マルチモーダルモデルを用いた公共空間滞在者の行動観測に関する研究

高森　秀司[1]・岡野　将大[2]・吉田　龍人[3]・藤井　純一郎[4]

[1]正会員　八千代エンジニヤリング株式会社 技術創発研究所（〒111-8648 東京都台東区浅草橋 5-20-8）
E-mail: takamori@yachiyo-eng.co.jp (Corresponding Author)

[2]非会員　八千代エンジニヤリング株式会社 技術創発研究所（〒111-8648 東京都台東区浅草橋 5-20-8）
E-mail: ms-okano@yachiyo-eng.co.jp

[3]正会員　八千代エンジニヤリング株式会社 技術創発研究所（〒111-8648 東京都台東区浅草橋 5-20-8）
E-mail: ry-yoshida@yachiyo-eng.co.jp

[4]正会員　八千代エンジニヤリング株式会社 技術創発研究所（〒111-8648 東京都台東区浅草橋 5-20-8）
E-mail: jn-fujii@yachiyo-eng.co.jp

国土交通省を中心に「居心地が良く歩きたくなるまちなか」の形成に向けた取組が進められている．「居心地の良さ」それ自体は主観評価であるが，その代替・補完として，まちなか空間の滞在者の行動・活動を分析する取組がある．その調査は人手作業に大きく依存している現実があるが，機械学習を用いた画像認識を適用し，その自動化や効率化に関する検討が進められている．しかし，教師あり学習の取組では教師データ作成のコストが問題となっている．

本稿では，マルチモーダルモデルを用い，学習コストを低減化して空間滞在者の活動を分析する手法の試行結果を報告し，土木計画での活用について考察する．具体的には公共空間の歩行者等を撮影した映像を一定間隔で切り出した画像に対し，マルチモーダルモデルで VQA により滞在者の活動の把握を試行したものである．推論の出力結果の精度は，量的・質的ともに実務適用には現時点での課題があることを確認したが，動画から推論の出力までを自動化する取組は重要であり引き続き，検証・改善検討を進める．

Key Words: *Walkable, Activity Survey, Multi Modal Model, Visual Question Answering*

1. はじめに

国土交通省は居心地のよいまちなか空間が形成されているかどうかの把握と，それを活用した「居心地が良く歩きたくなるまちなか」の形成を目的として「まちなかの居心地の良さを測る指標（以下「指標」）」[1)]を作成・公表している．その指標においては，滞在者・通行者数等の「量」の把握に加えて，どのような活動が行われているかの「質」への着目が強く打ち出されている．

現状のアクティビティ調査の方法は，現地での直接調査や撮影映像に基づく目視カウントなど，人手による調査が中心で，コスト負担は運用上の課題となっている．前述した「指標」においても現地での計測シートへの記入が基本となっており，調査体制の確保面での課題は残されている．一方，ICT 技術の進展・普及を背景に，人手調査を AI カメラなどのツール活用で代替する取組も進められている[2)]が，導入・運用コスト等の課題は人手調査に同様であり，簡易で低負担な手法の開発に対するニーズがあるものと考える．

AI を用いて歩道空間内の歩行者の人数や動線等を解析した先行研究として，高森ら[3)]による取組などがあるが，通行者の活動等の「質的評価」には至っていない．

表-1　指標[1)]の「活用の手引き」内で例示される活動の種類

1	遊んでいる	7	私服やスーツなど多様な服装で来ている	13	読書・スマホ操作をしている	19	パフォーマンスをしている
2	食事をしている	8	仕事をしている	14	景色を眺めている	20	音楽にのっている
3	会話をしている	9	写真や動画を撮っている	15	陽だまり、日陰で過ごしている	21	勉強をしている
4	横になっている	10	運動をしている	16	散歩をしている	22	絵を描いている
5	個々の趣味の練習をしている	11	ぼーっとしている	17	お茶をしている	23	ボードゲームをしている
6	ペットを連れている	12	座っている	18	買い物をしている		

その背景の一つとして，教師学習による開発を進めるうえでの学習コストの過大さがある．まちなかでの活動には相当数の種類があり，各活動に対する教師用データの入手や分析コストは高い障壁となっている．

ここで，まちなかでの活動の同定の省力化に向けて，ChatGPT 等に代表される「LLM(Large Languege Model)」の活用に着目する．LLM は，提供者が膨大なデータセットに基づいたトレーニングを済ませた状態で提供されているため，学習コストが大幅に低減できる．

