都市環境における歩行者支援のための画像説明文生成用データセットの作成

西村 千恵子 [1]　栗田 修平 [2]　関 洋平 [1]
[1] 筑波大学大学院　[2] 理化学研究所
s2221661@s.tsukuba.ac.jp　shuhei.kurita@riken.jp　yohei@slis.tsukuba.ac.jp

概要

視覚障害者の歩行支援では、周囲の視覚情報を他の感覚に変換して提供することが必要である。そこで本研究では、画像説明文生成のデータセットを作成し、視覚言語モデルのファインチューニングを行った。現在の画像説明文の生成技術は、写真の意味理解等の分野においては大きく進歩したが、写真から特定の目的を持って説明するものはまだ少ない。そこで本研究では、視覚障がい者や晴眼者の歩行支援を目的とし、街頭の歩行動画から取得した画像から、歩行時の注意文を生成した。性能評価では、ファインチューニングを施したモデルにて、ファインチューニング前よりも画像中の局所的な障害物について情報提供ができることが示唆された。

1　はじめに

本研究では、視覚障がい者やスマートフォンのながら歩きする人々などの安全な歩行を支援するための画像説明文生成のデータセットの構築を行う。これにより、視覚情報が制限される歩行者が、例えば「信号の色」や「点字ブロックの位置」といった具体的なテキストフィードバックを通じて、安全に移動できるような注意文を生成することが可能になる。

現在の歩行支援研究では、物体検出技術に焦点が当てられているものが多い [1, 2, 3, 4] が、これは特定の物体の位置やクラスを把握するのみに限られる。対照的に、画像キャプションはより詳細で文脈を持った情報を提供し、歩行者が直面する状況や障害物に関するより包括的な理解を可能にする。したがって、本研究では、歩行支援に特化した画像キャプションのデータセットを構築し、評価を行う。

本研究の主な目的は、視覚障がい者やスマートフォン使用者の歩行中の安全を支援することである。このため、特に視覚障がい者にとって重要なオブジェクトや状況に焦点を当てた画像キャプションのデータセットを収集する。その後、収集したデータセットを活用して機械学習モデルのファインチューニングを行い、実際の環境での歩行支援に役立つテキストフィードバックの生成を目指す。

図 1　提案手法

2　関連研究

本研究は、視覚障害者の歩行支援を目的とし、馬場ら [4] と Chou ら [5] の研究を基にしている。例えば馬場ら [4] は、日本の歩道上の移動に特化した物体検出データセットの開発を行った。公開されたデータセットを利用すれば、点字ブロックや横断歩道、歩行者用信号などを高精度に認識することが可能である。しかし、物体認識によるラベルや位置情報のみで歩行支援を行う場合、カメラ画像などから検出された人や物体が「人、手すり、看板」だと示すことができるとしても、それらの関係性までは使用者にフィードバックすることができない。Chou ら [5] はナレーション付きの 360° ビデオが与えられたときに、そのビデオが注視している点 Normal Field of View（NFoV）を自動的にグラウンディング（位置決め）することを目指した。ただし、この研究のデータセットは公開されておらず、状況も歩行支援という文脈ではない。また、マルチコンテクス

トな 360° 画像に対し、ユーザークエリに従って描写対象を決める画像キャプションモデルも存在する [6]。このように、画像理解とテキスト生成を用いた歩行支援では、「何に注目すべきか」と「それがどのような状態か」の二つを説明できることが望ましい。本研究では、これらの方法を参考に、より広範囲な情報提供を目指し、360° 画像や動画からの詳細な説明を生成できるデータセットを作成する。

3 データセットの作成

本研究では画像説明文の生成に焦点を当てている。扱いやすさを考慮し、撮影した動画から画像を抽出し、画像に対してキャプショニングを行った。

3.1 動画の収集

多くの動画データセットは屋外・屋内問わず、通常の RGB カメラで撮影されている。例えば、一人称視点動画データセットである Ego4D[7] や EpichKitchen[8] などは、GoPro などのカメラを用いて撮影されている。しかし、本研究では、歩行時に役立つ広範囲の情報を提供するため、360° カメラを用いて撮影を行い、人間が街を歩く際に近い広い視野角の情報を含んだ動画を収集することにし、360° カメラである Ricoh Theta を使用した撮影を行った。撮影者が映り込む部分は削除し、より自然な人間の視点を再現した。撮影手法は、カメラを持った撮影者が一定の距離を歩きながら、周囲の景色を歩行中に撮影するという方法を採用した。撮影場所は、日本とアメリカの二カ国の複数箇所である。最終的に、約 5 時間分の映像を収集した。[1)]

