Monitoring Human Activities in Riverine Space using 4K Camera Images with YOLOv8 and LLaVA: A Case Study from Ichinoarate in the Asahi River

Shijun PAN¹, Keisuke YOSHIDA², Yuki YAMADA³, Takashi KOJIMA⁴

¹Member of JSCE, Doctoral Student, Graduate School of Environmental and Life Science, Okayama University

(Tsushima-naka 3-1-1, Kita-ku, Okayama 700-8530, Japan)

E-mail: p4b36znn@s.okayama-u.ac.jp (Corresponding Author)

²Member of JSCE, Associate Professor, Graduate School of Environmental and Life Science, Okayama University

³Member of JSCE, Master Student, Graduate School of Environmental and Life Science, Okayama University

⁴Senior Researcher, TOKEN C. E. E. Consultants Co., Ltd

A uniform nationwide survey on riverine space utilization has been conducted approximately every five years as part of the "Census of Rivers and Waterfront Areas" in Japan, for properly promoting river projects and river management. Considering significant effort required for human tasks, the survey is commonly carried out for seven days per year. Then, the present river situation is estimated roughly through the year, based on the limited survey results. Therefore, it is challenging to grasp the actual conditions on weekdays, holidays, and at different times of the day. Accordingly, it is difficult to examine the effect of individual river maintenance work quantitatively over years. For this study, the authors in this research tried to recognize human activities on the river bank automatically from 4K camera images taken near the Asahi River diversion weir in Okayama Prefecture, using the object detection model YOLO (i.e., You Look Only Once) with the large-scale multimodal model LLaVA (i.e., Large Language-and-Vision Assistant). Results showed that the combination of these models has the potential to collect information on not only the number and location of people but also various human activities, such as walking, running, and skateboarding.

Key Words: LMM, YOLOv8, Human Activity Estimation, River Monitoring, 4K Camera

1. INTRODUCTION

As part of the Census of Rivers and Waterfront Areas, two essential surveys, the Survey on the Number of River Space Users and the River Report Card, are conducted to comprehensively understand river space utilization^{1), 2)}. The user count survey, guided by the Census Manual for Rivers and Waterfront Areas, occurs seven times (i.e., crowded days selected by river management administrators for monitoring the usage of the river) per year to capture the distinct characteristics of river use. For understanding the long-term water-friendly environment-based situation in the riparian area, despite providing a holistic view of general river usage, the current methodology faces challenges in capturing weekday patterns due to the balance required between survey frequency, time, and cost.

The existing broad categorization of use patterns

necessitates a more detailed breakdown to precisely understand river space utilization for the future riparian-environmental improvement, maintenance and citizen-convenience and protection (e.g. dogwalking-, kids-safety-based monitoring,), posing a challenge to surveyors from previous human-eyebased monitoring to automatic computeral analysis.

In recent years, many studies have been conducted to detect people using AI technology. As an example, a study³) using the object detection model YOLO⁴) to understand the status of walking and waiting space usage near bus stops showed the possibility of quantitatively understanding items such as user attributes (gender and age), the number of bus riders, and the number of people using waiting benches. On the other hand, human activities such as "walking" and "sports (including cycling and running)," which are items in the survey on river space use, were not detected. In addition, no research has been conducted to obtain information on people by combining object



Fig.1 Study site



Fig.2 A image of the Ichinoarate at 8:24 AM on November 10, 2021 during normal water from 4K camera dataset.

detection models and large-scale multimodal systems.

In response, this study proposes a detection method leveraging YOLOv8⁴⁾ and LLaVA⁵⁾ to address the limitations. Focused on riparian area, the innovative approach aims to unravel citizen behaviors using rivers. By merging the strengths of YOLOv8 (i.e., person-based number counting) and LLaVA (i.e., person-based activities analysis), the study seeks to provide an efficient detection approach, transcending current limitations that only one model cannot overcome the tasks like, the person-based number counting, activity analysis area selection and offering an insight into river space usage behaviors at the same time. Considering the current practical issues that need the personnel-based monitoring for a long term, the combination of the mentioned models-based non-personnel-monitoring automatic analysis results is an available method.

2. STUDY SITE

Fig.1(a) highlights the position of the study site in the Okayama Prefecture. In Fig.1(b), the location and the camera-direction of the device have been depicted. Fig.1(c) showed the 4K camera from the drone view. Fig.1(b) performed that Ichinoarate within the bifurcation area of the Asahi River. In Fig.1(b), the role of Ichinoarate becomes apparent as a crucial overflow weir responsible for channeling flood flow from the Asahi River into the distributary Hyakken River. This pivotal function underscores the importance of Ichinoarate in managing water flow in the lower reach of the Asahi River. To facilitate a comprehensive understanding, Fig.1(b) also provides insight into the perspective of the 4K camera, as depicted in Fig.1(c). This camera's positioning becomes instrumental in capturing the river behavior. especially during flood events. A sample image of 4K camera during normal water was shown in Fig.2.