そこで，本研究では，LLM を使用したマルチモーダルモデルを用いて画像とテキストの組み合わせによる公共空間内の滞在者の行動観測を試行し，今後の土木計画における適用可能性について考察することを目的とする．

2. 行動観測の VQA 試行環境

本論で用いるマルチモーダルモデルは，著者らによる報告 [4]で開発した内容を準用している．また，解析対象とした画像は，武蔵野市が 2023 年に JR 三鷹駅北口のパブリックスペースで実施したオープンテラスの社会実験時の撮影動画を切り出して用いている．

動画から画像の切り出しや画像に対する VQA(Visual Question Answering)の推論環境は，AWS の SageMaker 上で ml.g5.xlarge を使用し構築した．

(1) マルチモーダルモデルの概要

本研究では，マルチモーダルモデルには Qwen2-VL-2B[5]を用いた．Qwen2-VL は画像とテキストを入力とするマルチモーダルモデルで，画像に対するキャプション生成や VQA を実施することができる．Qwen2-VL は，アリババクラウド社によるオープンソースであり，動画や画像の視覚理解能力が高いと評価される新しいモデルであることから採用した．生成モデルには，出力が安定しない場合があるが，Qwen2-VL は学習時に「Answer the question using a Yes or No only.」のようにプロンプトで出力形式を明示する学習を行っているため，指定した形式を安定的に生成が可能となっている．これにより，分析対象とする「まちなかの活動」の出力のブレを抑制する効果が期待できる．

(2) 適用画像

動画は，2023 年 10 月 21 日に JR 三鷹駅北口において撮影したものを用いた．社会実験の内容としては，歩行者空間内にカフェセット（テーブルとイス）を配置し，通行者は自由に滞在してよいという取組である．

撮影に使用したビデオカメラは，SONY 社の HDR-AS50 でカメラの解像度は 1920×1080px，フレームレートは 30fps とし，MP4 形式で保存したものである．設置時は一脚を用いて街灯に固定（事前に道路占用許可を取得）した．撮影高さは地上約 2.6m とし，歩道を斜めに見下ろす画角で撮影を行った．また，撮影範囲の広角性を優先するため，手振れ補正機能はオフにしている．

本論では，VQA の性能確認を趣旨とするため，入力は動画ではなく静止画像とし，画像は，動画を 5 秒間隔（150 フレーム単位）で切り出したものを作成した．

なお，現地での撮影は AM8:00 から 10 時間程度行っているが，本論での検証においては，カフェセット利用が比較的見られた時間帯の動画ファイルのみを採用し，3 ファイル合計で 90 分程度の動画を 1,073 枚の画像に切り出したものを用いている．

利用した画像イメージを図-1 に示す．動画の撮影範囲は図-1 の全体であるが，VQA の試行段階において，遠隔地（画像の両端部に該当）の通行者の認識にブレが大きく出力に影響しうること，また画面下段のベンチ（社会実験とは関係なく常設されている）の利用者の挙動を除外する趣旨から【⓪】の画角で切り出すこととした．

また，今後の社会実装を想定すると，画角は広い方が使い勝手はよいものと考えるが，事前の推論の試行において，同一画像内に複数の推論対象が存在した場合の回答のブレが散見されたことから，対象範囲を区分した場合の推論結果との比較のために，⓪の画角から，主として通行空間に該当する【①】，社会実験のカフェセットが設置されている【②】，【③】の範囲をそれぞれ切り出して，個別に推論を実施することとした．

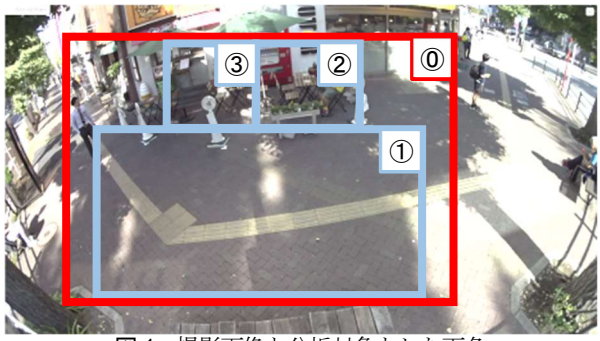

図-1 撮影画像と分析対象とした画角

(3) VQA の推論

切り出しを行った画像群を対象に，表-2 に示す 14 の質問を連続的に推論させ，推論結果を csv として出力する仕組を構築し，自動化を図った．推論に要した時間は入力画像によって差があり，図-1 の【⓪】の場合は約 6 秒/枚，【①】の場合で約 4 秒/枚，【②・③】の場合で約 3 秒/枚程度であった．30 分相当の動画から切り出した画像群の推論時間は，画像サイズが⓪の場合でも 30～40 分程度であり，人手によるカウントと比較（著者の人手カウントでは，10 分動画の場合で動画時間の数倍の時間を要した）して，省力化は十分に図られている．

