Leveraging YOLO-World and GPT-4V LMMs for Zero-Shot Person Detection and Action Recognition in Drone Imagery

Christian Limberg[1], Artur Gonçalves[1], Bastien Rigault[1] and Helmut Prendinger[1]

***Abstract*— In this article, we explore the potential of zero-shot Large Multimodal Models (LMMs) in the domain of drone perception. We focus on person detection and action recognition tasks and evaluate two prominent LMMs, namely YOLO-World and GPT-4V(ision) using a publicly available dataset captured from aerial views. Traditional deep learning approaches rely heavily on large and high-quality training datasets. However, in certain robotic settings, acquiring such datasets can be resource-intensive or impractical within a reasonable timeframe. The flexibility of prompt-based Large Multimodal Models (LMMs) and their exceptional generalization capabilities have the potential to revolutionize robotics applications in these scenarios. Our findings suggest that YOLO-World demonstrates good detection performance. GPT-4V struggles with accurately classifying action classes but delivers promising results in filtering out unwanted region proposals and in providing a general description of the scenery. This research represents an initial step in leveraging LMMs for drone perception and establishes a foundation for future investigations in this area.**

I. INTRODUCTION

Recent advances in Large Language Models (LLMs) have transformed many aspects of Machine Learning and AI [1], [2]. Previously, the most common approach involved gathering datasets that captured small contexts within specific task domains. However, with the advent of foundation models such as LLMs, trained on much larger datasets, this paradigm has shifted. These models can now be utilized by providing them with prompts that specify the domain and task. Thanks to their strong generalization abilities, these foundation models can often be applied in a zero-shot manner [3]. While LLMs were originally designed for processing text in Natural Language Processing (NLP) tasks, Large Multimodal Models (LMMs) have expanded their capabilities by incorporating additional modalities [4], [5], [6] such as images, sounds, and videos.

In this article, we delve into the application of two recent image-based LMMs within a drone setting. Firstly, we examine the YOLO-World model [7], which facilitates prompt-based object detection. Secondly, we utilize the more general vision model GPT-4V [4] for classifying the detected region proposals.

An important challenge in aerial robotics is ensuring that drones operate reliably across a wide spectrum of potential failures. This necessitates the acquisition of a high-quality, problem-specific dataset, which can be resource-intensive or even impractical to obtain. Moreover, conventionally trained models tend to excel only within the confines of their training data. Minor variations in the environment, such as changes in weather, seasonal fluctuations, or geographical differences, can lead to a significant decline in the robot's reliability.

LMMs, trained on a broader contextual scope, may not achieve competitive performance compared to their traditional counterparts within their narrowly defined training contexts [1]. However, their ability to generalize across domains, facilitated by their training on significantly broader ranges of data, enables them to better handle challenging conditions.

This preliminary study investigates the feasibility of applying YOLO-World and GPT-4V in a practical aerial robotic scenario involving person detection and action recognition. A real-world application could entail locating individuals in need following a disaster [8], [9]. Given the unpredictable nature of potential disasters, it is crucial to utilize a model with extensive generalization capabilities, capable of operating effectively across diverse settings. Both YOLO-World and GPT-4V are zero-shot approaches prompted with text. This means they could potentially be deployed in unforeseen scenarios, as the text prompts can be adjusted quickly, enabling the equipped robot to adapt to entirely different objectives instantly.

This manuscript is structured as follows: Section II provides an overview of the most important related works. Section III-A discusses the publicly available Okutama-Action dataset that we are utilizing for our evaluation. In Section III-B, we focus on detecting persons using YOLO-World. In Section III-C, we apply GPT-4V on the detected region proposals to recognize the persons' actions. Finally, Section IV summarizes our findings and concludes the paper.

II. RELATED WORK

From the very beginnings of computer vision research, object detection has been a prominent task of interest. First, hand-crafted features [10] were utilized for detecting and recognizing objects. With the rise of deep learning, convolutional neural networks, which derive features automatically from the training data, quickly overtook established hand-crafted methods in terms of accuracy and robustness [11]. "Two-stage" methods like R-CNN [12] and R-FCN [13] first detect candidate regions proposals, and then classify them.