3.2 前処理

本研究は視覚障害者の屋外歩行に役立つ情報を自動生成するための研究用データセットを作成することを目指しており、動画中の人物の特定などを目的とはしていない。しかし、屋外歩行の際に周辺にいる人物が映り込むことは、当該人物を撮影の主な対象にはしていないとしても、ある程度は避けられない。そこで本研究では、360° カメラで撮影した動画から画像を抽出し、プライバシー処理及び前処理を行った。まず最初に、プライバシー保護のため、顔やナンバープレートのぼかし処理とマスク処理を施した。性能については、いくつかのサンプルを手作業で精度を検証した。次に、アノテーションのために、動画を 30 秒単位で分割し、1 秒ごとの画像を抽出した。撮影された 360° 画像はパノラマ形式に変換し、物体検出により後のアノテーション作業でバウンディングボックスが表示できるようにした。最後に、撮影データを訓練、検証、テストに分けた。

3.3 アノテーション

この研究では、日本とアメリカの画像データに英語のアノテーションを行った。クラウドソーシングサービスの Amazon Mechanical Turk[2)] を利用し、ワーカーにアノテーションタスクを依頼した。ワーカーは主に英語圏の住民で、18 歳以上の登録者に限られ、報酬はシステムを通じて支払われた。アノテーション作成の手順は図 2 の流れに従い行われた。

1. 画像の選択　アノテーションの過程で、ワーカーは 30 秒間の動画から抽出された 30 枚の画像を確認し、関心を持った 3 枚に対してアノテーションを行うよう指示された。全ての画像にアノテーションを行わなかったのは、特定の画像に特徴的な要素がない場合、無理にアノテーションを加えると無意味な情報が含まれる可能性があるためである。

2. 言及する物体の選択　アノテーションの過程でワーカーは選択した画像内で言及すべき物体を選択した。画像には事前に作成されたバウンディングボックスが表示され、ワーカーはこれらのボックスを基に対象物を選択することができる。また、ボックスで囲まれていない物体についても、必要に応じてボックスを移動させて言及する物体を選択した。

3. アノテーションの作成　アノテーション作成では、「視覚障害者に伝えたいこと」というテーマで詳細な説明を付けるよう指示した。また、例文では、横断歩道や、階段、点字ブロックを例に挙げ、その方向や位置を説明するような内容を表示した。これは、物体検出だけでは伝わりにくい情報を視覚障害者に伝えることを目的としており、テキストフィードバックを通じて物体の危険性や位置などをより詳しく伝え、理解を深めることを目指すためである。さらに、歩行に関係しない遠くの窓や、画像と関係のない一般的なアドバイス「気をつけて歩く」などは不適切な例として記述しないように指示した。

1) 撮影の際には、人物の特定などにデータセットが悪用されないように、特定の人物にフォーカスした撮影は避け、街を歩いた際に常識的に視野に入る路上物体などを主な対象とするようにした。

2) https://www.mturk.com/

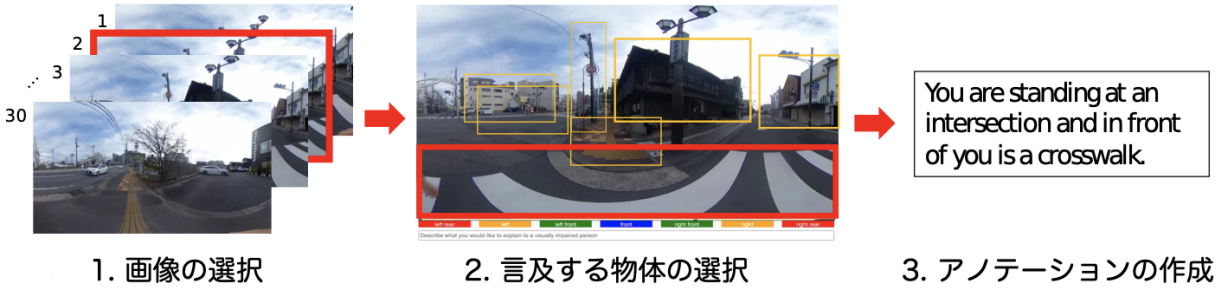

図 2　アノテーションの流れ

表 1　データセットの統計情報: ビデオの数、総時間、およびアノテーションの数

	ビデオ数	総時間	アノテーション数
訓練	915	7.6	7,744
検証	114	0.95	212
テスト	111	0.92	249
All	1,140	9.5	8,205