プロンプトは英文で作成することとし，事前に同じような設問を異なる表現で試行したうえで，ある程度安定的な回答が得られる表現を採用した．なお，より適切なプロンプトの作成・検討は今後の課題である．

また，各プロンプトの文頭には「Answer the question using a Yes or No only.」あるいは「Answer the question using a Number only.」を付記しており，回答の出力について制約をかけることで回答の扱いやすさを向上している．

表-2　VQA のプロンプト

No.	確認要素	プロンプト
1	人物の有無	Are there people in the image?
2	人数	How many people are in the image?
3	女性の有無	Are there any women in the image?
4	子供の有無	Are there children in the image?
5	座っている人の有無	Are people sitting on the chairs in the image?
6	座っている人数	How many people sitting on the chairs in the image?
7	座っている人の活動：食事	Are people sitting on the chairs eating in the image?
8	座っている人の活動：会話	Are people sitting on the chairs talking in the image?
9	座っている人の活動：携帯電話	Are people sitting on the chairs have a mobile phone in the image?
10	立っている人の有無	Are there people standing in the image?
11	立っている人数	How many people are standing in the image?
12	自転車に乗っている人の有無	Does anyone ride a bicycle?
13	自転車の台数	How many bicycles are in the image?
14	幸せそうに見える人の有無	Does anyone look happy in the image?

3．行動観測の VQA の結果

2．で構築した環境と画像群を用いて，プロンプトに基づく VQA の推論を行った結果は以下のとおりである．

(1)　同一条件で複数回実施時の出力のブレの有無

a)　画像群に対するプロンプトに対する出力

同じ画角【⓪】の全画像（1,073 枚）に対して，異なるタイミングで同じプロンプト（表-2）を与えた時の出力に差があるかについて確認を行った．その結果，すべての画像においてプロンプトに対応した出力結果は同一であり，安定的な回答を返したことを確認した．

今回のプロンプトにおいて，文頭で【Yes / No】や【数値のみ】を回答するよう指定したことは，出力の安定化にも寄与した可能性がある．

b)　自由回答形式でのプロンプトへの出力

Qwen2VL では自由回答による出力も可能である．例えば，プロンプトの文頭で出力形式を指定せず「座っている人は何をしているか；What is the person sitting in the image doing?」を推論させた結果の一例を図-2 に示す．

なお，ここでは「誤った出力」を例示している．

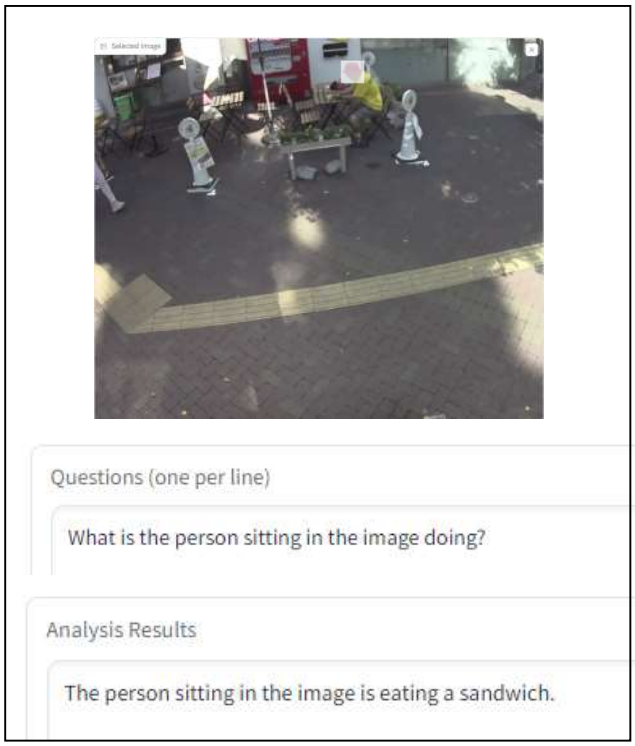

図-2　自由回答による推論結果の例（画角⓪）

画像上では，椅子に座った男性は携帯をテーブル上で触っているものと目されるが「座っている人は何をしているか」の問いに対し「サンドイッチを食べている」と出力された．また，複数回試行しても同じ出力であった．

