Efficient Video Object Detection of Indoor Furniture and Home Appliances

Fuangfar Pensiri
Department of Computer Science and Information
Faculty of Science at Sriracha
Kasetsart University, Sriracha Campus
Chonburi, Thailand
fuangfar.p@ku.th

Phasuwut Chunnapiya
Department of Computer and Information Science
Faculty of Applied Science
KMUTNB
Bangkok, Thailand
s6404062856048@email.kmutnb.ac.th

Wanida Khamprapai
Department of Computer Science and Information
Faculty of Science at Sriracha
Kasetsart University, Sriracha Campus
Chonburi, Thailand
wanida.kum@ku.th

Porawat Visutsak *
Department of Computer and Information Science
Faculty of Applied Science
KMUTNB
Bangkok, Thailand
porawatv@kmutnb.ac.th

2024 IEEE International Black Sea Conference on Communications and Networking (BlackSeaCom) | 979-8-3503-5185-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/BLACKSEACOM61746.2024.10646241

***Abstract*— Video object detection extends the principles of object detection in still images to the realm of moving pictures. It involves identifying and localizing objects of interest within each frame of a video sequence. In this work, we propose an efficient method for video object detection of indoor furniture and home appliances using YOLO, data augmentation, and a novel frame sampling technique. We tested our proposed method in a real-world environment, and performance measurements demonstrate a 24.73% increase in mean average precision (mAP50-95) and an 89.77% reduction in processing time compared to the standard YOLOv7 baseline.**

Keywords*— *video object detection, indoor video, intersection over union, yolo, deep learning

I. Introduction

Computer vision is a rapidly growing field in the technology and computer science world. One emerging technology in this field is video object detection. However, video object detection adds a layer of complexity by needing to track those objects across frames, maintaining their identities even when they move, become partially obscured, or change appearance. This technology has numerous applications, including self-driving vehicles, surveillance systems, traffic monitoring, and even analyzing sports footage. Video object detection technology is also expanding into various industries, and this paper focuses on its application in the indoor furniture and home appliances industry. The study aims to find ways to extract valuable information from indoor video to aid in room layout planning. Since SME furniture companies in Thailand involved have limited experience with deep learning, this paper suggests a method for quickly and effectively evaluating different video object detection models based on YOLO (You only look once). The process begins with creating a dataset of indoor furniture and home appliances and enhancing it through data augmentation techniques. Then, various open-source models (YOLO v5s, YOLO v6s, YOLO v7, and YOLO v8s) are tested on the dataset, with their performance measured.

We organize this work into 5 sections: 1) Introduction; 2) Literature Review; 3) Materials and Methods; 4) Result and Discussion; 5) Conclusion.

II. Literature Review

The motivation of this work begins with our previous works [1] and [2]. In [1], The paper describes a video object detection approach focused on leveraging depth information from sensors like the Kinect. To enhance accuracy, the paper proposes an ensemble method combining the benefits of both random forest and gradient boosting decision trees. Random forests offer improved accuracy by averaging multiple classifiers, while gradient boosting iteratively refines its models by correcting prior weaknesses. Additionally, it also introduces a grid-based optimization method for model development. This method employs a concept called “grid dominance” to compare and select solutions within the space of possible models. The grid structure facilitates a guided search process, leading to refinements in object detection performance. The combined use of depth data, ensemble learning, and grid-based optimization aims to create a more robust and accurate video object detection system. In [2], the paper focuses on video segmentation which is essential for identifying objects and events in videos. The paper proposes a method to improve video object segmentation, particularly for resource-limited environments. K-means clustering, a technique that groups similar data points together, is a popular tool for segmentation, but video data can be computationally expensive to process with standard K-means. The paper proposes using MATLAB's Just-In-Time (JIT) compilation to accelerate K-means clustering for video object segmentation. This JIT-based K-means approach aims to achieve faster

processing times while maintaining accuracy, making it more suitable for resource-constrained settings.