Later, "one-stage" methods such as SSD [14] and YOLO [15] established themselves, achieving higher processing speeds by detecting and classifying objects with one forward pass through the network. YOLO in particular has become extremely popular object detection method, with

*This work was supported by a fellowship within the IFI program of the German Academic Exchange Service (DAAD).

[1]National Institute of Informatics (NII), Tokyo, Japan cnlimberg@gmail.com

numerous iterations and variants [16], [17], [18], [19], [20]. New variants like YOLOX [19], [21] and FCOS [22] have found improvements by moving away from the concept of fixed anchor boxes [23], integrating new data augmentation techniques and optimizing for new training objectives.

Still, the basic training paradigm of those models is the same compared with traditional handcrafted approaches. There is a fixed task definition and the need to acquire a data set that captures this task definition very accurately. Then the model is trained on a large part of the data and evaluated on a smaller fraction [24], [25].

With the introduction of LLMs, there was also a shift in this training paradigm. OpenAI demonstrated this by autoregressively pre-training a large transformer-based model on a vast collection of internet text corpora [1], showcasing that the contained general knowledge can be effectively compressed. With a fine-tuning step the model can be tuned for a specific task. ChatGPT [26] and GPT-4 [2] are trained in this post-training step with supervised data and with reinforcement learning from human feedback to realize a conversational agent that can be prompted with a wide range of tasks.

Recently, OpenAI has released the vision-enabled variant GPT-4V [4], which shows promising capabilities on real-world visual understanding [3]. However, GPT-4V still underperforms in detecting humans within a drone context where humans are usually captured in a few pixels and recorded from steep angles. We could not replicate the person detection results shown in [3] (pp. 42, Figure 27) for those difficult conditions.

To compensate for these shortcomings we decided to explore whether YOLO-World [7] can be utilized for the detection task within this challenging application. YOLO-World is a very new approach within the YOLO family, published only two and a half weeks from the writing of this manuscript. While the approach incorporates traditional YOLO elements like the YOLO backbone and bounding box heads, the model employs a CLIP [27] text embedding and uses an intermediate vision-language network for fusing both modalities. These advancements allow prompting the model with new classes at inference time, letting it detect objects it was not explicitly trained to detect.

In the realm of aerial view person detection, several methods such as Mobilenetv2 [28], Faster R-CNN [29], SSPNet [30], and YOLOv3 [31] have been employed. These models are typically fine-tuned with bespoke datasets captured using drones and manually annotated. However, the significant effort required for recording and annotating such datasets prompts us to explore alternative approaches.

III. EVALUATION

A. DATASET

For our evaluation, we consider the Okutama-Action dataset [32] because it presents a challenging scenario and covers both our areas of interest: person detection and action classification. In the Okutama-Action dataset, multiple persons are captured from aerial views using two drones at various times of the day. The dataset comprises different flights, each averaging about one minute in duration. Each camera frame depicts up to 10 individuals engaged in 12 distinct actions. These actions include running, walking, lying, sitting, standing, reading, drinking, pushing/pulling, carrying, calling, handshaking, and hugging.

Fig. 1. Sample image of the Okutama-Action dataset. The image shows eight persons, performing the actions calling, carrying, push/pulling, lying, reading, sitting, and standing.

While the dataset includes a total of 43 minutes of recordings, we're focusing solely on the test dataset, which consists of 10 flights totaling 9 minutes and 32 seconds. An example image from the dataset is illustrated in Figure 1.