本画像を含めた複数の画像に対して，自由回答形式で同じ文章のプロンプトを複数回与えた結果では同一の回答が出力される結果が確認され，同じ入力条件であれば基本的に同じ出力を返す傾向を確認している．ただし，十分な試行回数を行ったとは言えないことから，自由回答形式に対する出力の安定性の検証は今後の課題としたい．なお，補足として，テーブル上に何かが置かれている場合に「食事」と出力される傾向が見られている．

また，図-2 の画面に対して，以下の 4 つのプロンプトを与えた場合の回答の出力は以下のとおりであった．

＜プロンプト＞

① What is the person sitting in the image doing?
② Answer the question using a Yes or No only. Are people sitting on the chairs eating in the image?
③ Answer the question using a Yes or No only. Are people sitting on the chairs talking in the image?
④ Answer the question using a Yes or No only. Are people sitting on the chairs have a mobile phone in the image?

<出力>
① The person sitting in the image is eating a sandwich.
② [食事しているか] Yes
③ [会話しているか] No
④ [携帯電話を手にしているか] Yes

出力の「①サンドイッチを食べている」と「②Yes：食事をしている」，また「③No：会話していない」の3項目の間に矛盾はないが，「④Yes：携帯電話を手にしている」は，①・②での「食事している」行動と同時に出力される結果となっている．

本論で構築した環境では，各プロンプトは独立して推論し，前の推論結果を踏まえた推論は行わない条件としていることから，上述したような同時に成立しにくい結果の両方を出力することがある．

⑵ 単純な設問間での矛盾回答の傾向

本論の推論で採用したプロンプト（**表-2**）は，【Yes-No】か【数値】による出力を指定しているが，出力される結果によっては矛盾回答となりうる組み合わせが存在する．例えば，「No.1：画像中に人はいるか」に対して【Yes：人がいる】と回答した画像に対して，「No.2：画像中に何人いるか」で【ゼロ人】と出力する場合である．（逆の組み合わせもある）．

画角⓪～③のそれぞれについて（画角範囲は**図-1** 参照），【No.1 と No.2】，【No.5 と No.6】，【No.10 と No.11】の組み合わせにおける矛盾回答の発生状況を整理した結果は，**表-3～表-6** のとおりである．

表-3 画角⓪の矛盾回答数 [N=1,073]

出力パターン	Yes／ゼロ人	No／1 人以上
No.1／No.2	7 （0.65%）	7 （0.65%）
No.5／No.6	9 （0.84%）	24 （2.24%）
No.10／No.11	1 （0.09%）	125 （11.65%）

表-4 画角①の矛盾回答数 [N=1,073]

出力パターン	Yes／ゼロ人	No／1 人以上
No.1／No.2	6 （0.56%）	1 （0.09%）
No.5／No.6	0 （0.00%）	59 （5.50%）
No.10／No.11	9 （0.84%）	24 （2.24%）

表-5 画角②の矛盾回答数 [N=1,073]

出力パターン	Yes／ゼロ人	No／1 人以上
No.1／No.2	0 （0.00%）	6 (0.56%)
No.5／No.6	0 （0.00%）	49 (4.57%)
No.10／No.11	0 （0.00%）	122 (11.37%)

表-6 画角③の矛盾回答数 [N=1,073]

出力パターン	Yes／ゼロ人	No／1 人以上
No.1／No.2	0 （0.00%）	18 （1.68%）
No.5／No.6	2 （0.19%）	23 （2.14%）
No.10／No.11	3 （0.28%）	34 （3.17%）

当初，分析対象となる画像の広さから，画角⓪，次いで画角①における矛盾回答が多いことを想定していたが，画角②での矛盾回答が最多の結果となった．また，矛盾回答の生じやすさは，「人がいるか」に対して【No：人はいない】と出力したあとに「何人いるか」に対して【1 人以上】を出力する組み合わせに偏っている．

