K., Hong, C.Y., Enkhbat, A., Utamininingrum, F.: Domain adaptation deep attention network for automatic logo detection and recognition in google street view. IEEE Access **9**, 102623–102635 (2021)