B. PERSON DETECTION

a) GPT-4V Experiments: In preliminary experiments, we wanted to evaluate whether GPT-4V can describe a drone image accurately. We prompted it with the image from Figure 1 and the following text prompt:

- **Prompt:** "You are an expert in classifying persons and their actions in pixelated images recorded from drones. Describe this picture in a few sentences. How many people are there, where are they in the image and what are they doing?"
- **Response:** "In the image, there appear to be five individuals and various objects that could be tables or benches. The people are dispersed across the frame with no two individuals close together, suggesting they may be engaging in independent activities. One person is sitting on a bench, another is near a table, and the others are standing. They seem to be in a public space, possibly a park, given the open grassy area and the park-like furniture visible. It's difficult to ascertain their specific actions due to the image's perspective and resolution, but they do not appear to be interacting with each other and are rather involved in separate tasks or leisure activities."

As the response suggests, GPT-4V can grasp the overall scene depicted in the image and has a basic understanding of what the people are doing and what might be the background. However, it struggles with accurately counting the number of people and identifying their locations within the image. Additional preliminary experiments also suggest that GPT-4V lacks the ability to determine the boundaries of individuals

Fig. 2. YOLO-World detection of aerial image. We prompt the model with a single class 'Person' and loaded pre-trained weights. We uploaded a detection video of the full test dataset here: `https://www.youtube.com/watch?v=QntgkMKVuVQ`.

accurately. In most cases, the boundaries were significantly misaligned and individuals were not captured by them at all. We attempted various prompts, including requesting relative image coordinates for axis-aligned bounding boxes, center coordinates of individuals, and indexing an imaginary grid, as well as overlaying an actual grid onto the image. However, none of those methods resulted in a significant improvement in detection quality.

It is also worth mentioning that GPT-4V often refuses to give an answer because of safety policies [33]. However, we could circumvent this issue by "distracting" it to reply in a specific JSON format. We assume that then, it interpreted the request more like a programming task, rather than a "detect and classify humans" task. However, it is likely that these safety-related fine-tuning steps are also negatively influencing the accuracy of GPT-4V.

b) YOLO-World Experiments: To compensate for the person detection shortcomings of GPT-4V, we apply YOLO-World with pre-trained weights on the dataset. We prompt the model just with a single-word text-prompt 'Person' and load the published weights YOLO-Worldv2-L-1280. We applied a basic non-maximum suppression to the predicted bounding boxes. Other than that, the predicted bounding boxes were calculated separately for each frame and we haven't applied any filtering, filling, smoothing or any other post-processing on them.

As depicted in Figure 2, YOLO-World accurately detects the persons in the images. However, the confidence scores associated with the bounding boxes are relatively low, likely because the model was primarily trained with close-ups or group portraits. To address this, we filter the bounding boxes using a very low confidence threshold of 0.01. This approach results in some false positive detections, which can be subsequently filtered out by GPT-4V. We delve deeper into this issue in Section III-C.

Table I depicts the performance of YOLO-World. We calculate the precision, recall and F1 scores of the detections. We classify a detection to be true if the bounding box is overlapping with the ground-truth bounding box by at least 10%. This low percentage is justified because the ground-truth bounding boxes are not very accurate. Only one in every 10 frames was manually annotated, while the rest were interpolated, and their quality per flight varies drastically – we could confirm that flight 2.2.1 and 2.1.8 had a very poor annotation, which is especially reflected in the mean Intersection Over Union (mIOU) column of Table I. The accuracy of YOLO-World stays consistent across the flights, which is shown in the rendered video linked in the caption of Figure 2.

TABLE I

YOLO-WORLD ZERO-SHOT PERSON DETECTION PERFORMANCE.

flight	precision	recall	F1	mIOU
2.2.10	0.643	0.530	0.572	0.228
1.2.1	0.946	0.575	0.707	0.494
2.1.9	0.888	0.880	0.868	0.622
1.2.3	0.866	0.748	0.785	0.468
1.1.9	0.802	0.839	0.791	0.507
2.2.3	0.866	0.814	0.820	0.651
2.2.1	0.369	0.317	0.333	0.117
2.1.8	0.067	0.066	0.066	0.033
1.2.10	0.674	0.642	0.654	0.281
1.1.8	0.728	0.818	0.742	0.539

It is noticeable that, performance-wise, YOLO-World cannot compete with a traditionally trained YOLOv3 approach [31], especially regarding mean Intersection over Union (mIOU). However, the achieved performance may be sufficient for drone use cases where detecting a person is more important than the prediction of an accurate image location.

