Text360Nav: 360-Degree Image Captioning Dataset for Urban Pedestrians Navigation

Chieko Nishimura, Shuhei Kurita, Yohei Seki
University of Tsukuba, RIKEN AIP, University of Tsukuba
1-2 Kasuga, Ibaraki, Japan
s2221611@s.tsukuba.ac.jp, shuhei.kurita@riken.jp, yohei@slis.tsukuba.ac.jp

Abstract

Text feedback from urban scenes is a crucial tool for pedestrians to understand surroundings, obstacles, and safe pathways. However, existing image captioning datasets often concentrate on the overall image description and lack detailed scene descriptions, overlooking features for pedestrians walking on urban streets. We developed a new dataset to assist pedestrians in urban scenes using 360-degree camera images. Through our dataset of Text360Nav, we aim to provide textual feedback from machinery visual perception such as 360-degree cameras to visually impaired individuals and distracted pedestrians navigating urban streets, including those engrossed in their smartphones while walking. In experiments, we combined our dataset with multimodal generative models and observed that models trained with our dataset can generate textual descriptions focusing on street objects and obstacles that are meaningful in urban scenes in both quantitative and qualitative analyses, thus supporting the effectiveness of our dataset for urban pedestrian navigation.

Keywords: Vision and Language, Natural Language Generation, Multimedia Document Processing

1. Introduction

Text feedback technology for city walking assistance is essential for multiple reasons. Individuals with visual impairments or those using smartphones while walking face visual information constraints. In such cases, text feedback improves understanding of the surroundings and informs pedestrians about safe paths and obstacles, supporting safe walking and instilling confidence in pedestrians' actions. Currently, datasets providing detailed text feedback focused on pedestrian navigation are lacking. By contrast, general domain datasets, such as MS-COCO (Lin et al., 2014), provide broad image descriptions, often lacking specific object details essential for individuals with visual impairments. Creating datasets for pedestrian navigation requires focusing on specific objects deemed useful for individuals with visual impairments, such as fences, tactile blocks, sidewalks, significant vehicles, bus stops, and elevator buttons. This could potentially enhance the safety and effectiveness of city walking, particularly for pedestrians, including those with visual impairments. In this study, we generated meaningful captions, focusing on obstacles on the streets, and created a dataset named Text360Nav to assist pedestrians in navigating urban environments. We conduct experiments on our dataset with image captioning models of BLIP (Li et al., 2022) and BLIP-2 (Li et al., 2023) and confirm that the models trained with TextNav360 concentrate on the salient objects in scenes that can be helpful for pedestrian navigation.

The contributions of our studies are as follows: (1) generating detailed object-specific captions beyond conventional datasets by utilizing the proposed dataset and (2) indicating the direction of technology to inform and assist often-overlooked objects during city walking.

2. Related work

2.1. Indoor datasets

Several studies have used indoor datasets focusing on object detection and region detection centered around 360-degree images (Zhao et al., 2020; Chou et al., 2020; Cao et al., 2022; Fu et al., 2019). Based on these studies, researchers have further explored indoor mobility (Li and Bansal, 2023; Wang et al., 2022; Cirik et al., 2020; Chen et al., 2019). Wang et al. (2022) proposed an optimal path from the current location to the destination based on the indoor information.Chen et al. (2019) and Cirik et al. (2020) aimed to reach the destination using instructions described in natural language and images within the current field of view. Kayukawa et al. (2023) created navigation movies for arbitrary destinations by synthesizing 360-degree videos without using object detection. Therefore, a wealth of research has focused on indoor walking and navigation. However, human activities are not limited to indoor environments. Additionally, unlike indoor situations, outdoor situations involve real-time changes such as traffic signals and vehicles, making straightforward applications unfeasible. In this study, we aimed to create an outdoor dataset by focusing on the elements that require attention during walking by considering such dynamic changes.

LREC-COLING 2024, pages 15783–15788
20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

Split	#Videos	Hours	#Annotations
Train	915	7.63	7,744
Val.	114	0.95	212
Test	111	0.92	249
All	1,140	9.50	8,205

Table 1: Statistics of the dataset: number of videos, total hours, and number of annotations.

