Framework for Motorcycle Risk Assessment Using Onboard Panoramic Camera

Natchapon Jongwiriyanurak[1] ✉
Department of Civil, Environmental and Geomatic Engineering, University College London, UK

Zichao Zeng ✉
Department of Civil, Environmental and Geomatic Engineering, University College London, UK

Meihui Wang ✉
Department of Civil, Environmental and Geomatic Engineering, University College London, UK

James Haworth ✉
Department of Civil, Environmental and Geomatic Engineering, University College London, UK

Garavig Tanaksaranond ✉
Department of Survey Engineering, Faculty of Engineering, Chulalongkorn University, Bangkok, Thailand

Jan Boehm ✉
Department of Civil, Environmental and Geomatic Engineering, University College London, UK

Abstract

Traditional safety analysis methods based on historical crash data and simulation models have limitations in capturing real-world driving scenarios. In this experiment, panoramic videos recorded from a motorcyclist's helmet in Bangkok, Thailand, were narrated using an image-to-text model and then put into a Large Language Model (LLM) to identify potential hazards and assess crash risks. The framework can assess static and moving objects with the potential for early warning and incident analysis. However, the limitations of the existing image-to-text model cause its inability to handle panoramic images effectively.

2012 ACM Subject Classification Information systems → Geographic information systems; Computing methodologies → Scene understanding

Keywords and phrases Traffic incident risk, Large Language Model, Vision-Language Model

Digital Object Identifier 10.4230/LIPIcs.GIScience.2023.44

Category Short Paper

Funding This research was supported by UCL Global Engagement and Faculty of Engineering, Chulalongkorn University Fund.

1 Background

Traffic incidents are a global issue that causes significant economic and social costs. Every year, millions of people die or are injured in road crashes worldwide, costing countries 3% of their Gross Domestic Product (GDP) on average, with most incidents happening in low- and middle-income countries [13]. According to the World Health Organisation (WHO), Thailand has one of the world's highest road traffic fatality rates, with an average of 22,000 deaths annually. Bangkok, the capital city, is a hotspot for traffic incidents, with almost 1 million casualties a year in 2020 and 2021, 90% of whom were motorcyclists. Tracking the cause of these incidents is challenging as they can often be attributed to multiple factors.

[1] corresponding author

12th International Conference on Geographic Information Science (GIScience 2023).
Editors: Roger Beecham, Jed A. Long, Dianna Smith, Qunshan Zhao, and Sarah Wise; Article No. 44; pp. 44:1–44:7
Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Safety analysis approaches for local roads in Bangkok face limitations due to incompleteness, unavailability, under-reporting, and a lack of comprehensive crash-related factors and behavioural information [3, 11]. Studies utilise advanced cameras, benefiting from improved camera quality, computational power, and AI integration for street scene analysis. However, reliance on static cameras may limit their ability to capture the complexities of real-world scenarios [4, 5]. Recent progress in visual understanding, particularly in Vision-Language (VL) models and Large Language Models (LLMs), has shown great potential in analysing image-text pairs. This opens up opportunities to leverage pre-training VL and LLMs for evaluating real-time videos and assessing the scenes and behaviours of motorcyclists, thereby enhancing traffic risk assessment [8, 2, 12].

This study presents a framework for investigating the hazardous environment and interactions involving motorcycle riders and their surroundings. Initially, panoramic videos will be recorded in Bangkok, Thailand, using a GoPro Max camera mounted on the rider's helmet. Subsequently, Image-to-Text, video captioning, and LLM will be integrated to extract valuable information to identify potential hazards.

This paper is organised as follows. Section 2 lists the related works before Section 3 elaborates on the methodology and experiment. The preliminary results are described in Section 4. The paper finishes with the conclusion and potential applications of this study.

