基盤モデルを用いた都市公園利用者数の自動計測手法の提案

○岡野 将大†, 武田 陸‡, 藤井 純一郎†
†: 八千代エンジニヤリング株式会社 技術創発研究所
‡: 八千代エンジニヤリング株式会社 九州支店 都市デザイン部
ms-okano@yachiyo-eng.co.jp

概要 : 本研究では，都市公園における利用者数の自動計測手法を提案する．都市公園の管理運営において利用者数を計測することに需要があるが，人手による計測は人的コストの高さに問題がある．自動計測手法の 1 つである CNN に基づく画像認識は，高精度を達成しているが，複数地点への展開には課題がある．基盤モデルによるゼロショット手法は追加学習なしで使える汎用性が期待されるが，基礎実験では背景の誤検知と密度変化時の精度変動が確認された．

そこで，SAM2 による背景除去，CLIP-EBC によるゼロショット群衆カウント，密度閾値処理を統合した 3 段階の処理で構成される手法を構築した．都市公園 6 地点で実験を行い，基礎実験では 16.92 人だった MAE が提案手法により 3.15 人に改善した．
＜キーワード＞ 群衆カウント，公共空間，基盤モデル

1. はじめに

都市公園の利活用や管理運営の最適化は重要な社会的課題となっている[1]．利用者数の把握は，管理最適化の基礎データとなるが，人手による計測は，時間的・人的コストが高いため，継続的な計測の実現に課題がある[2].

これらの課題に対して，画像処理技術を活用した自動計測手法の導入が検討されている．人数を計測するタスクとして群衆カウントタスクがあり，CSRNet[3]などの CNN ベースのモデルが高い計測精度を実現している．しかし，これらのモデルを屋外空間に適用する際には，樹木や建造物などの背景要素が誤カウントを生むことがある．また，新しい環境で高い精度にするためには大量の教師データとアノテーション作業が求められる．そのため複数地点への展開には課題がある[4].

本研究は，基盤モデルによるゼロショット手法に着目する．基盤モデルは大規模データで事前学習されているため，追加学習なしに様々なシーンで安定して高い精度が期待できる．しかし，基礎実験では都市公園環境においては（1）樹木や建造物による背景誤検知や，（2）閑散時と混雑時の利用者密度変化での精度変化によって，安定した人数把握が難しいことが予測される．

そこで本研究では，都市公園環境における背景誤検知抑制と密度変化に対応した，基盤モデルによるゼロショット人数計測手法を提案し，その計測精度を評価することを目的とする．

図 1:撮影画像例
※対象が画像全体と比べ小さいため切り取り画像

図 2:CLIP-EBC の密度マップの例

2. 群衆カウント基盤モデル（CLIP-EBC）

2.1. CLIP-EBC の動作原理

CLIP-EBC (CLIP Enhanced Blockwise Classification)[5]は, 大規模視覚言語モデル CLIP [6]を基盤として，ブロック単位での密度分類より群衆カウントを実現する．図 1 に撮影画像例を，図 2 に図 1 に対する CLIP-EBC の密度マップの例を

示す．従来手法が密度マップ回帰を用いるのに対し，CLIP-EBC は画像を複数のブロックに分割し，各ブロックの人数を分類問題として解く．この利点は，異なる密度スケールに対して柔軟に対応できる点と，大規模データで事前学習された CLIP の汎用的な特徴表現を活用しながら，追加学習なしで群衆カウントを可能にする点にある．

2.2. 都市公園環境での基礎実験

提案手法の基礎となる CLIP-EBC の性能を評価するため，都市公園 6 地点で撮影した画像データセットを用いて実験した．実験では，CLIP-EBC のみで，追加の前処理や後処理を行わない状態で計測精度を評価した．各地点から 30 分間隔で撮影した画像計 120 枚（各地点 20 枚）を使用し，人手による計測値を Ground Truth (GT)として精度を検証した．

精度評価には，MAE (Mean Absolute Error)・RMSE (Root Mean Square Error)・MAPE (Mean Absolute Percent Error)・MaxE (Max Error) の 4 つを採用した．MAE は平均的な誤差の大きさを，RMSE は大きな誤差をより強く反映する指標である．MAPE は誤差の相対的な割合を評価する．MaxE は最大誤差を示す．

2.3. 基礎実験結果の分析と問題

表 1 に基礎実験の地点別計測精度，データ全体での計測精度を示す．MAE は 16.92 人，RMSE は 23.35 人となった．地点ごとの精度にばらつきが見られ，地点 3 では MAE 2.40 と良好な精度を示した一方で，地点 1 で MAE 30.40，地点 5 で MAE 30.80 と大きな誤差となった．