C. ACTION RECOGNITION

In this section, we explore whether GPT-4V can classify the actions of the persons detected in the previous subsection. Our experimental setup is as follows: we select 10 images from each flight of the test dataset, and we utilize YOLO-World for person detection. For each prediction, we calculate the Intersection over Union (IOU) with all the ground-truth bounding boxes and assign the label of the bounding box with the highest IOU. If there is no overlap, we label the predicted bounding box as 'no_person'.

We define two recognition problems. Firstly, we assess whether GPT-4V can effectively filter out region proposals that do not contain a person, which is essentially a binary classification problem of 'person' or 'no_person'. Secondly, we investigate whether GPT-4V can accurately determine the correct action class for each person, representing a 13-class classification problem (12 action classes and the 'no_person' token). For each problem, we calculate the 0/1 accuracy and the F1 score.

For each problem, we have conducted four experiments:

- **Experiment XXX**: We used the simplest prompt: "Is there a person in this image? If yes, what activity is he or she doing? Return one of the following answers and nothing else: [action classes and 'no_person' token]."
- **Experiment EXX**: We applied an 'expert priming' by prepending the sentence: "You are an expert in classifying persons and their actions in pixelated images recorded from drones." to the XXX prompt.
- **Experiment EDX**: We requested an explanation for the recognized class by appending the following to the EXX

prompt: “Return the activity and a short statement why you think the person is doing this activity. Return in this format and nothing else: [”activity”, ”statement”].”

- **Experiment EDS**: We incorporated the preceding and subsequent image in the time series to consider temporal information (each of them is 10 frames, or 0.3 seconds, apart of the image of interest). We hope that by adding the time dimension, actions like walking can be classified more accurately because the model would recognize the movement of the legs.

TABLE II

GPT-4V SCORES FOR PERSON/NON-PERSON RECOGNITION AND THE RECOGNITION OF 13 ACTION CLASSES.

exp	valid responses		person rec		action rec	
	valid/total	ratio	0/1	F1	0/1	F1
XXX	436/447	0.975	0.725	0.362	0.248	0.255
EXX	435/447	0.973	0.724	0.381	0.266	0.274
EDX	436/447	0.975	0.734	0.333	0.262	0.268
EDS	360/447	0.805	0.750	0.483	0.277	0.344

The results of the four experiments are listed in Table II. In general, performance improves slightly with expert priming (Experiment EXX). Demanding an explanation from the model does not seem to enhance the recognition performance (Experiment EDX). Querying the model with an image sequence (Experiment EDS) delivered the best results but for this experiment 20% of the queries were rejected by OpenAI due to their safety policies (see Table II, column ‘valid responses’).

However, the model can not predict action class accurately. This might be because of the inherently challenging nature of the dataset, where one person can perform multiple actions simultaneously, but only the most prominent action is labeled. For instance, a person could be carrying an item while walking. Predicting the class ‘walking’ would be considered incorrect because the prominent class would be ‘carrying’. We have depicted a confusion matrix to highlight this characteristic of the dataset in Figure 3. However, there are instances where unambiguous actions are also misclassified. For experiments EDX and EDS we verified the explanations the model returned for some examples. Sentences like “The person appears to have one foot in front of the other, suggesting movement typical of walking.” or “The person appears to have one hand raised to their ear in a manner that suggests they are holding a phone, which indicates they might be making a call.” indicates, that the model has a good understanding about the images and the prompted action classes. However, often the explanations were just wrong. We have illustrated the action recognition results for the previously discussed image in Figure 4.

