A FINE-GRAINED OBJECT DETECTION MODEL FOR AERIAL IMAGES BASED ON YOLOV5 DEEP NEURAL NETWORK

Chinthakunta Srilatha[1], Subramanian K.M[2], Mohammed Waheduddin Hussain[3]
[1]PG Scholar, Dept of CSE, Shadan College of Engineering and Technology, Hyderabad, Telangana,
psrilathareddysrilathareddy@gmail.com
[2]Professor, Dept of CSE, Shadan College of Engineering and Technology, Hyderabad, Telangana.
kmsubbu.phd@gmail.com
[3]Professor, Dept of IT, Shadan College of Engineering and Technology, Hyderabad, Telangana.
mwaheeduddinhussain@gmail.com

ABSTRACT

A real-time visual tracking system with an active camera is implemented and described in the study. The system's purpose is to track human mobility indoors. Frame differencing and camera motion compensation are the foundations of quick and easy motion detection techniques that enable real-time tracking. The outcomes of the online person tracking are shown. The multiple objects tracking approach maintains a graph structure where it maintains multiply hypotheses on the number of aids the object trajectories in the video have, based on initial object detection results in each image, which may have missing and/or inaccurate detection. The graph is extended and pruned based on the image information, which also establishes the optimal hypothesis. Although the tracking process confirms and validates the detection over time, image-hued object detection makes a local judgment. As a result, it can be seen of as temporal detection, which makes a global decision over time. The multiple object detection approach provides the object detection module with feedback in the form of object position predictions. Thus, object detection and hauling are strongly integrated in the technique. Multiple objects tracking results are provided by the most plausible hypothesis. Results from the experiment are shown.

INTRODUCTION

The deployment of a real-time visual tracking system with an active camera that is intended for indoor human motion tracking is described in the study. This system achieves real-time tracking capabilities by using quick and effective motion detection processes, mainly based on frame differencing and camera motion compensation. The study's main focus is online person tracking, and the report presents the findings from this real-time tracking method.

In order to overcome issues like missing or erroneous detections in initial object detection findings, the study presents a multiple object tracking technique that preserves a graph structure and permits several hypotheses on the quantity and trajectory of objects in the video. In order to identify the optimal hypothesis, the procedure entails expanding and trimming the graph according to picture data.

In contrast to image-based object identification, which makes decisions locally, tracking successfully does temporal detection and makes global decisions over time by validating and confirming these detections. This integrated approach closely combines object detection and tracking since the object detection module receives feedback from the multiple object tracking method in the form of object location predictions. This collaborative procedure produces the multiple object tracking data, which are the most likely hypothesis. To confirm the efficacy of the suggested strategy, experimental findings are shown at the end of the paper.

OBJECTIVE

The proposed system with an active camera for aerial picture object detection aims to develop an efficient and responsive tracking framework. In order to accomplish low-latency tracking, the system optimizes quick motion detection techniques such frame differencing and camera motion correction. Online person tracking, which ensures that people are regularly and comprehensively monitored in indoor situations, is another aspect of the focus.

To solve problems of missing or incorrect detections, a multiple object tracking strategy is explored, which uses a graph structure to test and evaluate theories about the quantity and routes of items. A primary goal is the seamless integration of object identification and tracking, with a feedback loop between the multiple item tracking technique and the object detection module to continuously increase accuracy. With the use of image-hued object detection, the system is designed to exhibit temporal detection capabilities. Local conclusions are then confirmed over time for global decision-making. To show the system's practical effectiveness, extensive experimental validation is conducted, and its adaptability is highlighted by examining cross-domain applicability.

PROBLEM STATEMENT

In order to follow people's movements indoors, the project focuses on creating a real-time visual tracking system. It is not possible to manage the complexity of monitoring many objects with missing or

erroneous initial detection findings using conventional techniques like frame differencing and camera motion compensation. A tracking system that is accurate, effective, and able to manage several objects at once while adapting to changing conditions is the biggest obstacle. In order to recognize and track objects, the system must overcome problems like occlusion, cluttered backdrops, and changing lighting. In order to ensure tracking accuracy and robustness even in cases when detection results are not ideal, it is crucial to keep track of object trajectories consistently across time.