3. METHODS

(1) Specifications of video-taken device

In this research, data collection was derived from an 4K camera (Hikvision, DS-7600 Series) as performed in **Fig.1** with the resolution of 3840 × 2160 px and 20 fps (i.e., frames per second), these river monitoring cameras were previously installed to provide real-time information on river conditions during floods from 3^{rd} , Aug. to 13^{th} , Dec. 2021. During the non-flood seasons, the 4K cameras also continue working for the data collection.

(2) Outline of artificial intelligence models

In this research, there are two models that have been applied for the person-related number and activities analysis from the computer-vision- and the LMM- based perspectives, individually. The computer vision model, YOLOv8 (i.e., You Only Look Once version 8) is a state-of-the-art deep learning framework for object detection, segmentation, pose estimation and classification tasks, the YOLOv8 cannot yet identify the specific activities without additional training. The multimodal model, LLaVA (i.e., Large Language and

Table 1 Specifications of LLaVA and YOLOv8 applied in this research.

	LLaVA-v1.5-7b	YOLOv8x-seg
GPU memory (GB)	10.0	19.2
Image size for inference (px)	3840×2160 px	4000×2250 px
Confidence	-	0.1 (Group A) 0.5 (Group B & C)
Classes	-	Person, Bicycle, Dog, Backpack, Umbrella, Handbag
Inference time (s · frame-1)	4.0	$0.5 \sim 0.9$



Fig.3 Flow chart of investigating human activities by using YOLOv8 and LLaVA: a) 4K camera-based data collection, b) Analyzing the person-based numbers using YOLOv8 and LLaVA, c) Analyzing the person-based activities using LLaVA, d) Practical analysis of the person-based numbers and activities using LLaVA.

Vision Assistant) is a multi-modal model that combines vision and language for general-purpose visual and language understanding. The parameter setting is written in Table 1.

(3) Work Flow of investigating human activities

In Fig.3 (a), the data collection phase begins by deploying a 4K camera, strategically positioned to meticulously capture the targeted area of interest. Moving to the analysis stage, as depicted in Fig.3 (b) and (c), the YOLOv8 framework was for the object detection, effectively locating various persons within the recorded 4K footage. The integration of LLaVA

Table	2	36	cho	sen	case	s for	the	analy	/sis	deri	ved	from	mul	tiple
		sit	tuati	ons	(i.e.,	mon	ths,	week	s, w	ater	situa	ations	and	time
		pe	rioc	ls)										

Mont	Week-	Water	Morning	Noon	Afternoon		
	related	situation	(6:00~12:00) (12:00~15:00)(15:00~18:00)		
		Water	1	2	3		
	Weekday	effluence	_	2	5		
	Weekuay	Normal	4	5	6		
Δυσ		water	_	5			
Aug.		Water	7	8	0		
	Holiday*	effluence		0	/		
	Honday	Normal	10	11	12		
		water		11	12		
		Water	_	_	_		
	Wookday	effluence	_	-	-		
	weekuay	Normal	13	14	15		
Son		water		14	15		
Seb.		Water	_	_	-		
	Holiday	effluence	_	-	-		
		Normal	16	17	19		
		water		17	18		
	Weekday	Water					
		effluence	_	-			
		Normal	10	20	21		
Oct		water		20	21		
ou.	Holiday	Water					
		effluence	_	-	-		
		Normal	22	23	24		
		water	22	23	24		
		Water					
	Weekday	effluence	_	-	-		
		Normal	25	26	27		
Nov		water		20	27		
INOV.	Haliday	Water					
		effluence	_	-	-		
	nonuay	Normal	28	20	20		
		water	20	29	50		
		Water					
	Wookdow	effluence	_	-	-		
	weekday	Normal	21	32	22		
Dec.		water	51	32	33		
		Water					
	Halidar	effluence	-	-	-		
	попаау	Normal	24	25	26		
		water	54	33	30		
*. Wo	akand and	Vegetion					

further enhances the analytical depth, facilitating analysis that unveils an understanding of the detected activities. By synergizing the capabilities of YOLOv8 and LLaVA, capturing intricate details of human interactions with heightened accuracy. As the final step in Fig.3 (d), the results are meticulously compiled and processed, giving rise to comprehensive reports and visual representations that highlight identified human activities.