2 Related works

2.1 Traditional Traffic Risk Assessment

Traditionally, safety analysis has relied on historical crash data [11], which unfortunately may suffer from limitations such as incompleteness, unavailability, under-reporting, and a lack of comprehensive behavioural information, as well as the omission of important crash-related factors [3]. Simultaneously, simulation methods may not accurately represent non-lane-based mixed traffic conditions [1]. Although recent advancements in camera quality, Artificial Intelligence (AI), and computational power have enabled the development of simulation models for driving behaviour at intersections, most studies still heavily rely on static cameras, which may not fully capture the intricate complexities of real-world driving scenarios [4, 5].

2.2 Large Vision and Language Pre-trained Models in Traffic Scenes

Since the introduction of Contrastive Language-Image Pre-training (CLIP), VL Pre-training (VLP) models have rapidly advanced, relying on large text-image datasets [6, 9, 7]. Large VLP models achieve competitive performance on benchmark datasets, even without specific training, through zero-shot learning [9, 6]. Additionally, they have the capability for zero-shot Visual Question Answering (VQA) in the context of traffic image understanding [14]. However, while these large VLP models can efficiently extract textual information from image features, they face challenges when it comes to correlating relevant textual information and performing deeper interpretation, particularly in complex scenes involving multiple objects.

After OpenAI proposed ChatGPT [14], the use of LLMs expanded to specific tasks, excelling in summarising prompts and completing questions, explanations, and captions. However, LLMs lack the ability to extract visual features. By combining LLMs with VLP models, they can effectively interpret text features from images and gain detailed information through VQA. This combination overcomes the limitations of large VLP models in explaining phenomena and the inability of LLMs to extract image features without additional training. It is like a visually impaired person relying on an interpreter for specific tasks to understand

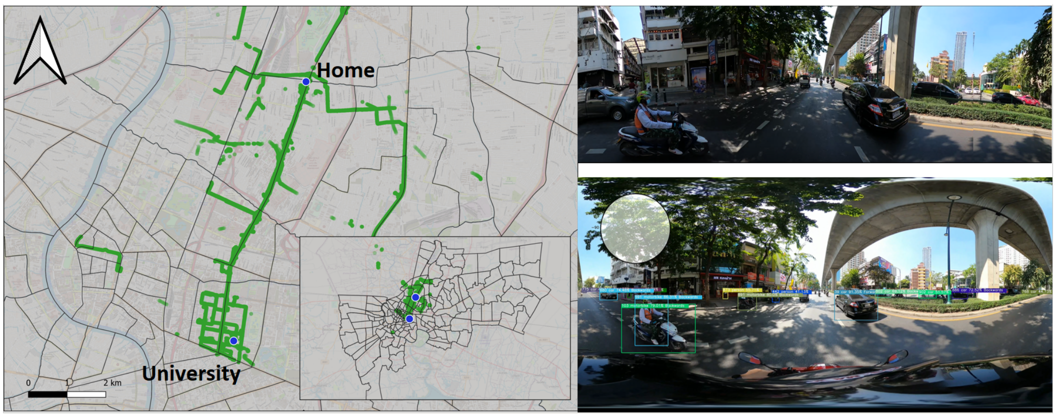

Figure 1 Paronamic video dataset coverage in Bangkok, Thailand (left) and an example of object detection (right).

their surroundings. While Large Language-and-Vision Assistant (LLaVA) has shown good interpretation of scenes in many scenarios [7], it falls short in object tracking in videos. To address this, a novel framework proposed in this study incorporates object detection and instance segmentation at each keyframe before applying VLP models. This framework aims to interpret VLP models for analysing traffic scenes, with potential applications in real-time traffic interpretation for early warning of potential risks and incident analysis by stakeholders and planners.

3 Methodology and Experiment

3.1 Data collection

This study involved collecting panoramic videos using a GoPro Max camera mounted on a helmet while riding a motorcycle from December 18, 2022, to January 16, 2023. The camera was set to 360 video mode with 5.6k resolution and 30 Frames per second. The journeys mostly covered the route from home in the Phaya Thai district to Chulalongkorn University in central Bangkok, as indicated by the green dots on the map with examples of captured scenes in Figure 1 (left) with an example of object detection (right). It is important to note that these journeys were routine activities and did not put the user at increased risk. The study obtained ethics approval from UCL and Chulalongkorn University. The dataset is the 360-view of street scenes from a motorcyclist's helmet across Bangkok's streets during diverse times of day, including peak and off-peak periods on weekdays and weekends.