設問の組み合わせでは【No.10：立っている人はいない】が【No.11：立っている人は1人以上】が多く，次いで【No.5：座っている人はいない】が【No.6：座っている人は 1 人以上】の組み合わせで，矛盾した推論が多くなる傾向がある．

a) 【Yes／ゼロ人】の組み合わせの例

表-3～表-6 において【Yes／ゼロ人】のパターンで矛盾した出力は比較的少ないが，該当した画像を以下に例示する．

【No.1：Yes】／【No.2：0 人】|【No.5：Yes】／【No.6：0 人】
図-3 矛盾回答の画像例（画角⓪）

図-3 の左図では，左上の店舗で注文している方が No.1 では認識されたものの No.2 では対象外となったこと，右図では，オープンテラス席への着席はないが，画面右下の方が No.10 で「座っている」，No.11 のカウント段階では対象外となったことなどの可能性が考えられる．

なお，当該画角の下部（画面外）には，歩道に備え付けのベンチがあり，実際には座っている状態であるが，画面だけではわからない条件（入力画像からは座っていると考えてもよい条件）となっている．

b) 【No／1 人以上】の組み合わせ

【No／1 人以上】のパターンで矛盾回答に該当した画像を以下に例示する．

【No.1：No】／【No.2：1 人】|【No.10：No】／【No.11：2 人】
図-4 矛盾回答の画像例（画角⓪）

図-4 の左図では，画面右端の歩行者（半身のみ）に対する評価がズレたものと想定される．右図は「No10：立っている人はいない」と回答されつつ「NO.11：立っている人は 2 人」という出力の組み合わせとなっているものである．画面左端の歩行者は足のみが映っており，それに対する評価がズレたと考えれば図-4 左図と同様の事象と考えられるが，一方，画面右側にいる歩行者が「No.10：立っている人はいない」と推論されてしまう場合がある点には留意が必要である．

参考として，図-4 右図と同じような画像の図-5 に対する推論では「No.10：立っている人はいる」「No.11：立っている人数は 1 人」で正常な出力を得ている．

図-5　正常な推論の画像例（画角⓪）

また，「No.11：座っている人数」の推論が，連続する 2 枚の間で異なる結果となった画像を以下に示す．

図-6　座っている「人数」にズレのある画像（画角①）

図-6 は両方とも「No.5：座っている人はいない」と回答された画像（左図から 5 秒後が右図）であるが，座っている人数は「左図：0 人」，「右図：1 人」として出力されている．

「座っている」「立っている」のいずれにしても出力には誤差がありうること，人数カウントをさせた場合に評価が変わることがあることにも留意が必要である．

本論では，「推論時に画像のどこを見ているか」については未確認のため，今後の確認が必要である．

(3)　人手評価と推論結果との比較

以下，画角②（画面右側のカフェセット周辺を切り出した画像）を対象に，各プロンプトに対する「人手での評価結果」と「推論による出力結果」を比較した結果を表-7 に示す．

人手による評価は，画角②の撮影動画を対象とした．複数名が入りやすい画角⓪や①では，どの人物を対象に行動の推論が行われているかの想定が困難であることから，比較的狭い領域を対象とする趣旨である．広範な対象領域に対する評価は今後の課題としたい．

画角②の動画データのうち，カフェセットでの滞留が比較的長く見られた時間帯を含めた 30 分相当（5 秒間隔で 360 枚）を切り出した画像を対象に行った．以下，その評価結果を「正」とする．

なお，「No.14：幸せそうに見える人」は評価対象外としているが，参考として，360 枚中 5 枚で出力（ただしカフェセットを利用中の同一人物）されている．

表-7　人手調査結果との比較　　[N=360]

v	確認内容	エラー出力数	備考
1	人物の有無	6	[正] 人物有：202 枚
2	人数	65	ー
3	女性の有無	17	[正] 女性有：23 枚
4	子供の有無	79	[正] 子供有：0 枚
5	人が座っている	28	[正] 座っている：165 枚
6	座っている人数	90	ー
7	食事	90	[正] 食事有：6 枚
8	会話	66	[正] 会話有：0 枚
9	携帯電話	69	[正] 携帯有：107 枚
10	立っている人	23	[正] 立っている人：66 枚
11	立っている人数	53	ー
12	自転車利用	3	[正] 自転車利用有：5 枚
13	自転車の台数	3	[正] 自転車利用有：5 枚

a)　No.1：人物の有無に関する推論結果

基本的にはエラー率は低く，「No.1：人物の有無」でエラー出力となった 6 枚中 5 枚は，ほぼ連続する画像（当該動画の撮影開始から 380 秒～405 秒の間に含まれる 5 枚）であった．画像の例を図-7 に示す．三角コーンを人物と誤認した可能性も考えられるが，別の時間帯の同様な画像においては「人物はいない」との出力が支配的であり，本推論の原因については現時点で不明である．