Recent works related to video object detection are also investigated. In [3], the paper tackles the challenges of extending traditional image-based object detection methods to the realm of video sequences. Video object detection introduces issues like motion blur and unusual object poses, degrading feature quality and detection accuracy. The proposed solution merges deep reinforcement learning for image-based detection with a correction module that utilizes object tracking algorithms. This combined approach treats a video as a series of images, where a reinforcement learning agent generates potential object regions. The correction module then refines these regions using tracking information derived from previous frames. Experimental results on the MOT 15 dataset demonstrate a significant improvement in detection performance compared to using the Faster-R CNN image detector alone. In [4], video object detection is difficult due to constantly changing object appearances and poor quality in some video frames. However, video provides an advantage - information from nearby frames can be used to improve detection in a specific frame. Existing methods for combining information across frames are complex and require expensive two-stage detectors. The paper proposes a new, efficient strategy for video object detection. It focuses processing power on important regions identified by a one-stage detector, avoiding wasted effort on low-quality areas. Additionally, it analyzes the relationship between the target frame and reference frames to guide the information aggregation process. This method achieves significant accuracy gains with minimal extra processing cost compared to current video object detection approaches. Notably, the YOLOX-based model using this strategy achieves impressive performance (over 87.5% accuracy) while maintaining real-time processing speeds, making it ideal for large-scale or real-time applications. The code for this method is also publicly available. In [5], the study explores using automated techniques for wildlife monitoring. The high costs and time commitment of traditional methods are being addressed with advancements in drones, sensors, and machine learning. The study trained YOLOv5 neural networks to identify interesting areas, hares, and roe deer in thermal aerial footage. It also developed a method to manually assess detection accuracy beyond the typical mean average precision (mAP) metric. This method revealed that a high mAP doesn't guarantee perfect detection and allows for a deeper understanding of the factors influencing model precision. Finally, the paper proposes a basic algorithm for real-time object detection using drones equipped with thermal sensors, zoom lenses, and laser rangefinders. This approach holds promise as a valuable tool for monitoring animals that are difficult to detect visually, particularly nocturnal or well-camouflaged species. In [6], underwater object detection is difficult because underwater videos are blurry and lack contrast. YOLO models are commonly used, but they struggle with these poor-quality videos and don't consider how objects relate across video frames. To address this, the paper proposes a new model called UWV-Yolox. UWV-Yolox first enhances video contrast, then uses a new module to improve how the model represents objects of interest. It also incorporates a new loss function and a frame-level optimization module that leverages connections between frames to refine detections. This combination significantly improves video detection performance. Experiments on a custom underwater video dataset (UVODD) show that UWV-YoloX achieves an mAP50 of 89.0%, outperforming the base YOLO model by 3.2%. The model also provides more stable object predictions compared to other approaches, and these improvements can be applied to other detection models.

We also reviews recent works of indoor object detection. In [7], indoor object detection is a growing field due to the vast amount of image and video data being used today in many applications especially in home decoration and room layout planning. Traditional methods were limited in training efficiency and could not leverage transfer learning. Deep learning and neural networks have revolutionized object detection, enabling real-time processing, multi-class detection, and transfer learning. The paper proposes a new object detection system based on the single shot detector (SSD) algorithm with MobileNetv2 for feature extraction. This system has potential applications in various fields like e-commerce, hospitality, security, real estate, self-driving cars, and inventory management. In [8], the paper presents a neural network-based method for robots to detect storage space within household furniture. The method involves automatically identifying and marking storage areas within 3D furniture models, then generating numerous depth images with labeled storage space locations. These images are then used to train a neural network for real-world storage detection using depth cameras. The paper also offers a dataset of depth images with storage space annotations for further research. This approach introduces a unique research area and demonstrates the potential to adapt object detection networks for identifying empty or filled storage spaces. In [9], the paper explores the use of object detection for kitchen planning purposes. Given the complexity of choosing the right deep learning model, the paper proposes a methodology for quickly and reliably evaluating different object detection models in the context of a kitchen industry use case. The process involves building a kitchen image dataset, applying data augmentation techniques, and then testing various freely available models (Faster R-CNN, SSD, EfficientDet) from the TensorFlow Object Detection API. The mean average precision (mAP) metric is used to determine the best-performing model. The paper aims to demonstrate the feasibility of object detection for kitchen planning within a company that may have limited deep learning experience.