3.2 Framework

This section introduces the proposed framework for identifying motorcycle crash risks from panoramic videos outlined in section 3.1. The framework overview is shown in figure 2 and considers two types of objects: environment or static objects and interactions or moving objects by using different VL models.

The environments or static objects are described using an image-to-text model, which provides information on non-moving objects, such as the number of lanes, flow density, road surface conditions, weather, and lighting.

The framework tracks and filters the interactions between the rider and surrounding objects. It excludes smaller boxes far from the rider to reduce computational costs and contribute less to risk. This exclusion is done to reduce computational costs. Each object is

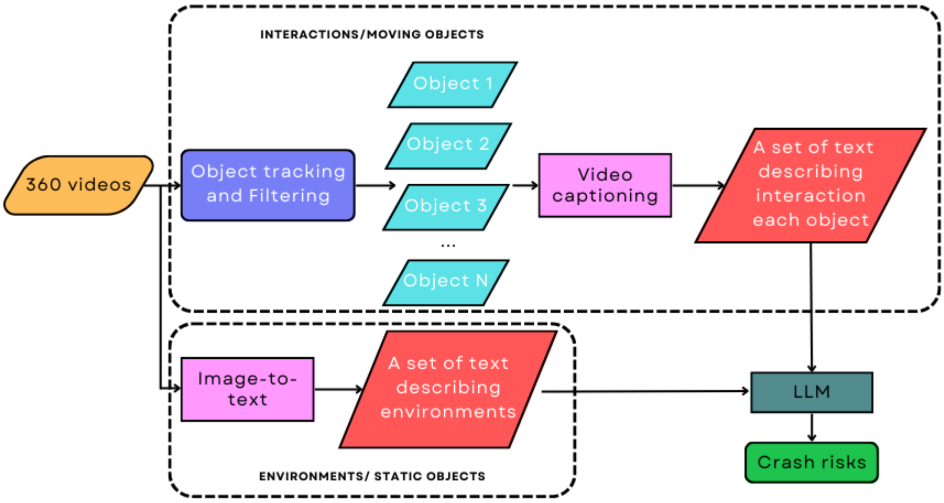
INTERACTIONS/MOVING OBJECTS
360 videos
Object tracking and Filtering
Object 1
Object 2
Object 3
...
Object N
Video captioning
A set of text describing interaction each object
Image-to-text
A set of text describing environments
ENVIRONMENTS/ STATIC OBJECTS
LLM
Crash risks

Figure 2 Framework for panoramic video crash risk analysing.

considered separately within the model, with other objects blurred. This approach focuses on capturing the interaction between the object itself and the rider. A video captioning model is employed to generate descriptions for each moving object. This model processes specific seconds of video footage and generates a set of sentences that describe the interactions between the surrounding objects and the motorcyclist.

The descriptions generated from both moving and static objects are input into a LLM to assess the potential crash risks. The existing LLM will be implemented and fine-tuned to rate a quantitative score that quantifies the level of risk. The visual risk score obtained from the LLM is then combined with trajectory information derived from the camera's GPS, which integrates with Geographic Information System (GIS) data. This GIS data may include historical incident records and Points-of-Interest. Through this integration, the final risk score is computed. This framework can potentially alert the rider when the risk score surpasses a predefined threshold. Such a system could be a valuable tool for motorcyclists, particularly when the camera is equipped with a GPU for real-time processing capabilities.