EXISTING SYSTEM

In computer vision research, multiple item detection has proven to be a difficult subject. In addition to the challenges associated with multiple object tracking, such as inter-object occlusion and multi-object concision, it must address the challenges of single object tracking, such as changing appearances, non-rigid motion, dynamic illumination, and occlusion. Many efforts have been made to visual track multiple objects. Their method for tracking a set number of objects is a sampling algorithm. To track numerous individuals, Tao et al. provide an effective hierarchical approach.

Disadvantage of Existing System

- Limited Cross-Platform Support
- Performance Restrictions, and Advanced Image Processing Methods

PROPOSED SYSTEM

The implemented technique functions in object tracking by embracing the probability obtained from initial object detection. As seen graphically in Figure 2, the tracking procedure entails preserving several hypotheses about object paths inside a structured graph. The number of objects being tracked and the accompanying trajectories they follow are contained in each hypothesis. The method starts the tracking process by adding the most recent object detection results to the current graph. This means coming up with a wide range of theories that take into consideration possible paths based on the recently obtained object detection data.

A closer look demonstrates how the system dynamically adjusts to changing item detection data. In order to allow for the examination of multiple possible trajectories that correspond with the identified items, it carefully integrates these new discoveries into the current graph structure. As a result, the algorithm generates several hypotheses, each of which represents a tenable collection of object trajectories. The algorithm's flexibility and adaptability are improved by this method, which enables it to investigate various scenarios during the continuous tracking process and take object detection uncertainties into consideration. As a result, the algorithm is able to manage a wide range of tracking situations, which makes it appropriate for applications where object trajectories may be unpredictable and variable.

Advantages of Proposed System

- Simple to Use
- Adaptable, and Compatible with Multiple Architecture

RELATED WORKS

In the literature, a number of methods for tracking numerous objects have been put forth. Previous approaches, which included strategies like Bayesian tracking and Kalman filters, were mainly concerned with tracking individual objects. However, these methods have trouble managing several things at once, particularly when those objects are partially or completely obscured. The tracking problem is now modeled as a set of hypotheses in graph-based tracking strategies, which are the more current approaches. For every object trajectory, these methods keep track of several hypotheses that can be expanded or reduced in response to fresh frame data, enabling increasingly precise monitoring over time. To improve object tracking outcomes, methods like Maximum Likelihood Estimation (MLE) and the Hungarian algorithm for data association have been incorporated. Combining these techniques improves object detection and tracking accuracy across frame sequences by enabling the tracking system to make decisions globally across time. Temporal detection is a crucial notion in this context because it offers a feedback loop that improves detection outcomes and preserves object trajectory coherence.

METHODOLOGY OF PROJECT

Dataset Preparation:

The preparation of the dataset is necessary in order to train YOLOv5 for object detection in aerial photos. The dataset must be representative and varied. This dataset ought to contain annotated samples of the required items, such buildings, cars, or other properties.

Model Configuration:

Customers can select a model size based on the trade-off between speed and accuracy because Yolov5 is available in small, medium, and large sizes. The process of configuring the model entails choosing a suitable YOLOv5 variation and modifying anchor box sizes and aspect ratios in accordance with the features of the aerial dataset..

Pre-trained Models:

Starting with pre-trained models on large general datasets (such as COCO) can improve YOLOv5's performance. Adapting the model to the particular characteristics and difficulties of aerial images is

facilitated by fine-tuning it on the particular aerial dataset

Object Classes:
List the different classes of items that are shown in the aerial photos. These could include infrastructure, vegetation, buildings, cars, etc., depending on the application. In the training data, each class must have an annotation.

Data Augmentation:
Different lighting conditions, perspectives, and scales can be seen in aerial photos. To enhance the training dataset's diversity and the model's capacity for generalization, data augmentation techniques including rotation, flipping, and scaling are used.