3.4　品質管理

本研究で使用したクラウドソーシングサービスは、18 歳以上であれば誰でも登録できるため、タスクを正しく理解していないワーカーも存在した。そのため、低品質なアノテーションの削除が課題となった。最初は評価が高いワーカーを指定するオプションを試みたが、十分な量のデータ収集ができなかった。そこで今回は、ワーカーごとに 10 件ほどの内容を目視で確認し、不適切な作業をしたワーカーのデータを全て除外するという方法を用いた。最終的に表 1 のように 7,744 件の学習用データ、249 件の検証用データ、212 件のテストデータを収集した。

3.5　品質管理

本研究で作成したデータセットの詳細を表 1 に示す。今回、30 秒の動画を合計で 1,140 の動画を収集した。そのうち、923 本（合計 7.6 時間）はアメリカのニューヨーク市、217 本（合計 1.8 時間）は日本の東京および関東地域の都市のものである。平均キャプション長は 14.72 単語、最短 7 単語、最長 53 単語、標準偏差は 5.25 となった。

表 2 は、アノテーションに含まれる上位 30 単語を分類分けしたものである。その結果、位置や方向に関連する単語が多く含まれていることが明らかになった。これは、物体の情報だけでなく、それについて具体的な情報を提供することを指示したためと考えられる。また、横断歩道や歩道に関する言及はある程度行われていたが、信号機やバス停、駅などに関する言及は少なかったことが明らかになった。この原因としては、遭遇率や、信号の視認の難しさの問題が考えられた。このように、位置や方向に関する単語が多く集まっていることが確認された他、これらを組み合わせることで、例えば「フェンスまでの大まかな距離」のような表現も獲得できた。

表 2　上位 30 単語の性質の分類分け

単語の性質	収集された単語
位置・場所	方向: front, right, left, 距離: steps, paces dozen, feet, rear その他: side, edge
交通設備 (交差点, 信号 横断歩道, 歩道 バス停や駅)	歩道, 道: sidewalk, street, road, path, way 横断歩道: crosswalk
障害物	car, steps, tree, pole, wall 階段: steps, stairs 建物: building, door, store
その他	person, set, curb, metal, sign

4　実験

作成したデータセットを使用して、ファインチューニングを受けたモデルの性能を定量的評価とアンケートによる質的評価の両方で行った。

4.1　定量評価

自動評価指標を用い、テキスト生成タスクの性能を調査した。ファインチューニングに使用したモデ

表 3　今回作成したデータセットに関する画像キャプション指標との性能比較

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE_L	METEOR	CIDEr	SPICE
ゼロショット								
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
ファインチューニング								
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

ルは BLIP[9] の base と large、BLIP-2[10] の OPT-2.7B と OPT-6.7B で、ファインチューニングの有無で性能を比較した。評価指標には、BLEU、ROUGE_L、METEOR、CIDEr、SPICE を使用した。

結果は表 3 のようになった。ファインチューニングしたモデルとゼロショットモデルの性能を比較し、ファインチューニングした BLIP-2 が最高の性能を示した。また、OPT-2.7B を使用した BLIP-2 モデルは、より大規模な OPT-6.7B よりも優れた性能を示した。通常、大規模モデルがより良い性能を示すとされる中、この逆の結果が見られた理由は、データセットの規模不足による OPT-6.7B の過学習が原因と考えられる。

4.2　定性評価

ファインチューニングしたモデルの文章生成の質をアンケート調査を通じて評価した。ファインチューニングした文章の評価を調べるため、比較対象としてゼロショットモデルで生成した説明文とワーカーによって作成された文章も評価した。なお、これらが表示される順番はランダムである。評価は Amazon Mechanical Turk を使用し、正確性、詳細さ、全体の 3 つの指標で 5 段階評価を行った。10 画像ごとの 3 つの文章に対する評価を 141 人のワーカーが行い、合計 1,410 の評価結果を収集した。