III. Materials and Methods

This study focuses on video object detection of indoor furniture and home appliances. We aim to train YOLO models (YOLO v5s, YOLO v6s, YOLO v7, and YOLO v8s) with augmentation techniques such as blurring, filtering, flipping, and rotating to increase the number of images in the dataset. Our dataset consists of seven classes of indoor furniture and home appliances, totaling 18,333 images. Our preliminary study used the dataset from [10]. The seven classes of indoor furniture and home appliances used in this study include: Chair; Sofa; Table; Battery; Extinguisher; Air conditioner; Router. Figure 1 shows some examples of our dataset.

Fig. 1. Indoor furniture and home appliances dataset [10]

We use augmentation technique [11] to increase our images in dataset. Mostly, the augmentation in video includes:

1) Add objects: We add more furniture or home appliance objects to the image. This could give the model more data to train on and improve its ability to detect these objects in videos. For example, we could add a couch, a chair, a table, or a router.

2) Modify existing objects: We modify the existing objects in the image in different ways. For example, we could change the color of the furniture, rotate the objects to different positions, or occlude part of the object with another object.

3) Change the background: We change the background of the image to simulate different environments where furniture and home appliances might be found. For example, we could change the background to a living room, a kitchen, or an office.

5) Add text labels: We add text labels to the image to identify the different objects in the image. This would help the model learn the association between the object and its name.

6) Blur, filter, flip, and rotate: As mentioned in the prompt, we use techniques like blurring, filtering, flipping and rotating the image. This can create more variations of our dataset and help the model generalize better to unseen data.

By using these augmentation techniques, we can create a more diverse dataset that will help our YOLO models better detect furniture and home appliances in videos. Our augmentation algorithm in Python is shown in figure 2.

```
import albumentations as A
import cv2

image = cv2.imread('furniture_image.jpg')

transform = A.Compose([
    A.HorizontalFlip(p=0.5),  # Flip horizontally with 50% probability
    A.RandomBrightnessContrast(p=0.3),  # Adjust brightness/contrast
    A.Blur(blur_limit=3, p=0.2)  # Add some blur])

augmented_image = transform(image=image)['image']
cv2.imshow('Augmented Image', augmented_image)
cv2.waitKey(0)
```

Fig. 2. Augmentation pseudo-code

Where the 'p' values in the example control the probability of applying each transformation. Trial these values to find what works best for dataset. To train the video object detection models, we fine tune the following YOLOs: YOLO v5s, YOLO v6s, YOLO v7, and YOLO v8s.

YOLOv5s builds upon the popular YOLOv5 object detection framework, prioritizing speed and efficiency. It's perfect for real-time applications and devices with limited computational power. Architecturally, it leverages a CSPNet backbone and PANet neck for extracting image features. YOLOv6s introduces a significant shift from the YOLOv5 architecture, aiming for easy deployment and industrial use. It improves accuracy compared to YOLOv5 but might be slightly slower. Its architecture incorporates an EfficientRep backbone, Rep-PAN neck, and a decoupled head for greater flexibility. YOLOv7 refines the YOLOv6 design further for optimal speed and accuracy balance. It introduces trainable bag-of-freebies (BoF) to boost performance without sacrificing inference speed and uses an extended label assignment strategy (ELAN). YOLOv7 achieves state-of-the-art results on many benchmarks. YOLOv8s, the latest release, pushes the focus heavily towards accuracy while remaining competitive in speed. It introduces anchor-free heads, cutting-edge loss functions, and model scaling methods. YOLOv8s demonstrates remarkable accuracy gains, outperforming most other object detectors. We also compare and contrast each YOLO in table I.