Training Process:
YOLOv5 is trained on the dataset of tagged aerial images. The model's parameters are optimized during training in order to precisely anticipate the bounding boxes and class labels for the objects in the pictures. The model is trained until it performs satisfactorily on the validation set.

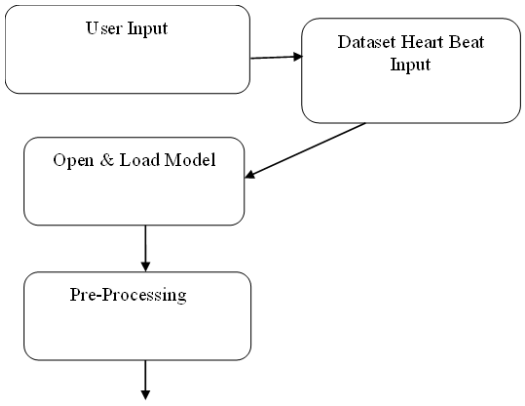
Inference on Aerial Images:
After training, the model can be used to make inferences on fresh, untested aerial photos. Bounding box coordinates and related class predictions are provided by the model once it has identified and located objects of interest.

ALGORITHM USED IN PROJECT
YOLOv5:
To use YOLO (You Only Look Once), an image is divided into a grid, and predictions are made for each grid cell. The task of estimating bounding boxes and related class probabilities falls to each cell. This indicates that the network concurrently predicts an object's location and class. YOLOv5 expands on this core idea by improving the architecture to increase precision and effectiveness. Improved backbone networks, sophisticated attention algorithms, and more intricate feature pyramids are among the optimizations it brings to better capture context from objects of various sizes. YOLOv5 can now function better in demanding settings with intricate backgrounds and a range of item sizes thanks to these improvements. Furthermore, YOLOv5 employs a more straightforward and computationally efficient structure than its predecessors, enabling quicker inference times without sacrificing detection quality. Because of this, it may be used in real-time in a variety of fields, such as surveillance, autonomous driving, and aerial images.

High-resolution Aerial Images:
since of their high spatial resolution, aerial photos are ideal for applications requiring precise object detection since they capture minute features. These photos offer an aerial perspective of the terrain, which is highly beneficial for uses such as environmental monitoring, urban planning, and surveillance. YOLOv5's capacity to effectively manage massive amounts of data and capture spatial relationships in a single network pass makes it especially well-suited for processing such high-resolution photos. Even in intricate situations with disparate object sizes and densities, it can swiftly recognize objects thanks to its improved feature extraction capabilities and streamlined architecture. For activities like following moving objects, identifying changes in land use, and keeping an eye on natural disasters from above, YOLOv5 is therefore appropriate for real-time analysis of aerial photography. Because of its rapid speed and precision of detection, it is the perfect instrument for making decisions and taking action in urgent situations.

Single-pass Inference:
By analyzing the entire image in a single pass, YOLO's "You Only Look Once" philosophy highlights its capacity for real-time item detection. This indicates that YOLO can process an image quickly and effectively, predicting class probabilities and bounding boxes in a single forward network run. In order to preserve its real-time capabilities, YOLOv5 significantly optimizes the architecture, carrying on this legacy. This effectiveness is especially useful for aerial imagery, where prompt decision-making and action are necessary for time-sensitive applications including traffic monitoring, surveillance, and disaster response. YOLOv5 is the perfect option for a variety of real-time applications in aerial environments because it makes crucial information instantly available to decision-makers by lowering latency and speeding up processing time.