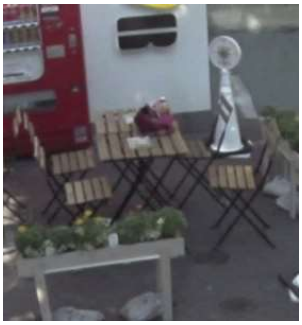

図-7　人物が「いる」と出力された画像の例（画角②）

残る 1 枚は，人物が「いない」と推論されたものである．画像を図-8 に示す．画面左端に，通過後の人物の背中が映っている状態であるが，当該画像範囲だけでの推論のエラーとしては許容範囲と考える．

上記の原因不明の挙動もあるが「人物の有無」は概ね信頼できる推論であると考える．

図-8　人物が「いない」と出力された画像（画角②）

b)　No.2：人数に関する推論結果

65 ケースのエラー出力のうち，6 ケースは a) No.1 のケースに同様である．「①：1 人を 2 人と出力」したもので 50 ケース，また「②：2 人を 3 人と出力」したものが 9 ケースである．

なお，①・②の計 59 ケースのうち 58 ケースは，図-9 に映る利用者を含む画像で発生しており，なんらかの誤認要素が含まれているものと考えられるが，詳細の分析は今後の課題とする．また，当該利用者は，118 枚連続（約 10 分間）カフェセットを利用している．

図-9　人数カウントが増加しやすい画像の例（画角②）

c)　No.3・No.4：女性・子供の有無

「No.3：女性の有無」のエラーでは，「①：女性がいないのにいると出力」したのが 10 ケース，「②：女性がいるのにいないと出力」したのが 7 ケースである．

また，「No.4：子供の有無」は，実際には子供がいないのに子供ありと出力したケースが全て（全 79）である．なお，その全てが，図-9 の利用者の画像であった．当該利用者のみが映る場合の画像で「子供は無」との出力もあり，一概には言えないが，対象者の服装の色味等から判断された可能性も考えられる．当該画像内の誤認要素については，改めて確認する必要がある．

d)　No.5・No.6：座っている人の有無・人数

「No.5：座っている人の有無」のエラーでは「①：座っていないのに座っていると出力」したのが 6 ケース，「②：座っているのに座っていないと出力」したのが 22 ケースである．

①の 6 ケース中 5 ケースは，a) No.1 の図-7 で例示したケースに同様である．残る 1 ケースは，図-10 のとおりである．ベンチではないが，自転車に座っている状態であり，エラーとしては許容範囲と考える．なお，②の 22 ケースは，全て図-9 の利用者を含む画像である．

図-10　（椅子に）座っていると出力された画像（画角②）

「No.6：座っている人数」のエラー出力のうち 5 ケースは a) No.1 の図-7 で例示したシーンに同様である．また人数の増加も，ほぼ b) No.2 で示したケースに該当しており，1 人が 2 人としてエラー出力されているケースが多い．異なるシーンで人数のエラー出力があった画像を図-11 に示す．

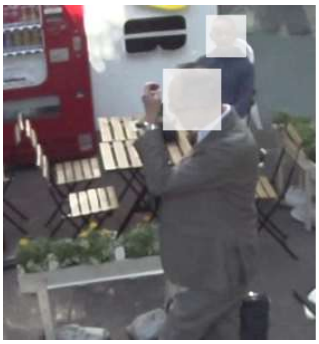

図-11　座っている人数がゼロと出力された画像（画角②）

e)　No.7・No.8・No.9：座っている人の行動内容

「No.7：食事」「No.8：会話」「No.9：携帯電話」はマルチモーダルモデルの活用により，学習なしでの把握可能性を試行する「行動」に該当する．

本論で行動内容の評価対象とした画像群（30 分）において，オープンテラスを利用した利用者は 2 名である．うち，図-9 に例示した利用者が約 10 分，また，図-11 に例示した利用者が約 3.5 分である．ただし，図-11 の利用者は，缶コーヒーを飲んだ以外は，ほぼ休憩（座っている）であり，「No.7～No.9 をしているか？」の各問いに