4. RESULTS AND DISCUSSION

To optimize computational costs and refine the analysis area, the authors strategically structured two distinct groups for testing purposes. Group-A specifically aimed to assess the detection capabilities in full-screen and individual cropped images, focusing on the varying locations concerning the 4K camera, namely closer-distance and longer-distance locations. This evaluation aimed to discern the effectiveness of detection in different spatial

contexts, providing insights for the accurate personbased numbers.

Alternatively, Group-B was dedicated to investigating the impact of the analysis area's size on confirming person-interacted activities. This group delved into understanding how adjusting the size of area influenced the outcomes. the analysis particularly in scenarios involving interactions among individuals. By categorizing the tests into these two groups, the authors addressed locationrelated considerations in Group-A and size-related considerations in Group-B, collectively contributing to a more efficient and cost-effective analysis process.

Images were selected by classifying cases in terms of weekday/holiday, morning/afternoon/night, and water-effluence/normal-water, covering most river conditions during the shooting period. As shown in **Table 2**, 36 cases were selected as Group-C. A single targeted image was chosen for each case, confirming the presence of people on the river bank. Only in August, there were water effluence situations.

(1) Group-A: Analysis on person in full-screen images.

In **Fig.4**, a comparative analysis of results obtained from LLaVA and YOLOv8 reveals that both methods area. However, LLaVA faces challenges in extracting any information from the full-screen image of the longer-distance area. In contrast, YOLOv8 exhibits the capability to detect three persons in this more distant region. As illustrated in **Fig.5**, a straightforward prompt was employed to verify LLaVA's ability to count person-based numbers on individual cropped images (i.e., upper-half and downside-half parts of the original image). Evidently, the longer-distance area also encounters limitations in extracting this information using LLaVA (i.e., misclassified 3 persons to 2 persons).

The findings from **Fig.4** were subsequently revealed that the closer-distance output effectively extracted person-related data using both YOLOv8 and LLaVA. Conversely, the longer-distance output exhibited a divergence in performance, with YOLOv8 yielding partially correct results and LLaVA misclassifying person-related numbers entirely.

(2) Group-B: Analysis on person-interacted activities.

Fig.6 upper performed a woman that is running, **Fig.6 middle** presents a depiction of two individuals, a woman and a child, engaged in walking or running along the river bank in the upper frame. Notably, when the cropping area is expanded, as demonstrated in the **Fig.6 downside**, an additional element is introduced: a woman with a leashed dog. This augmentation in the cropping area induces a discernible shift in the outcomes generated by LLaVA.

Drawing insights from the LLaVA-derived results, it becomes apparent that when analyzing a group of persons, particularly when interactions among these individuals come into work, careful consideration need to be given to the selection of the cropped area. The behaviors introduced by the inclusion of a leashed dog underscores the importance of judiciously defining the spatial boundaries of the analysis. As such, the implications of alterations in the cropping area on the interpretability of the results must be taken into account, ensuring a nuanced and comprehensive understanding of person-related activities within a given scene.

(3) Group-C: Application on the practical cases.

To avert misclassification of activities resulting from incomplete person coverage, a deliberate choice was made to employ a cropping process depicted in **Fig.7**. In consideration of the diverse periods encompassing study site, the authors judiciously selected 36 cases, meticulously detailed in **Table 2**, derived from various situations. In consonance with the findings from **Group-A**, the focus was placed on closer-distance areas.

This methodology ensures that the images encapsulate the entirety of interacting individuals, thereby enhancing the precision of LLaVA's analysis.

Subsequently, the requisite analysis from the images was methodically extracted, as illustrated in Fig.8, denoted by the delineated yellow areas (the Bounding-Box- and Mask-based results on the individual objects shown in Fig.9). In Fig.10, the authors applied the prompt in detecting the personbased number and activities. The outcomes of person-related activities and numbers across different as discerned through LLaVA, cases, were systematically compiled in Table 3. Half of the person-related activities and numbers were accurately detected across all 36 cases comparing with GT, that were counted by the author without twice counting. From the output derived from the Table 3, the prompt is not totally accurate for the person-based activities and numbers. In the future, the other prompts are in the need for the accurate detection.