2.2. Outdoor datasets

360-degree videos offer abundant data for autonomous driving and self-driving robots. Liao et al. (2022) collected videos taken from cars and annotations related to autonomous driving. Martin-Martin et al. (2021) collected 360-degree images by using robots to create a dataset for on-road object recognition. Data collection via robots is quite common. Although a dataset for on-road walking was reported by Xiao et al. (2012), it is currently not available. In this study, we collected data while walking in two countries: Japan and the United States. Furthermore, as a distinguishing factor from the 360-degree image dataset (Chou et al., 2018), we incorporated questions regarding what the researchers wanted to convey specifically to visually impaired individuals. We acquired captions for elements considered crucial during walking. Therefore, street objects (people, cars, etc.) were preferentially mentioned. Handrails and similar features were also referenced, providing unique values in the context of 360-degree video data obtained during walking.

3. Dataset

This section describes the methods for creating Text360Nav dataset and the related statistics.

3.1. Video collection and preprocessing

We collected equirectangular images of urban scenery using omnidirectional cameras of Ricoh Theta. We shot the videos in multiple locations in the Kanto region, Japan, and New York City, United States, resulting approximately five hours of footage. These videos, however, sometimes includes privacy-sensitive urban scenary, such as portraits of other pedestrians or number plates of vehicles. We apply two stage privacy processing for carefully concealing them. We first apply face blur[1] for disabling face identification. Then we apply the Detic (Zhou et al., 2022) for predicted regions of pedestrian portraits and license plates and we carefully blur such regions. After privacy processing, we segmented the original videos into short 30 seconds sample videos and extracted 30 images from them at one frame per second (1 FPS) that are used for image-wise evaluation based on image captions. Subsequently, during image annotation, to facilitate reference selection, we used a pretrained object detection model to display the bounding boxes. We applied Segment Anything (Kirillov et al., 2023) to generate candidate object boxes that can detect a broader range of classes than conventional models such as Detic, including Braille blocks. Furthermore, we divided the videos into training, validation, and testing sets at a 8 to 1 to 1 ratio. In video splitting, we ensure that the 30 seconds-length short videos that are extracted from the same original source video always appear in the same split of the dataset.

Table 1 presents the statistics of the dataset. For the training set, we densely annotated captions for captured frames with multiple workers. Similar to the data augmentation approach of the training data, we densely annotated the training data because we considered that a lot of image and caption pairs is required for the model training. For validation and test sets, however, it is not plausible to generate new different captions in a few frames. Therefore, we decided to annotate their captions per 30 second videos.

3.2. Annotation procedure

The annotation was performed using Amazon Mechanical Turk as follows (Figure 1). First, we displayed the previously extracted 30 images in the 30 seconds videos on MTurk and asked the workers to select three images from 30 images for the annotation. This selection aimed to choose distinctive images, particularly in scenes lacking distinct objects, instead of forcing annotations. The selected images were annotated by the workers. The images contained the bounding boxes created in the preprocessing step, and the workers described the objects selected from these bounding boxes. Here, the workers could manually adjust the bounding boxes to fit them to the objects being described. In our task, we concentrated on specific object-related descriptions for visually impaired persons, rather than describing the entire scene. To encourage workers to create object-specific descriptions, we instructed them to find objects based on the theme "What you would want to convey to a visually impaired person?" This method enabled them to write object-specific annotations for visually impaired persons instead of descriptions of the entire scene. We provided sample sentences related to crosswalks and cars for references. Furthermore, the workers were instructed to avoid mentioning distant objects like building windows or pedestrians clothing, which were less relevant for visually impaired persons when they walked.