これらの結果より，次の 2 つが確認された．

1. 樹木や建造物による背景誤検知
   地点 1（図 3）と地点 5 では樹木の葉や設置されている物体を人物として誤検知した．CLIP-EBC の密度マップでは，これらの背景要素に対しても密度値が割り当てられており，実際の人数を上回る予測値が出力された．
2. 利用者密度の変化による精度変化
   表 1 の MAPE が示すように，同一手法でも利用者数によって精度が大きく変動した．地点 1（GT：100 人）のような閑散時には MAPE が 371.62%と過大評価の傾向が顕著であったのに対し，地点 6（GT：1,014 人）のような混雑時には MAPE が 10.66%であった．図 4 に示すように，低密度環境では背景ノイズのわずかな誤検知でも相対誤差が大きくなった．

これらの結果から，CLIP-EBC を都市公園環境で実用化するためには，樹木や建造物による背景誤検知の抑制と，利用者密度変化に対応した計測手法が求められることを確認した．

表 1：CLIP-EBC 単体での地点別計測精度

地点	MAE [人]	RMSE [人]	MAPE [%]	MaxE [人]
1	30.40	34.51	371.62	54
2	12.40	14.15	40.21	23
3	2.40	3.41	11.53	8
4	15.00	16.38	66.03	24
5	30.80	38.15	99.95	80
6	10.50	12.04	10.66	18
全体	16.92	23.35	99.99	80

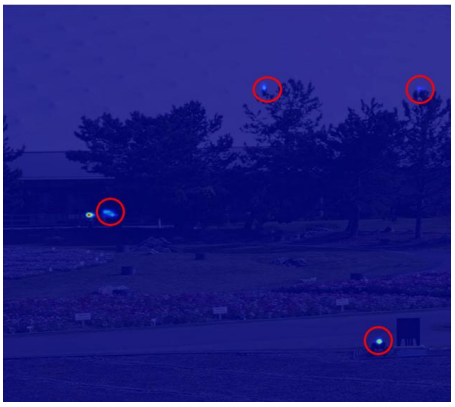

図 3：1. 樹木や建造物の誤検知例
※対象が小さいため切り取り画像

(1) 利用者：高密度時　　(2) 利用者：低密度時
図 4：異なる利用者密度画像に対する密度マップ比較

3. 提案手法

図 5 に本手法の処理フローを示す．都市公園環境での適用時における問題に対し，基盤モデルを用いた 3 段階処理手法を提案する．

① SAM2 による背景誤検知の抑制
   固定カメラ環境の利点を活かし，事前に樹木や建物にポイントプロンプトを設定して SAM2 (Segment Anything Model 2)[7]でマスク

画像を生成し，人数計測の対象外領域として設定することで背景誤検知を抑制する．

② 基盤モデルによる群衆カウント

基盤モデルである CLIP-EBC で，ゼロショット群衆カウントを行う．

③ 閾値の調整による密度変化への対応

CLIP-EBC 出力における密度環境変化時の精度変動を抑制するため，密度閾値処理を導入する．各ブロックの予測人数に対し閾値を設定し，閾値以下の予測値は 0 に補正することで，低密度時の誤検知を除去する．最終的な人数は補正後の全ブロックの人数の総和として算出する．本研究では，閾値を人手で行った．

本手法により，追加学習なしで屋外特有の背景誤検知と密度変化の両問題に対応した．

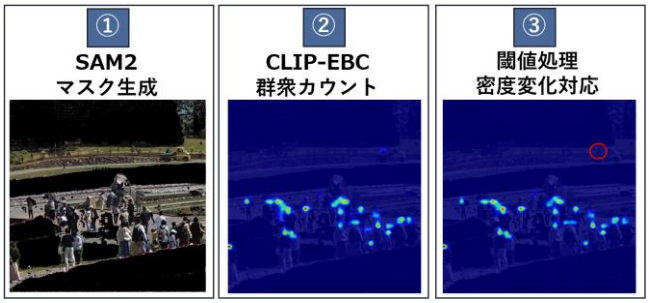

図 5:提案手法処理フロー図

4. 実験概要

本研究では，山口きらら博記念公園内の 6 地点に設置した固定カメラで撮影した画像を用いる．画像は 30 分間隔で撮影している．各画像は樹木密度や建物の種類などが異なる．

実験は人手計測値を GT として評価した．基礎実験での CLIP-EBC の結果をベースラインとし，各処理が計測精度に与える影響を調査するためアブレーション分析を行う．アブレーション条件として以下の 5 つを設定した．

1) SAM2 のみ
2) 密度閾値を一定のみ
3) 密度閾値の調整のみ
4) SAM2＋密度閾値一定
5) SAM2＋密度閾値調整

精度評価には，MAE・RMSE・MAPE・MaxE の 4 つを採用した．これにより，SAM2 の処理と密度閾値処理が計測精度に与える影響を定量的に検証し，都市公園環境における提案手法の精度を評価した．