Fig.4 Results of detecting the person-related activities were derived from full-screen using LLaVA, YOLOv8x-seg. Closer- and Longer- distance outputs were extracted by authors derived from Full-screen results, individually. (i.e., red- and yellow-line stand for the closer- and longer-distance result, separately)



Fig.5 Results of detecting the person-related numbers derived from closer- and longer- distance images using LLaVA, YOLOv8x-seg, individually. (i.e., red- and yellow-lines are same as in Fig.4)

/// Intelligence, Informatics and Infrastructure, Volume 5, Issue 1, 2024

Cropped Images	LLaVA output
	The image features a woman walking down a sidewalk, wearing a scarf and carrying a handbag. She appears to be in motion, possibly running or walking quickly. The sidewalk is located near a street, and there are some plants nearby. The woman is the main focus of the scene, and her actions suggest that she might be in a hurry or engaged in some activity. The image features a woman and a young boy walking down a sidewalk. The woman is walking ahead of the boy, and they appear to be enjoying their time together. The sidewalk is made of concrete, and there is a potted plant nearby, adding some greenery to the scene. The woman is carrying a handbag, which is visible as she walks.
	The image features a woman walking her dog on a leash, accompanied by a young boy. They are walking down a sidewalk, possibly near a river or a beach. The woman is holding the dog's leash, ensuring the dog. The scene also includes a few other people in the background, possibly enjoying the outdoors or engaging in their own activities.

Fig.6 Results of detecting the person-interacted activities derived from cropped images using LLaVA. It showed a problem of LLaVA, when the cropped area changed, the activities also change.



Fig.7 Process of how to choose the cropped images for the LLaVA analysis: 1. Cropping the down-half of the whole image area (i.e., Closer-distance area), 2. Using YOLOv8 to extract the necessary factors (i.e., person, bike, handbag, dog), 3. Collecting the detailed positions for cropping the images derived from the YOLOv8-based results (i.e., x-min, y-min, x-max, y-max), 4. Cropping the image based on the detailed positions.

// Intelligence, Informatics and Infrastructure, Volume 5, Issue 1, 2024



Fig.8 Table 2-based 36 chosen analysis area were selected using the yellow boxes derived from the process (i.e., W.E. and N.W. stand for water effluence and normal water, individually)



Fig.9 Results of B-Box (i.e., Bounding Box) and Mask-based YOLOv8 on multiple classes from 36 cases shown in Fig.8.





Fig.10 Results of detecting the person-interacted activities using prompt-assisted LLaVA.

Table 3 Person-related activities and numbers detected in this research using the 4K camera from 36 cases (i.e., green color stands for the correctly detected). GT stands for the meaning of ground truth that were judged by the authors about person-related activities and numbers.

Case Welling		D		Photo-		Bike-		Dog-			
Case	wan	king	Run	ning	taki	taking		ng	walking		
(4K)	GT	LLaVA	GT	LLaVA	GT	LLaVA	GT	LLaVA	GT	LLaVA	
1	2	2	0	0	0	0	0	0	0	0	
2	1	1	0	0	0	0	0	0	0	0	
3	2	2	0	0	0	0	0	0	0	0	
4	0	1	0	0	0	0	2	2	0	0	
5	0	1	1	1	0	1	0	0	0	0	
6	0	1	0	1	0	1	6	5	0	1	
7	3	1	0	0	0	0	0	0	1	1	
8	0	1	0	1	1	0	0	0	0	0	
9	8	2	0	1	0	1	0	0	1	0	
10	0	0	0	0	0	1	1	1	0	0	
11	0	1	0	0	0	0	1	1	0	0	
12	1	1	0	0	0	0	0	0	0	0	
13	1	1	0	0	0	0	0	0	0	0	
14	0	1	1	1	0	1	0	0	0	0	
15	1	1	0	0	0	0	0	0	0	0	
16	1	1	0	0	0	0	0	0	0	0	
17	3	1	0	0	0	0	1	1	0	0	
18	1	1	0	0	1	1	1	1	0	0	
19	1	1	0	0	0	0	0	0	0	0	
20	0	1	1	1	0	1	0	0	0	0	
21	2	2	0	0	0	0	0	0	0	0	
22	0	0	1	1	0	0	0	0	0	0	
23	0	0	0	0	0	0	1	1	0	0	
24	2	2	0	0	0	0	0	0	0	0	
25	1	1	0	0	0	0	0	0	0	0	
26	3	1	0	0	0	0	0	0	0	0	
27	0	1	0	0	0	0	0	0	1	1	
28	0	1	0	1	0	1	0	0	1	0	
29	0	1	0	0	0	0	1	1	0	0	
30	2	2	0	0	0	0	0	0	0	0	
31	1	1	0	0	0	0	0	0	0	0	
32	0	1	0	0	0	0	0	0	1	1	
33	2	2	0	0	0	0	0	0	0	0	
34	0	1	1	1	0	1	0	0	0	0	
35	0	1	1	1	0	1	0	0	0	1	
36	2	2	0	0	0	0	0	0	0	0	