対しては，おおむね安定的に「していない」を出力する結果となった．なお，缶コーヒーを飲用したシーンは「食事していない」と出力され，エラーとして扱った．

図-9 の利用者は，着席している時間帯は，ほぼ携帯電話を触っていたが，テーブル上にモノが置いてあることで「食事」や，利用者と通行人の頭部が近接した場合に「会話」として出力されている傾向がある．

図-12　会話していると出力された画像（画角②）

3 種類の行動の把握を志向したが，座っているだけの場合に，各行動を「していない」出力は一定の精度を有しているが，「している」場合のブレ幅は大きく，現時点では実務適用に課題がある．

f)　No.10・No.11：立っている人の有無・人数

「No.10：立っている人の有無」のエラーでは「①：いないのに立っている人がいると出力」したのが 17 ケース，「②：立っている人がいるのにいないと出力」したのが 6 ケースである．23 のエラーケース中 20 ケースが図-9 の利用者のシーンである．図-4 でも記述したとおり，明確に歩行者と思われる場合でも「いない」と評価されるケースが，数としては多くないが発生している．

図-13　立っている人がいないと出力された画像（画角②）

「No.11：立っている人数」では 53 のエラーケースがある．「①：立っている人がいないシーンで「1 名～2 名」と出力」が 39 ケース，「②：1 名が立っているシーンで「2 名」と出力」が 14 ケースである．なお，53 ケース中 50 ケースが図-9 の利用者のシーンであり，エラーの発生が偏在する傾向がある．

g)　No.12・No.13：自転車利用者・台数

自転車に関連する画像は多くないことから，全体を図に示す．図-14 および図-16 は，自転車を認識して出力された画像，図-15 は自転車利用なしと出力された画像である．図-17 はエラー出力である．本試行での自転車利用の検出精度は 5 割程度であり，高くない結果であるが，図-16 程度の画像でも検出する点には可能性がある．


図-14　自転車利用者として出力された画像（画角②）

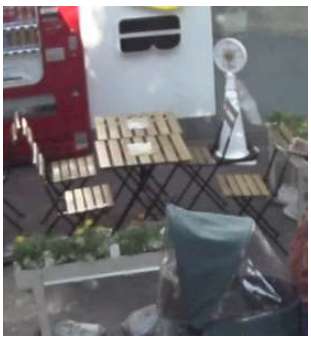


図-15　自転車利用なしと出力された画像（画角②）

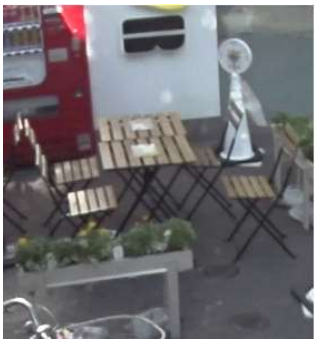

図-16　自転車利用者（No.12）はいないが自転車台数（No.13）にはカウントされた画像（画角②）

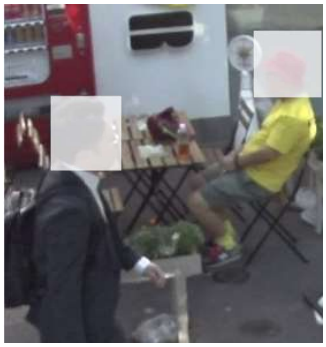

図-17　自転車利用者として出力された画像（画角②）

４．推論結果を踏まえた考察

本論では，人手による評価結果を正とし，マルチモーダルモデルによる推論結果との比較を行った．

事前学習等のコストを要せず，また画像の切り出しから推論までを自動化することにより，動画のインプットから推論のアウトプットまでの作業時間や手間については，人手調査に比較して相当の低減化が図られている．

一方，画像からの行動把握の精度には課題がある．本論での試行の範囲では「No.1：人物の有無」は十分な精度を有しているが，それ単体では有用な情報とは言えない．まちなかの実態調査において，定量的な把握は基本的な要件であるが，現時点で人数等の定量的な把握にはエラー出力が多く，現時点での結果からは実務適用には課題がある．定量把握の精度向上に向けては，行動把握の精度向上の工夫を行ったうえで，プロンプトでの矛盾回答を許容せず，回答結果を踏まえた分岐を繰り返して細分化していくやり方なども考えられる．また，滞在時間の把握においては，本論では 5 秒間隔で画像の切り出しを行っており，5 秒以内に人の入れ替わりは起こりにくいとしてよければ，出力の連続性の整理から整理が可能になると考える．いずれにしても定量的な分析の精度向上は必要なため，例えばマルチモーダルモデルの適用のみに限定せず，先行研究 [3) 6)]等での物体検出と追跡による手法も参考に，複数の手法の組み合わせも選択肢に検討を深化したい．なお，本論での試行では，エラー出力が特定のシーンに偏在する傾向も確認されているため，その要因の分析も今後の課題である．