[1] https://github.com/ORB-HD/deface

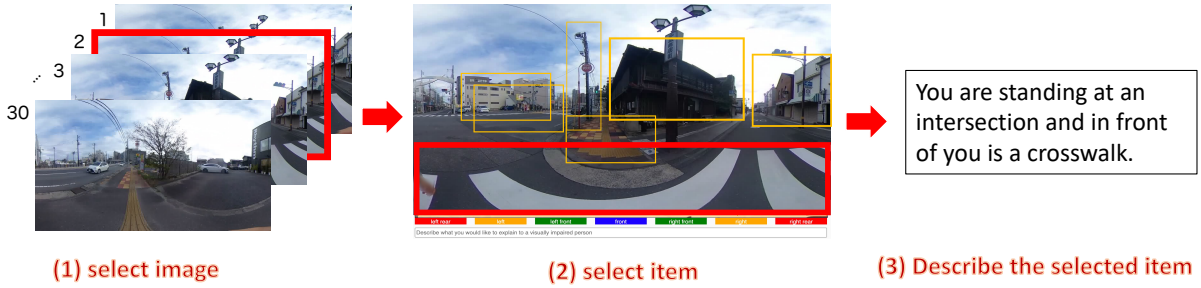

Figure 1: MTurk annotation flow

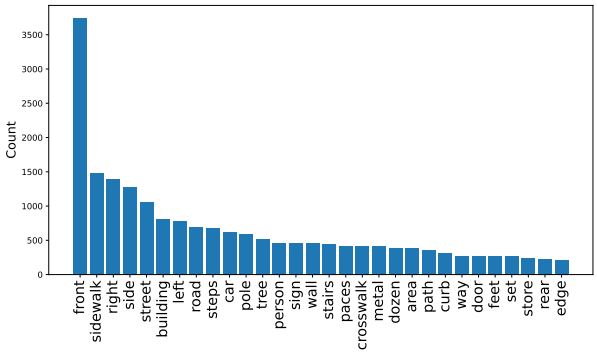

Figure 2: Frequent words in captions

3.3. Quality control

Crowdsourcing services may involve workers with inadequate skills. Therefore, we reviewed the annotation results of each worker. We configured the system to prevent assigning tasks to the workers who did not complete the tasks or showed evident grammatical errors in their responses. Additionally, we consider that some of the annotation writings are not necessarily suitable for visually impaired persons. For example, Hoogsteen et al. (2022) specified that there are mismatches of the attentions between visually impaired people and sighted people when sighted persons write up some descriptions for visually impaired persons. Therefore we manually checked the written descriptions and excluded those that were not in line with the images or suitable for visually impaired individuals.

3.4. Statistics

The statistics of the dataset are presented in Table 1. Out of the total, 1,140 videos collected, 923 videos (total 7.6 hours) were from New York City, USA, and 217 videos (total 1.8 hours) were from the urban areas of Tokyo and the Kanto region, Japan. We provide an overview of the word count distribution and statistics of the collected annotations. The average caption length was 14.72 words, with the minimum, maximum, and standard deviation being 7, 53, and 5.25 words, respectively. The content varied according to the caption length. Captions of 20 words or fewer described street objects and their directions. Captions containing 20–30 words included distinctive descriptions of the objects. Captions exceeding 30 words included predictions and warnings regarding what might occur as the subject progresses in addition to detailed object descriptions. Figure 2 presents an aggregation of commonly used words. It is evident that words denoting locations such as “left,” “front,” and “right” are frequently used. Moreover, words like “sidewalk” and “steps” describing street conditions are also frequent, indicating a strong understanding of both directions and the associated objects.

In the annotation process, 66 workers in worked for 8,205 writings, averaged 124.3 instances per worker. Different workers may annotate the same videos. As an open-ended description annotation, workers can choose different target objects in different frames to be described even when they are assigned to the same video. By doing so, we obtained diverse descriptions. Throughout the annotation process, we collected the annotation writings, with 7,744 writings for training, 249 for test, and 212 for validation.

4. Experiment

We evaluate the effectiveness of the created dataset through experiments using the following procedure.

4.1. Task

Our data consisted of captions and videos; however, worker annotations were performed on the images extracted from the videos. Given such images, we formulate our task as an image captioning task. In real applications, this corresponds to the images being automatically sampled from omnidirectional cameras and the models generating their textual descriptions.

4.2. Model

We apply state-of-the-art image captioning models like BLIP (Li et al., 2022) and BLIP-2 (Li et al.,

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE_L	METEOR	CIDEr	SPICE
Zero-shot								
BLIP (base)	15.22	5.28	1.95	0.71	16.37	6.17	10.96	5.33
BLIP (large)	10.03	3.46	1.37	0.58	15.39	4.80	8.24	5.52
BLIP-2 OPT-2.7B	12.02	3.48	1.21	0.49	14.33	4.82	8.82	4.34
BLIP-2 OPT-6.7B	10.08	2.46	0.80	0.36	13.02	4.24	7.67	4.30
Finetuned								
BLIP (base)	9.98	4.34	2.11	7.66	18.19	5.84	10.23	6.41
BLIP (large)	9.98	4.35	2.11	0.96	18.19	5.85	10.24	6.42
BLIP-2 OPT-2.7B	23.47	12.00	5.78	2.84	24.33	9.02	26.93	10.20
BLIP-2 OPT-6.7B	20.75	9.85	4.81	2.41	22.38	8.27	21.90	8.44

Table 2: Performance comparison with image captioning metrics on the test set of Text360Nav.

Annotated Caption:
A metal banister on the right of the staircase leading downstairs.