The ongoing study aims to implement image-to-text, video captioning and LLMs for risk rating purposes. As part of this framework, the image-to-text approach was tested using the LLaVA model [7]. The LLaVA model prompts 6 questions to gather information on various factors that are considered critical in assessing motorcycle crash risks. These factors include flow density, number of lanes, weather conditions, traffic signs (specifically speed limits), road surface conditions, and lighting conditions. By incorporating these crucial risk factors, the study seeks to enhance the accuracy and effectiveness of the risk rating process by validating 16 and 8 images during day and night time, respectively. Researchers manually supervise by rating 1 as correct, 0.5 as partially correct and 0 as wrong, then will calculate the accuracy of returning captions.

4 Preliminary results

The preliminary results are presented in table 1, revealing the accuracy from 6 prompts against manual supervision. The model performed well in classifying flow density, weather, and lighting. However, it showed relatively poor performance in identifying the number of lanes, road surface conditions, and traffic signs. Additionally, the model showed slight differences in performance between day and night time.

Table 1 Accuracy (in %) of captions tested against manual supervision.

Class	% (all)	% (day)	% (night)
Number of lanes	39.6	40.6	37.5
Flow density	81.3	75.0	93.8
Road surface	25.0	37.5	0.0
Traffic sign	16.7	15.6	18.8
Weather	77.1	75.0	81.3
Lighting	83.3	75.0	100.0

Lighting conditions are the easiest to identify, as shown in table 1. This is attributed to the straightforward evaluation of lighting based on red, green, and blue (RGB) values during visual decoding. The model also demonstrated high accuracy in identifying flow density. This is because VLP models have learned vehicles from the abundance of traffic images for a large dataset, and LLM can also easily understand traffic congestion at a textual level. The model effectively determined the weather conditions due to the substantial coverage of sky or weather in traffic images, as the model was pre-trained in a large weather-related sample size.

However, The accuracy of counting the number of lanes is relatively poor in panoramic images, which introduce horizontal misalignment between lane lines and vehicles, leading to confusion for the pre-trained model. Previous work has shown that image understanding tasks trained mainly on rectilinear images benefit from re-projecting equirectangular images to rectilinear before the visual task is performed [10]. Identifying road surface conditions poses a challenge due to the diverse range of colours, conditions, and materials found in different countries. This difficulty is exemplified by the misclassification of cement surfaces as wet surfaces in this study. The most challenging class to identify is traffic signs. It is worth noting that traffic signals (red/green/yellow lights) were considered traffic signs in the textual decoding, and tail lights from vehicles are often labelled as traffic signals, further contributing to the difficulty in accurately identifying traffic signs.

5 Conclusion and Future work

In this study, we proposed a framework to examine motorcycle incident risk by using VLP and LLM models from a panoramic video dataset. The video data was collected in Bangkok, Thailand, by mounting a 360 camera on a motorcyclist's helmet to record the interaction between the surroundings and the rider. A VLP model, LLaVA, is tested on a series of panoramic images in the daytime and nighttime. Promptings related to traffic incident risks are used. The results show the potential of using the pre-trained model to describe safety related features, from testing flow density, weather and lighting conditions, and images for prompting the LLM to rate the incident risk. On the other hand, the results reveal the limitations of using panoramic images when counting the number of lanes, road surfaces, and traffic signs.

In future developments, the framework will incorporate distortion correction to mitigate potential misinterpretations caused by distorted geometries. The objective is to describe critical risks associated with stationary objects and environments accurately. Moreover, there will be a strong emphasis on understanding traffic scenes within the framework model, achieved through the fine-tuning and training of pre-training VL and LLM for visual traffic comprehension and textual analysis. The risk analysis will transition from image-to-text to video captioning, integrating the detection and tracking of moving objects. Unrelated objects will be disregarded or given lower weights using depth estimation techniques to enhance accuracy. The overarching goal of this comprehensive framework is to comprehend crash risks for motorcyclists and provide real-time notifications to the rider when equipped with graphics processing units on the panoramic camera or edge device. While the framework holds the potential for transferability to other cities, careful consideration must be given to factors such as the environment, vehicles, behaviours, and contextual risks.