Considering that CCTV cameras are more widely used in river monitoring in Japan than 4K cameras, the authors performed similar validations by resizing the video captured with a 4K camera from its original resolution ($3840 \times 2160 \text{ px}$) to a standard CCTV camera resolution ($1920 \times 1080 \text{ px}$) in **Table 4**. As a

result, no significant differences were found between original and resizing datasets in the YOLO and LLaVA analyses. This implies that the earlier results of this study may be applicable in cases using CCTV cameras installed in most Japanese rivers. //// Intelligence, Informatics and Infrastructure, Volume 5, Issue 1, 2024

 Table 4 Person-related activities and numbers detected in this research using images resized to CCTV resolution from 36 cases (i.e., green color stands for the correctly detected). GT stands for the meaning of ground truth that were judged by the authors about person-related activities and numbers.

Case Walkir		dina	Running		Photo-		Bike-		Dog-	
(CC		ung	Kunning		taking		riding		walking	
TV)	GT	LLaVA	GT	LLaVA	GT	LLaVA	GT	LLaVA	GT	LLaVA
1	2	2	0	0	0	0	0	0	0	0
2	1	1	0	0	0	1	0	0	0	0
3	2	2	0	0	0	0	0	0	0	0
4	0	1	0	1	0	1	2	2	0	0
5	0	1	1	1	0	1	0	0	0	0
6	0	1	0	1	0	1	6	4	0	1
7	3	1	0	0	0	0	0	0	1	1
8	0	1	0	1	1	0	0	0	0	0
9	8	2	0	1	0	1	0	0	1	0
10	0	0	0	0	0	0	1	1	0	0
11	0	1	0	0	0	0	1	1	0	0
12	1	1	0	0	0	0	0	0	0	0
13	1	1	0	0	0	0	0	0	0	0
14	0	1	1	0	0	0	0	0	0	0
15	1	1	0	0	0	0	0	0	0	0
16	1	1	0	0	0	0	0	0	0	0
17	3	2	0	0	0	0	1	1	0	1
18	1	1	0	0	1	1	1	1	0	0
19	1	1	0	0	0	0	0	0	0	0
20	0	1	1	0	0	1	0	0	0	0
21	2	2	0	0	0	0	0	0	0	0
22	0	1	1	1	0	0	0	0	0	0
23	0	1	0	0	0	0	1	1	0	0
24	2	2	0	0	0	0	0	0	0	1
25	1	1	0	0	0	0	0	0	0	0
26	3	3	0	0	0	0	0	0	0	0
27	0	1	0	0	0	0	0	0	1	1
28	0	1	0	0	0	0	0	0	1	1
29	0	0	0	0	0	0	1	1	0	0
30	2	2	0	0	0	0	0	0	0	0
31	1	1	0	0	0	0	0	0	0	0
32	0	1	0	0	0	0	0	0	1	1
33	2	2	0	0	0	0	0	0	0	0
34	0	1	1	0	0	0	0	0	0	0
35	0	1	1	1	0	0	0	0	0	0
36	2	2	0	0	0	0	0	0	0	0

5. CONCLUSION

The current study introduces an innovative methodology that integrates YOLOv8 and LLaVA for the daily observation of person-based activity in the riverine space. Leveraging the implementation of prompts, this approach successfully detected person-related activities and numbers in 18 out of the 36 cases examined. This collaborative fusion of YOLOv8 and LLaVA, augmented by the guiding influence of prompts, marks a significant stride towards enhancing the efficacy of understanding the long-term water-friendly environment-based situation in the riparian area.

ACKNOWLEDGMENT: The authors appreciate the Chugoku Regional Development Bureau, Ministry of Land, Infrastructure, Transport and Tourism for offering materials related to river management tasks and 4K video images along the Asahi River and Hyakken River.

REFERENCES

- 1) https://openjicareport.jica.go.jp/pdf/1000047169_01.pdf
- Tsuruta, M. and Nakamura, K.: Cultural ecosystem service characteristics of rivers ascertained from national census on river environments. Advances in river engineering, Vol. 27, pp.323-328, 2021.
- 3) Hironori, K., Masashi Y., Yusuke N., Toshiki H., Akihito O., Aya U.: Research on analysis of public space usage using object detection technology. Artificial Intelligence and Data Science, Vol. 4, Issue 3, pp.83-89, 2023.
- 4) https://github.com/ultralytics/ultralytics
- 5) https://github.com/haotian-liu/LLaVA

(Received January 15, 2024) (Accepted February 29, 2024)