7. DATA FLOW DIAGRAM



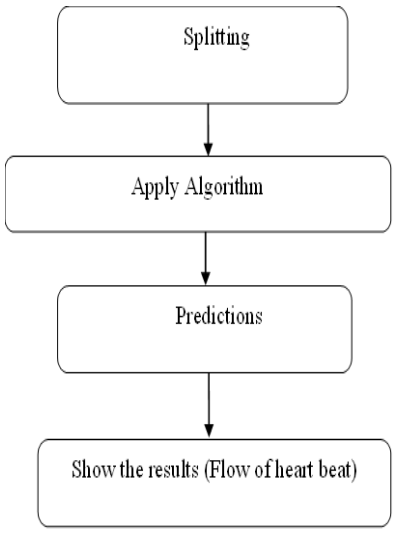


Fig: 7 Flow Diagram

8. SYSTEM ARCHITECTURE

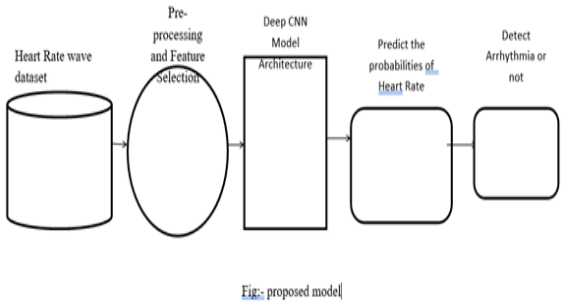


Fig: 8 System Architecture

9. RESULTS

FUTURE ENHANCEMENT

Enhanced Generalization and Accuracy:
The goal of ongoing research and development is to improve YOLOv5's architecture and training methods in order to increase its accuracy and generalizability across a variety of aerial datasets. Researchers want to make YOLOv5 even more resilient in processing intricate and diverse aerial imagery by refining feature extraction procedures, integrating cutting-edge strategies like attention mechanisms, and fine-tuning the model's layers.

The effectiveness of edge devices:
Prioritizing YOLOv5 optimization for deployment on edge devices, like drones or embedded systems on aerial platforms, may be a future improvement. Model compression methods and effective inference approaches to deal with resource limitations may be used in this.

Connecting Multimodal Data:
An even more thorough understanding of the aerial environment would be possible through the integration of YOLOv5 with additional sensors and data sources, such LiDAR or hyper spectral images. This could use complementary data modalities to improve object detection.

Combining Semantic Segmentation:
The model's comprehension of the context of observed objects in the scene may be improved by combining YOLOv5 with semantic segmentation approaches. In intricate aerial scenes with overlapping objects, this could be especially helpful.

Robustness against adversaries:
Future advancements could concentrate on strengthening YOLOv5's resistance to hostile attacks, guaranteeing its dependability in practical situations where possible aerial images manipulations might take place.

Active Education and Gradual Training:
By using active learning strategies and allowing for incremental training, YOLOv5 may be able to adjust to changing aerial datasets more effectively, picking up new skills and enhancing its functionality over time

The concepts of interpretability and explain ability:
Improvements could be made to give the model's predictions more compelling explanations. This is especially significant for situations where it is essential to comprehend the logic underlying object detection judgments.

CONCLUSION

To sum up, the use of YOLOv5 for object detection in aerial photos is a potent combination of aerial monitoring and deep learning. For activities ranging from environmental monitoring and disaster response to surveillance and urban planning, YOLOv5 is a powerful tool because to its real-time processing capabilities, accurate object location, and flexibility to a variety of objects. It is perfect for real-time analysis in crucial situations due to its effective handling of enormous amounts of data and its capacity to detect objects quickly and accurately.

Future developments in this area are probably going to concentrate on improving the model's generalization and accuracy. In order to ensure that YOLOv5 can function well on platforms with limited resources, such drones or satellites, researchers may try to optimize it for deployment on edge devices. A more thorough grasp of aerial surroundings could be obtained by integrating multimodal data sources, such as optical images with thermal or radar data, which would enhance detection and analysis capabilities even more. Furthermore, it is crucial to enhance explain ability, which makes the model's decision-making process easier for end users to comprehend. The reliability and applicability of the model across many use cases will be improved by addressing robustness to adversarial assaults and integrating domain-specific pre-trained models. As technology develops, the combination of YOLOv5 and aerial images presents exciting opportunities for applications that need to analyse high-resolution aerial data quickly and accurately.