ファインチューニングのキャプションの品質を比較した。図 3 のようにファインチューニングされたモデルは、特定の対象（例えば、車や障害物）への言及においてより詳細で正確な情報を提供する傾向が見られた。これは、本研究の主な目的である、歩行に関連する局所的な対象を特定し、それらについて説明することが達成されていることを示している。一方で、ゼロショットモデルは画像全体の説明する傾向にあるが、図 4 駅構内や交差点のような場面での位置把握においては評価が高く、有効とされる傾向にあることも明らかになった。

生成モデル	評価	キャプション
ファインチューニング	4	a food truck parked on the side of the road
ゼロショット	2	a panoramic view of a food cart on a city street

図 3　ファインチューニングの高評価例

生成モデル	評価	キャプション
ゼロショット	5	a panoramic view of a city intersection
ファインチューニング	2	a person walking on the sidewalk near the crosswalk

図 4　ゼロショットの高評価例

5　おわりに

この論文では、歩行支援を目的とした画像説明文生成のために特化した 360 度都市画像のデータセットを作成し、その効果を評価した。アノテーションは 360° カメラで撮影した動画から抽出した画像を用い、Amazon Mechanical Turk を活用して、重要な交通設備や障害物を中心としたデータを作成した。定量評価ではゼロショットより高い評価を記録し、定性評価では、ファインチューニングされたモデルが歩行に関連する局所的な対象を特定し、それらについて言及することに優れていることが示唆された。

謝辞

本研究は JST さきがけ JPMJPR20C2 および科研費 JP22K17983, Microsoft Accelerate Foundation Models Research の支援を受けたものです.

参考文献

[1] Nazirah Hassan, Kong Wai Ming, and Choo Keng Wah. A comparative study on hsv-based and deep learning-based object detection algorithms for pedestrian traffic light signal recognition. In **International Conference on Intelligent Autonomous Systems (ICoIAS)**, pp. 71–76, 2020.

[2] Rodrigo F. Berriel, André Teixeira Lopes, Alberto F. de Souza, and Thiago Oliveira-Santos. Deep learning-based large-scale automatic satellite crosswalk classification. **IEEE Geoscience and Remote Sensing Letters**, Vol. 14, No. 9, pp. 1513–1517, 2017.

[3] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulò, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In **2017 IEEE International Conference on Computer Vision (ICCV)**, pp. 5000–5009, 2017.

[4] Tetsuaki Baba. Vidvip: Dataset for object detection during sidewalk travel. **Journal of Robotics and Mechatronics**, Vol. 33, No. 5, pp. 1135–1143, 2021.

[5] Shih-Han Chou, Yi-Chun Chen, Kuo-Hao Zeng, Hou-Ning Hu, Jianlong Fu, and Min Sun. Self-view grounding given a narrated 360 video. **Proceedings of the AAAI Conference on Artificial Intelligence**, Vol. 32, No. 1, 2018.

[6] Koki Maeda, Shuhei Kurita, Taiki Miyanishi, and Naoaki Okazaki. Query-based image captioning from multi-context 360$cdegree$ images. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, **Findings of the Association for Computational Linguistics: EMNLP 2023**, pp. 6940–6954, Singapore, December 2023. Association for Computational Linguistics.

[7] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Abrham Gebreselasie, Cristina González, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jáchym Kolář, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Ziwei Zhao, Yunyi Zhu, Pablo Arbeláez, David Crandall, Dima Damen, Giovanni Maria Farinella, Christian Fuegen, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4d: Around the world in 3,000 hours of egocentric video. In **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)**, pp. 18995–19012, June 2022.

[8] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. **International Journal of Computer Vision (IJCV)**, Vol. 130, p. 33–55, 2022.

[9] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation, 2022.

[10] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023.

倫理声明

この研究では都市部のビデオデータセットとキャプションを所属機関のルールに従って収集した。プライバシー処理を慎重に適用して、顔やナンバープレートを隠す処理を行った。私たちの研究の目的は都市のシーンにおける人々を記述することではなく、シーンからのテキストフィードバックを可能にすることであるため、これらのプライバシー処理がデータセットの目的を妨げることはないと考えている。

アノテーションの作業者には、都市部で視覚障害者が歩行するのに役立つオブジェクトを選択し、説明を付けるよう指示しました。しかし、私たちのデータセットが本当に視覚障害者にとって有効であるかを検証するためには、HCI 手法を用いた視覚障害者を対象としたユーザースタディ実験を行う必要があるが、今回はデータセット論文として発表する。