定量的な評価に加え，質的な評価（行動内容の把握）が本論の主な目的であった．指標 [1)]も念頭に，オープンテラスで観察される可能性が高いと思われる３行動（飲食，会話，携帯電話）を選定して試行した．検出精度には課題があるが，本論での人手による評価を行った画像群における滞在者の行動の種類がそもそも少なかった（携帯電話と休憩に特化していた）こともあることから，まずは比較検証を拡大することから始めたい．並行して，プロンプトの工夫も重要と考える．

今回の分析では，広範囲の画角⓪ではなく，限定的な画角②の範囲を対象に比較を行った．その背景は，複数の人物がいる場合に，どの対象に対する出力であるかが不明確になりやすく，出力に対する解釈等が別途必要になることを回避しようとしたことにある．自動化を図るうえで，単純な出力とすることを志向した結果であるが，一方でまちなかでの行動把握のための取組は，一定以上の広範囲を対象としていくことが求められる．推論時にどの領域を重視しているかに関する分析等や，並行して最新の技術動向調査も行いながら，まちなかの行動把握と分析に資する効果的な枠組みの構築を目指したい．

近年のスマートシティやデジタルツイン形成の機運において，まちなかのデータ収集の高密度化等は加速的に進展していくことが想定される．人口減少時代においてこそ，歩きやすく居心地のよいまちなか空間を形成し，にぎわいや活力，あるいは高質な暮らしを創出していく取組が重要になる．まちの変容には一定の時間が必要であるが，その観点からも EBPM を常態化していく取組が重要になる．データの量も質も適切に扱えるようにするための技術開発・検討を継続する．

謝辞：本研究で活用したビデオデータは，武蔵野市都市整備部まちづくり推進課の全面的なご協力により撮影したものです．ここに記し，謝意を表します．

REFERENCES

1) まちなかの居心地の良さを測る指標（改訂版 ver.1.1）[Indicator to measure the comfort of the town center(ver.1.1)]
2) 例えば柏市での歩行者通行量調査の取組 https://www.city.kashiwa.lg.jp/chushinshigaichi/kamerahokousya.html (参照 2024-09-30) [e.g. Pedestrian Traffic Survey Efforts in Kashiwa City, https://www. city. kashiwa. lg.jp/chushinshigaichi/kamerahokousya.html (accessed on Sep.30, 2024)]
3) 髙森 真紀子，大久保 順一，藤井 純一郎：都市空間での人流解析における深層学習の応用，第 2 回 AI・データサイエンス論文集，2021. [Takamori, M., Okubo, J., Fujii, J., Application of deep learning in human flow analysis in urban spaces, Proceedings of the 2nd AI and Data Science, 2021]
4) 岡野 将大，吉田 龍人，藤井 純一郎，高森 秀司，天方 匡純：マルチモーダルモデルによる公共空間の行動認識，第 38 回人工知能学会全国大会，2024 [Okano,M., Yoshida, R., Fujii, J., Takamori, S., Amakata, M., Multimodal Model for Recognizing Behavior in Public Spaces, The JSAI International Symposia on AI, 2024]
5) https://arxiv.org/abs/2409.12191(参照 2024-09-30) [https://arxiv.org/abs/2409.12191 (accessed on Sep.30, 2024)]
6) 高森 秀司，吉田 龍人，堀井 大輔，菊池 恵和，大久保順一：AI による人流解析結果を介した歩道空間の特性把握の可能性に関する研究，AI・データサイエンス特別シンポジウム「デジタルツイン」論文集，2023 [Takamori,S., Yoshida, R.,Horii, D., Kikuchi, Y.,Okubo, J., Research on the possibility of understanding the characteristics of sidewalk space through the results of human flow analysis by AI, Proceedings of the Special Symposium on AI and Data Science “Digital Twin”, 2023

(Received October 4, 2024)