IV. CONCLUSIONS

We evaluated two zero-shot LMMs for their recognition applicability within a drone context. Our results suggest that YOLO-World achieved good detection performance. Subsequently, we employed GPT-4V to classify the predicted region proposals with action classes. However, the model

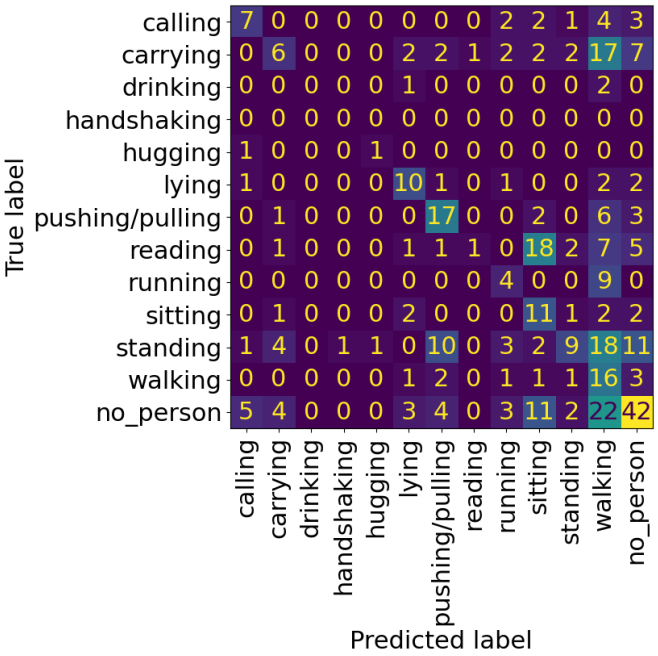

Fig. 3. Confusion matrix for action recognition using GPT-4V.

Fig. 4. GPT-4V classification of YOLO-World region proposals.

struggles to provide accurate zero-shot predictions for the 12 action classes. Nevertheless, it could potentially be utilized to filter out unwanted region proposals or to provide a general description of the scenery. While the accuracy may not yet be comparable to traditional approaches, there is a significant advantage in not having to train the models but simply prompting them. By changing just one word in the two prompts, a robot could be applicable for an entirely different use case, such as finding dogs or other objects of interest. This increased flexibility could potentially compensate for the current limitations in recognition quality.

In our future work we aim to assess whether the two models are applicable in a rescue drone use case. Our preliminary investigation has revealed that GPT-4V possesses a detailed basic understanding of drone images, but it struggles to accurately determine the location and actions of people. The latter problem could potentially be addressed by prompting the model with supervisory input in a few-shot learning manner, providing it with additional prior task knowledge.

REFERENCES

[1] T. B. Brown, B. Mann, N. Ryder, *et al.*, "Language models are few-shot learners," 2020.

[2] OpenAI, :, J. Achiam, S. Adler, S. Agarwal, *et al.*, "GPT-4 technical report," 2024.

[3] Z. Yang, L. Li, K. Lin, J. Wang, C.-C. Lin, Z. Liu, and L. Wang, "The dawn of LMMs: Preliminary explorations with GPT-4V(ision)," 2023.

[4] OpenAI, "GPT-4V(ision) technical work and authors," 2023. [Online]. Available: https://openai.com/contributions/gpt-4v

[5] J. Copet, F. Kreuk, I. Gat, T. Remez, D. Kant, G. Synnaeve, Y. Adi, and A. Défossez, "Simple and controllable music generation," in *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[6] A. Blattmann, R. Rombach, H. Ling, T. Dockhorn, S. W. Kim, S. Fidler, and K. Kreis, "Align your latents: High-resolution video synthesis with latent diffusion models," 2023.

[7] T. Cheng, L. Song, Y. Ge, W. Liu, X. Wang, and Y. Shan, "YOLO-world: Real-time open-vocabulary object detection," 2024.

[8] S. M. S. M. Daud, M. Y. P. M. Yusof, C. C. Heo, L. S. Khoo, M. K. C. Singh, M. S. Mahmood, and H. Nawawi, "Applications of drone in disaster management: A scoping review," *Science & Justice*, vol. 62, no. 1, pp. 30–42, 2022.