Zeroshot Generated Caption (OPT-2.7B):
a fisheye lens view of a hallway and stairs.

Finetune Generated Caption (OPT-2.7B):
a stairway leading down to the bottom of the building.

Annotated Caption:
There are bouquets of flowers on the wall of a booth near the right.

Zeroshot Generated Caption (OPT-2.7B):
a group of people walking around with umbrellas.

Finetune Generated Caption (OPT-2.7B):
a flower stand on the sidewalk in front of a person holding an umbrella.

Figure 3: Qualitative comparison of captions.

2023) to our task. They are based on the combination of a pretrained image recognition model and a large-scale language model, such as OPT (Zhang et al., 2022) with a Q-former. OPT has notable features such as high zero-shot transfer capability and language generation ability. It also achieves high performance with fewer trainable parameters compared to existing methods. In this study, we choose this model for image captioning experiments using a dataset created from scratch. We prepare two experimental settings: zero-shot and finetuned. In zero-shot, we evaluate the off-the-shelf performance of models trained using MS-COCO. In fine-tuned, we finetune these models with the Text360Nav dataset and report the performance. For finetuning, we followed the default hyperparameters of the those models. In BLIP, We used the global batch-size of 64. The initial learning rate is $2e-6$ and min learning rate is 0, weight decay is 0.05, and we use the linear warup cosine learning rate.

4.3. Quantitative analysis

Table 2 presents the experimental results for BLIP trained using the proposed dataset. We use automatic metrics evaluated using pycocotool [2]. We observe a relatively small performance difference among the zero-shot models compared to those of the finetuned models. We attribute this to the domain shift between the MS-COCO training images and Text360Nav urban street images. With fine-tuning, BLIP-2 with OPT models achieves the highest performance. Interestingly, BLIP-2 with OPT-2.7B performe better than with OPT-6.7B. It is notable that compared to BLIP, BLIP-2 often refers to the direction of objects such as “right,” “left,” “front” rather than the descriptions of objects. This may affect the simple BLEU-based metrics. To investigate the reasons for this, we examine several outputs from fine-tuning OPT-6.7B on a few images. The results reveal that the zero-shot model often makes references to the “fisheye lens view” as observed in Figure 3. Similar patterns are also observed in the finetuned model results. Therefore, we consider

[2]https://pypi.org/project/pycocotools/

that the model's performance does not improve during finetuning despite its large size. It is also plausible that the dataset's scale might not align with OPT-6.7B to overcome this pretraining prior. Consequently, in the next section, we compare and examine models trained with OPT-2.7B.

4.4. Qualitative analysis

In Figure 3, we present examples of descriptive captions generated by our model. These captions exhibit a tendency to emphasize objects within the urban environment, particularly focusing on "stairs," rather than providing an overall description of the entire image (see Figure 3 top). Notably, there is a strong emphasis on mentioning objects in the front. This can be attributed to the prevalence of the word "front" in our annotations, as indicated in Figure 2. Furthermore, at the bottom of Figure 3, worker annotations include references to in-store products, such as "bouquets of flowers on the wall all of a booth." Similarly, captions generated by our fine-tuned model also make references to the store, as evidenced by phrases like "flower stand." This suggests the potential to encourage mentions of urban scenery that would typically go unnoticed and promote serendipitous discoveries.
However, there are certain limitations in the generated captions. One of these is the model's tendency to prioritize references to people over objects related to pedestrian activities, such as stairs. However, our annotations show a higher frequency of references to objects other than "people," as seen in Figure 2. The observed outcome may be attributable to the inherent properties of the base model before fine-tuning.

5. Conclusion

We created a dataset of textual feedback within 360-degree images of urban environments to enhance pedestrian mobility support in cities. Traditional image captions are limited to describing the entire image and lack detailed information regarding street objects. Using the suggested dataset of Text360Nav, we can generate textual descriptions of street objects, offering valuable guidance to pedestrians, including those with visual impairments, to better understand urban scenes and navigate safely.

Ethics Statement

In this study, we gather an urban video dataset and captions obeying our institutional rules. We carefully apply privacy processing as described in Sec. 3.1 to conceal identification information. As our research purpose is not on describing people on urban scenes but enabling textual feedbacks from scenes, we consider these privacy processing doesn't distract our dataset purpose.

We instructe workers to select objects and attach descriptions that are helpful for visually impaired individuals to walk in urban scenes. However, to verify whether our dataset is truly effective for visually impaired individuals, it is helpful to conduct user study experiments with the visually impaired individuals in HCI methodology, which is out-of-scope in this dataset paper. During annotation creation, the workers were instructed to focus on aspects to which visually impaired pedestrians should pay attention.