REFERENCES:

[1] Cheng, P. Zhou, and J. Han, “Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing image,” IEEE Transactions on Geoscience and Remote Sensing, vol.54, no.12, pp.7405–7415, 2016.

[2] .Li, G. Wan, G. Cheng, L. Meng, et al., “ Object detection in optical remote sensing images: A survey and a new benchmark,” ISPRS Journal of Photogrammetry Remote Sensing, vol.159, pp.296–307, 2020.

[3]T.Y. Lin, M. Maire, S. Belongie, et al., "Microsoft COCO: Common objects in context," in Proceedings of European Conference on Computer Vision, Springer, Cham,pp.740–755, 2014
[4] M. Everingham, L. Van Gool, C. K. Williams, et al., "The PASCAL visual object classes (VOC) challenge," International Journal of Computer, vol.88, no.2, pp.303–338, 2010.
[5] A. Bochkovskiy, C. Y. Wang, and H. Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," arXiv preprint, arXiv: 2004.10934, 2020.
[6] K. Duan, S. Bai, L. Xie, et al., " Centernet: Key point triplets for object detection," in Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, pp.6568–6577, 2019.
[7] R. Girshick, J. Donahue, T. Darrell, et al., " Rich feature hierarchies for accurate object detection and semantic segmentation," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Venice, Italy, pp.580–587, 2014.
[8]T .Y. Lin, P. Goyal, R. Girshick, et al., "Focal loss for dense object detection," in Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, pp.2980–2988, 2017.
[9] W. Liu, D. Anguelov, D. Erhan, et al., " SSD: Single shot multibox detector," in Proceedings of the European Conference on Computer Vision, Springer, Cham, pp.21–37, 2016.
[10] J. Redmon, S. Divvala, R. Girshick, et al., "You only look once: Unified, real-time object detection," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, pp.779–788, 2016
[11] M. Azimi, E. Vig, R. Bahmanyar, et al., "Towards multiclass object detection in unconstrained remote sensing imagery," in Proceedings of Asian Conference on Computer Vision, Springer, Cham, pp.150–165, 2019
[12] . Zhang, S. Lu, and W. Zhang, " CAD-Net: A contextaware detection network for objects in remote sensing imagery," IEEE Transactions on Geoscience Remote Sensing,vol.57, no.12, pp.10015–10024, 2019.
[13] . Han, J. Ding, N. Xue, et al., "ReDet: A rotation-equivariant detector for aerial object detection," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, pp.2768–2795, 2021.
[14] X. Yang, J. Yang, J. Yan, et al. " SCRDet: Towards more robust detection for small, cluttered and rotated objects," in Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, pp.8231–8240, 2019.
[15] J. Ding, N. Xue, Y. Long, et al., "Learning roi transformer for oriented object detection in aerial images," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA,pp.2844–2853, 2019.
[16] J. Han, J. Ding, J. Li, et al., " Align deep features for oriented object detection," IEEE Transactions on Geoscience and Remote Sensing, vol.60, pp.1–11, 2021.
[17] . Yang, J. Yan, Z. Feng, et al., " R3Det: Refined singlestage detector with feature refinement for rotating object," in Proceedings of the 35th AAAI Conference on Artificial Intelligence, Virtual Event, pp.3163–3171, 2021.
[18] X. Yang and J. Yan. " Arbitrary-oriented object detection with circular smooth label," in Proceedings of European Conference on Computer Vision 2020, LNCS, vol.12353,Springer, Cham, pp.677–694, 2020.
[19] G. S. Xia, X. Bai, J. Ding, et al., " DOTA: A large-scale dataset for object detection in aerial images," in Proceedings of the IEEE Conference on Computer Vision and Pat- tern Recognition, Salt Lake City, UT, USA, pp.3974–3983, 2018.
[20] X. Sun, P. Wang, Z. Yan, et al., " FAIR1M: A benchmark dataset for fine-grained object recognition in high-resolution remote sensing imagery," ISPRS Journal of Photogrammetry Remote Sensing, vol.184, pp.116–130, 2022.
[21] X. Yang, H. Sun, K. Fu, et al., " Automatic ship detection in remote sensing images from google earth of complex scenes based on multiscale rotation dense feature pyramid networks," Remote Sensing , vol.10, no.1, article no.132, 2018.
[22 K. Fu, Z. Chang, Y. Zhang, et al., " Rotation-aware and multi-scale convolutional neural network for object detection in remote sensing images," ISPRS Journal of Photogrammetry Remote Sensing, vol.161, pp.294–308, 2020.
[23] Z. Liu, H. Wang, L. Weng, et al., " Ship rotated bounding box space for ship extraction from high-resolution optical satellite images with complex backgrounds," IEEE Geoscience Remote Sensing Letters, vol.13, no.8, pp.1074– 1078, 2016.
[24] S. Ren, K. He, R. Girshick, et al., " Faster R-CNN: To- wards real-time object detection with region proposal net- works," IEEE Transactions on Pattern Analysis and Ma- chine Intelligence, vol.39, no.6, pp.1137–1149, 2017.
[25] L. Zhou, H. Wei, H. Li, et al., " Objects detection for re- mote sensing images based on polar coordinates," arXiv pre- print, arXiv: 2001.02988, 2020.
[26] J. Yi, P. Wu, B. Liu, et al., " Oriented object detection in aerial images with box boundary-aware vectors," in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, pp.2149– 2158, 2021.
[27] W. Li, Y. Chen, K. Hu, et al., "Oriented reppoints for aerial object detection," in Proceedings of the

IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, Louisiana, pp.1829–1838, 2022.
[28] X. Yang, X. Yang, J. Yang, et al., "Learning high-precision bounding box for rotated object detection via kullbackleibler divergence," Advances in Neural Information Processing Systems, vol.34, pp.18381–18394, 2021.
[29] X. Yang, J. Yan, Q. Ming, et al., " Rethinking rotated object detection with Gaussian Wasserstein distance loss," in Proceedings of the International Conference on Machine Learning, Vienna, Austria, pp.11830–11841, 2021.
[30] J. Hu, L. Shen, and G. Sun, " Squeeze-and-excitation net-works," in Proceedings of the IEEE Conference on Com- puter Vision and Pattern Recognition, Salt Lake City, UT,
[31] S. Woo, J. Park, J. Y. Lee, et al., " CBAM: Convolutional block attention module," in Proceedings of the European Conference on Computer Vision (ECCV), Munich, Ger- many, pp.3–19, 2018
[32] H. Zhao, J. Jia, and V. Koltun, "Exploring self-attention for image recognition," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, pp.10073–10082, 2020.
[33] A. Srinivas, T. Y. Lin, N. Parmar, et al., "Bottleneck trans- formers for visual recognition," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, pp.16514–16524, 2021.
[34] A. F. Agarap, " Deep learning using rectified linear units (ReLU)," arXiv preprint, arXiv: 1803.08375, 2018.
[35] B. Xu, N. Wang, T. Chen, et al., " Empirical evaluation of rectified activations in convolutional network," arXiv pre- print, arXiv: 1505.00853, 2015.
[36] D. Misra, " Mish: A self regularized non-monotonic activa- tion function," arXiv preprint, arXiv: 1908.08681, 2019.
[37] J. Deng, W. Dong, R. Socher, et al., "A large-scale hierarch- ical image database," in Proceedings of IEEE Computer Vision and Pattern Recognition, Miami, FL, USA, pp.248–255, 2009.
[38] M. Tan and Q. Le, " Efficientnet: Rethinking model scaling for convolutional neural networks," in Proceedings of the International Conference on Machine Learning, Long Beach, California, USA, pp.6105–6114, 2019.
[39] N. Ma, X. Zhang, M. Liu, et al., "Activate or not: Learning customized activation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, pp.8028–8038, 2021.