TABLE I. YOLO OBJECT DETECTION USED IN OUR STUDY

Model	Architecture Changes	Focus	Strengths	Weaknesses
YOLO v5s	Based on YOLO v5	Speed	Fast, efficient, suitable for real-time	May slightly compromise accuracy
YOLO v6s	Significant shift	Deployment	Improved accuracy, ease of use	Potentially slower than YOLO v5
YOLO v7	Refinements of YOLO v6	Speed & Accuracy	Excellent balance, state-of-the-art	Slightly harder to train than some models
YOLO v8s	Further evolution	Accuracy	Even higher accuracy, flexibility	Complexity, resource requirements

IV. RESULT AND DISCUSSION

To evaluate our YOLO models, we consider mAP (mean Average Precision) which is the standard metric for object detection. It measures the average precision across all object classes, considering both correct detections and false positives (higher mAP is better). The following mAP parameters are considered:

1) mAP@0.5: mAP calculated with an IoU (Intersection over Union) threshold of 0.5. This means a predicted bounding box is considered correct if it overlaps with the ground truth box by at least 50%.

2) mAP@0.5:0.95: mAP calculated across different IoU thresholds from 0.5 to 0.95, giving a more comprehensive view of the model's performance.

Therefore, YOLOv5s demonstrates the highest mAP@0.5, indicating excellent object detection with standard overlap requirements. However, YOLOv8s has the highest

mAP@0.5:0.95, suggesting it delivers the most accurate bounding box predictions across a wider range of overlap thresholds. This makes YOLOv8s the most precise model overall. YOLOv7's lower mAP@0.5:0.95 could imply challenges stemming from the dataset or potential room for training optimization. When comparing the running times of YOLOv5 (2.095 hours) and YOLOv8 (2.520 hours), YOLOv5 emerges as the more suitable choice for testing video object detection of indoor furniture and home appliances in real-world scenarios due to its faster processing speed. To further illustrate this point, let's consider the running times of YOLOv5 (2.095 hours) and YOLOv7 (20.471 hours). This comparison reveals a significant 89.77% reduction in runtime with YOLOv5. Table II shows the validation results of our models. The confusion matrix of YOLO v5s is shown in figure 3.

TABLE II. THE VALIDATION RESULTS OF YOLO MODELS

Model	Precision	Recall	mAP@0.5	mAP@0.5: 0.95	F1 Score
YOLO v5s	0.965	0.971	0.986	0.852	0.967
YOLO v6s	0.982	0.964	0.985	0.898	0.971
YOLO v7	0.931	0.903	0.95	0.736	0.916
YOLO v8s	0.974	0.99	0.917	0.918	0.981

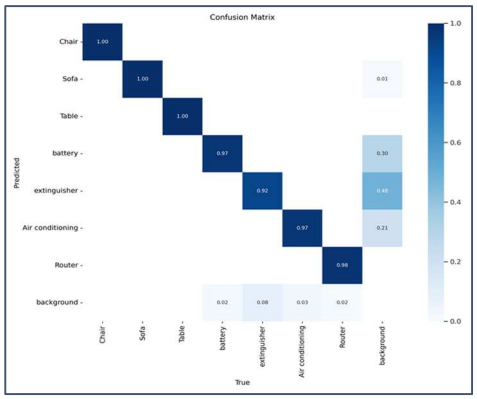

Fig. 3. The confusion matrix of YOLO v5s

V. CONCLUSION

In this study, we investigate YOLO-based video object detection models for extracting information from indoor videos to assist with room layout planning. Our findings suggest that YOLOv8s might be the most accurate model overall, achieving the highest mAP@0.5:0.95 (0.918), indicating excellent precision in bounding box predictions across various overlap thresholds. However, YOLOv5s remains a strong contender due to its impressive mAP@0.5 (0.986) and significantly faster processing speed (2.095 hours) compared to YOLOv8s (2.520 hours). Notably, YOLOv5s offers an 89.77% reduction in runtime compared to YOLOv7 (20.471 hours), making it more suitable for real-world scenarios with resource limitations.