[9] H. Zhao, F. Pan, H. Ping, and Y. Zhou, "Agent as cerebrum, controller as cerebellum: Implementing an embodied lmm-based agent on drones," 2023.

[10] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1. IEEE, 2005, pp. 886–893.

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[12] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 580–587.

[13] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object detection via region-based fully convolutional networks," in *Advances in Neural Information Processing systems*, 2016, pp. 379–387.

[14] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *European Conference on Computer Vision*. Springer, 2016, pp. 21–37.

[15] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 779–788.

[16] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[17] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," 2020.

[18] G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics YOLO," Jan. 2023. [Online]. Available: https://github.com/ultralytics/ultralytics

[19] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "YOLOX: Exceeding YOLO series in 2021," *arXiv preprint arXiv:2107.08430*, 2021.

[20] C.-Y. Wang, I.-H. Yeh, and H.-Y. M. Liao, "YOLOv9: Learning what you want to learn using programmable gradient information," 2024.

[21] C. Limberg, A. Harter, A. Melnik, H. Ritter, and H. Prendinger, "Deep detection dreams: Enhancing visualization tools for single stage object detectors," *Springer Communications in Computer and Information Science*, 2024.

[22] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: fully convolutional one-stage object detection," *CoRR*, vol. abs/1904.01355, 2019. [Online]. Available: http://arxiv.org/abs/1904.01355

[23] C. Limberg, A. Melnik, A. Harter, and H. Ritter, "Yolo–you only look 10647 times," *arXiv preprint arXiv:2201.06159*, 2022.

[24] C. Limberg, H. Wersing, and H. Ritter, "Beyond cross-validation—accuracy estimation for incremental and active learning models," *Machine Learning and Knowledge Extraction*, vol. 2, no. 3, pp. 327–346, 2020.

[25] ——, "Accuracy estimation for an incrementally learning cooperative inventory assistant robot," in *International Conference on Neural Information Processing (ICONIP)*, Nov. 2020.

[26] P. P. Ray, "ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope," *Internet of Things and Cyber-Physical Systems*, vol. 3, pp. 121–154, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S266734522300024X

[27] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," 2021.

[28] R. Geraldes, A. Goncalves, T. Lai, M. Villerabel, W. Deng, A. Salta, K. Nakayama, Y. Matsuo, and H. Prendinger, "UAV-based situational awareness system using deep learning," *IEEE Access*, vol. 7, pp. 122 583–122 594, 2019.

[29] G. L. Hung, M. S. B. Sahimi, H. Samma, T. A. Almohamad, and B. Lahasan, "Faster R-CNN deep learning model for pedestrian detection from drone images," *SN Computer Science*, vol. 1, pp. 1–9, 2020.

[30] M. Hong, S. Li, Y. Yang, F. Zhu, Q. Zhao, and L. Lu, "Sspnet: Scale selection pyramid network for tiny person detection from uav images," *IEEE geoscience and remote sensing letters*, vol. 19, pp. 1–5, 2021.

[31] S. Speth, A. Gonçalves, B. Rigault, S. Suzuki, M. Bouazizi, Y. Matsuo, and H. Prendinger, "Deep learning with RGB and thermal images onboard a drone for monitoring operations," *Journal of Field Robotics*, vol. 39, no. 6, pp. 840–868, 2022.

[32] M. Barekatain, M. Martí, H.-F. Shih, S. Murray, K. Nakayama, Y. Matsuo, and H. Prendinger, "Okutama-Action: An aerial view video dataset for concurrent human action detection," in *Computer Vision and Pattern Recognition (CVPR) Workshops*, 06 2017, pp. 28–35.

[33] Y. Wu, X. Li, Y. Liu, P. Zhou, and L. Sun, "Jailbreaking GPT-4V via self-adversarial attacks with system prompts," 2024.