Acknowledgments

This work was supported by JSPS Grant-in-Aid for Young Scientists (#22K17983), JSPS Fostering Joint International Research (A) (#22KK0184) and by JST PRESTO (#JPMJPR20C2). This work was partially supported by a JSPS Grant-in-Aid for Scientific Research (B) (#23H03686), and a JSPS Grant-in-Aid for Challenging Exploratory Research (#22K19822).

6. Bibliographical References

Miao Cao, Satoshi Ikehata, and Kiyoharu Aizawa. 2022. Field-of-view iou for object detection in 360° images. *CoRR*, abs/2202.03176.

Howard Chen, Alane Suhr, Dipendra Misra, Noah Snavely, and Yoav Artzi. 2019. Touchdown: Natural language navigation and spatial reasoning in visual street environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Shih-Han Chou, Yi-Chun Chen, Kuo-Hao Zeng, Hou-Ning Hu, Jianlong Fu, and Min Sun. 2018. Self-view grounding given a narrated 360 video. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Shih-Han Chou, Cheng Sun, Wen-Yen Chang, Wan-Ting Hsu, Min Sun, and Jianlong Fu. 2020. 360-indoor: Towards learning real-world objects in 360deg indoor equirectangular images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*.

Volkan Cirik, Taylor Berg-Kirkpatrick, and Louis-Philippe Morency. 2020. Refer360°: A referring expression recognition dataset in 360° images. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7189–7202, Online. Association for Computational Linguistics.

Jianglin Fu, Ivan V. Bajić, and Rodney G. Vaughan. 2019. Datasets for face and object detection in fisheye images. *Data in Brief*, 27:104752.

Karst M. P. Hoogsteen, Sarit Szpiro, Gabriel Kreiman, and Eli Peli. 2022. Beyond the cane: Describing urban scenes to blind people for mobility tasks. *ACM Trans. Access. Comput.*, 15(3).

Seita Kayukawa, Keita Higuchi, Shigeo Morishima, and Ken Sakurada. 2023. 3DMovieMap: An interactive route viewer for multi-level buildings. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI EA '23, New York, NY, USA. Association for Computing Machinery.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. 2023. Segment anything. *CoRR*, arXiv:2304.02643.

Jialu Li and Mohit Bansal. 2023. Panogen: Text-conditioned panoramic environment generation for vision-and-language navigation. *Advances in Neural Information Processing Systems (NeurIPS)*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 12888–12900. PMLR.

Yiyi Liao, Jun Xie, and Andreas Geiger. 2022. KITTI-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. *Pattern Analysis and Machine Intelligence (PAMI)*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Roberto Martin-Martin, Mihir Patel, Hamid Rezatofighi, Abhijeet Shenoi, JunYoung Gwak, Eric Frankel, Amir Sadeghian, and Silvio Savarese. 2021. JRDB: A Dataset and Benchmark of Egocentric Robot Visual Perception of Humans in Built Environments. *IEEE transactions on pattern analysis and machine intelligence*.

Su Wang, Ceslee Montgomery, Jordi Orbay, Vighnesh Birodkar, Aleksandra Faust, Izzeddin Gur, Natasha Jaques, Austin Waters, Jason Baldridge, and Peter Anderson. 2022. Less is more: Generating grounded navigation instructions from landmarks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15428–15438.

Jianxiong Xiao, Krista A Ehinger, Aude Oliva, and Antonio Torralba. 2012. Recognizing scene viewpoint using panoramic place representation. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2695–2702. IEEE.

Jinyu Yang, Mingqi Gao, Zhe Li, Shang Gao, Fangjing Wang, and Feng Zheng. 2023. Track anything: Segment anything meets videos. *CoRR*, abs/2304.11968.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *CoRR*, 2205.01068.

Pengyu Zhao, Ansheng You, Yuanxing Zhang, Jiaying Liu, Kaigui Bian, and Yunhai Tong. 2020. Spherical criteria for fast and accurate 360° object detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):12959–12966.

Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. 2022. Detecting twenty-thousand classes using image-level supervision. In *European Conference on Computer Vision*, pages 350–368. Springer.

Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Wang, Lijuan Wang, Jianfeng Gao, and Yong Jae Lee. 2023. Segment everything everywhere all at once. *Advances in Neural Information Processing Systems (NeurIPS)*.