Future Work: To build upon this study, several promising directions exist. Fine-tuning YOLOv5s and YOLOv8s on the indoor furniture and appliance dataset could further enhance their accuracy for this specific domain.

ACKNOWLEDGMENT

"This research was funded by King Mongkut's University of Technology North Bangkok, Contract no. KMUTNB-67-BASIC-27".

REFERENCES

[1] P. Visutsak and F. Pensiri, "Optimization of Hue, Brightness, Luminance, and Saturation Parameters for Video Segmentation Based on Evolutionary Algorithms," ICIC Express Letters - An International Journal of Research and Surveys, ISSN 1881-803X, vol. 18, No. 6, June 2024.

[2] W. Khamprapai, F. Pensiri, and P. Visutsak, "Efficient Video Object Segmentation using JIT K-Means Clustering" Proc. of the 2024 16th International Conference on Knowledge and Smart Technology (KST 2024), February 28 - March 2, 2024, Krabi, Thailand [Online].

[3] R. Fei and L. Ou, "A Video Object Detector Based on Deep Reinforcement Learning and Correction Module," 2020 IEEE 3rd International Conference on Electronics and Communication Engineering (ICECE), Xi'An, China, 2020, pp. 1-5. doi: 10.1109/ICECE51594.2020.9353034.

[4] Y. Shi, N. Wang, and X. Guo, "YOLOV: making still image object detectors great at video object detection," In Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence (AAAI'23/IAAI'23/EAAI'23), Vol. 37. AAAI Press, Article 251, 2023, pp. 2254–2262. doi: 10.1609/aaai.v37i2.25320.

[5] P. Povlsen, D. Bruhn, P. Durdevic, D. O. Arroyo, and C. Pertoldi, "Using YOLO Object Detection to Identify Hare and Roe Deer in Thermal Aerial Video Footage—Possible Future Applications in Real-Time Automatic Drone Surveillance and Wildlife Monitoring," Drones 8, no. 1: 2, 2024. doi: 10.3390/drones8010002.

[6] H. Pan, J. Lan, H. Wang, Y. Li, M. Zhang, M. Ma, D. Zhang, and X. Zhao, "UWV-Yolox: A Deep Learning Model for Underwater Video Object Detection," Sensors 2023 (Basel, Switzerland), 23(10), 4859. doi: 10.3390/s23104859.

[7] N. Darapaneni et al., "Object Detection of Furniture and Home Goods Using Advanced Computer Vision," 2022 Interdisciplinary Research in Technology and Management (IRTM), Kolkata, India, 2022, pp. 1-5, doi: 10.1109/IRTM54583.2022.9791508.

[8] M. Hržica, P. Pejić, I. H. Tolić, and R. Cupec, "Detection of Household Furniture Storage Space in Depth Images," Sensors 2022, 22, 6774. doi: 10.3390/s22186774.

[9] B. Stecker and H. B. Pook, "Application of Object Detection Models for the Detection of Kitchen Furniture - A Comparison," In Artificial Intelligence and Soft Computing: 22nd International Conference, ICAISC 2023, Zakopane, Poland, June 18–22, 2023, Proceedings, Part II. Springer-Verlag, Berlin, Heidelberg, 91–101. doi: 10.1007/978-3-031-42508-0_9.

[10] https://universe.roboflow.com/teleradtech/homeinterior_object_detection [Online].

[11] P. Visutsak, X. Liu, K. H. Ryu, N. Bussabong, N. Sirikong, P. Intamong, W. Sonnui, S. Boonkerd, J. Thongpiem, M. Poonpanit, A. Homwiseswongsa, K. Hirunwannapong, C. Suksomsong, R. Budrit, "Transfer Learning for Caladium bicolor Classification: Proof of Concept to Application Development," KSII Transactions on Internet and Information Systems, Vol. 18, No. 1, pp. 126-146, 2024. doi: 10.3837/tiis.2024.01.008.