References

1 Gowri Asaithambi, Venkatesan Kanagaraj, and Tomer Toledo. Driving Behaviors: Models and Challenges for Non-Lane Based Mixed Traffic. *Transportation in Developing Economies*, 2(2):19, October 2016. doi:10.1007/s40890-016-0025-6.

2 Jun Chen, Deyao Zhu, Kilichbek Haydarov, Xiang Li, and Mohamed Elhoseiny. Video ChatCaptioner: Towards Enriched Spatiotemporal Descriptions, April 2023. arXiv:2304.04227 [cs]. URL: http://arxiv.org/abs/2304.04227.

3 Rupam Deb and Alan Wee-chung Liew. Missing Value Imputation for the Analysis of Incomplete Traffic Accident Data. In Xizhao Wang, Witold Pedrycz, Patrick Chan, and Qiang He, editors, *Machine Learning and Cybernetics*, volume 481, pages 275–286. Springer Berlin Heidelberg, Berlin, Heidelberg, 2014. Series Title: Communications in Computer and Information Science. doi:10.1007/978-3-662-45652-1_28.

4 Nopadon Kronprasert, Chomphunut Sutheerakul, Thaned Satiennam, and Paramet Luathep. Intersection Safety Assessment Using Video-Based Traffic Conflict Analysis: The Case Study of Thailand. *Sustainability*, 13(22):12722, November 2021. doi:10.3390/su132212722.

5 Gabriel Lanzaro, Tarek Sayed, and Rushdi Alsaleh. Can motorcyclist behavior in traffic conflicts be modeled? A deep reinforcement learning approach for motorcycle-pedestrian interactions. *Transportmetrica B: Transport Dynamics*, 10(1):396–420, December 2022. doi:10.1080/21680566.2021.2004954.

6 Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models, May 2023. arXiv:2301.12597 [cs]. URL: http://arxiv.org/abs/2301.12597.

7 Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual Instruction Tuning, April 2023. arXiv:2304.08485 [cs]. URL: http://arxiv.org/abs/2304.08485.

8 Jesus Perez-Martin, Benjamin Bustos, Silvio Jamil F. Guimarães, Ivan Sipiran, Jorge Pérez, and Grethel Coello Said. A Comprehensive Review of the Video-to-Text Problem, November 2021. arXiv:2103.14785 [cs]. URL: http://arxiv.org/abs/2103.14785.

9 Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision, February 2021. arXiv:2103.00020 [cs]. URL: http://arxiv.org/abs/2103.00020.

10 E. Sanchez Castillo, D. Griffiths, and J. Boehm. SEMANTIC SEGMENTATION OF TERRESTRIAL LIDAR DATA USING CO-REGISTERED RGB DATA. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLIII-B2-2021:223–229, June 2021. doi:10.5194/isprs-archives-XLIII-B2-2021-223-2021.

11 Chamroeun Se, Thanapong Champahom, Sajjakaj Jomnonkwao, and Vatanavongs Ratanavaraha. Motorcyclist injury severity analysis: a comparison of Artificial Neural Networks and random parameter model with heterogeneity in means and variances. *International Journal of Injury Control and Safety Promotion*, pages 1–16, June 2022. doi:10.1080/17457300.2022.2081985.

12 Junke Wang, Dongdong Chen, Chong Luo, Xiyang Dai, Lu Yuan, Zuxuan Wu, and Yu-Gang Jiang. ChatVideo: A Tracklet-centric Multimodal and Versatile Video Understanding System, April 2023. arXiv:2304.14407 [cs]. URL: http://arxiv.org/abs/2304.14407.

13 WHO. Global status report on road safety 2018. Technical Report 2, WHO, 2018. ISBN: 9789290496977 ISSN: 00142972 Publication Title: World Health Organization Volume: 3. doi:10.18041/2382-3240/saber.2010v5n1.2536.

14 Ou Zheng. ChatGPT Is on the Horizon: Could a Large Language Model Be All We Need for Intelligent Transportation? *Computation and Language*, March 2023. doi:10.48550/arXiv.2303.05382.