表2:全データでの精度評価

実験内容	MAE [人]	RMSE [人]	MAPE [%]	MaxE [人]
CLIP-EBC	16.92	23.35	99.99	80
+SAM2	3.65	4.87	16.62	15
+閾値一定	10.50	15.42	67.42	50
+閾値調整	4.78	6.18	24.81	18
+SAM2 +閾値一定	4.00	5.12	14.73	14
+SAM2 +閾値調整	**3.15**	**4.29**	**12.84**	**10**

(1) 処理前　　(2) 処理後

図6:密度閾値処理前後の出力

5. 結果

実験の結果を表 2 に示す．SAM2 による背景除去を適用した場合，MAE は 3.65 人，RMSE は 4.87 人，MAPE は 16.62%，MaxE は 15 人となり，全ての指標で精度が向上した．特に MAPE は CLIP-EBC の 99.99%から SAM2 適用時の 16.62%へと約 83%減少し，背景誤検知の改善効果が確認された．MaxE も CLIP-EBC の 80 人から SAM2 適用時の 15 人へと減少し，誤差が改善された．

密度閾値処理のみを適用した場合，閾値一定では MAE 10.50 人, RMSE 15.42 人となり, CLIP-EBC の MAE 16.92 人と比較して 6.42 人，RMSE 23.35 人と比較して 7.93 人改善した．一方，閾値調整では MAE 4.78 人，RMSE 6.18 人となり，CLIP-EBC の MAE と比較して 12.14 人，RMSE は 9.24 人改善した．

SAM2 と密度閾値処理を組み合わせた場合，SAM2+閾値一定で MAE 4.00 人，RMSE 5.12 人，MAPE 14.73%，MaxE 14 人となった．SAM2+閾値調整では更に精度が向上し，MAE 3.15 人，RMSE 4.29 人，MAPE 12.84%，MaxE 10 人となり，全ての評価指標において最も低い誤差を示した．

6. 考察

提案手法による背景除去と密度閾値の結果を考察する．実験結果から，提案手法の各構成要素が都市公園環境での人数計測精度向上に異なる寄与をしていることがわかった．

SAM2 による背景除去は単独で CLIP-EBC MAE を CLIP-EBC の 16.92 人から 3.65 人へ削減し，全手法の中で最も大きな改善効果を示した．この大幅な改善は，SAM2 が画像ごとにセグメンテーションを行う特性によるものと考えられる．SAM2 は同一のポイントプロンプトを用いた場合でも，各画像の状況に応じて動的にセグメンテーションできる．この特性により，人物に対する検出精度を維持しつつ，樹木や建造物による誤検知を抑制したことが，精度向上の要因となったと推察される．

図 6 に密度閾値処理前後の出力例を示す．処理前では人が存在しない領域においてモデルの予測密度が不当に増加した．これに対して閾値処理後は密度の増加が抑制され，実際に人が存在する位置の密度だけが増加した．したがって閾値調整処理が精度改善に資するものだと判断した．

密度閾値処理については，一定値処理の MAE 10.50 人と比較して，画像ごとの調整の MAE 4.78 人が 5.72 人の改善を示した．密度閾値処理によって，低密度領域での誤検知除去は確認されたことから，この処理は特に閑散時の過大評価抑制に寄与したと考えられる．一定値では全ての画像に同じ閾値を適用するため，密度変化への対応が不十分であったのに対し，画像ごとの調整により各シーンの密度特性に応じた処理が可能となった．ただ，閾値一定の手法にも利点があり，事前に一つの閾値を設定するだけで運用できるため実装が簡便である．SAM と組み合わせた場合，閾値一定でも MAE 4.00 人，MAPE 14.73%の精度を達成しており，許容誤差が 15 人程度の用途では，運用の簡便性を優先して閾値一定を選択することも考えられる．

SAM2 と密度閾値調整の組み合わせにより，MAE 3.15 人という最高精度を達成した．SAM2 単独の MAE 3.65 人や密度閾値調整単独の MAE 4.78 人と比較して，改善が見られた．SAM2 は樹木や建造物に起因する誤検知を画像ごとのセグメンテーションすることで背景誤検知を抑制した．一方，密度閾値処理は SAM2 で除外されなかった領域内での低密度時の誤検知を画像ごとに抑制した．固定カメラ環境では背景要素は不変である一方，利用者数は時間帯により変動するため，静的な背景除去と動的な閾値調整を組み合わせることで，単独適用時よりも高い精度改善が得られる可能性が考えられる．

MAPE の変化を見ると，CLIP-EBC のみの 99.99%から最終的に 12.84%まで改善された．この大幅な改善は，特に利用者数が少ない環境での相対誤差が抑制されたことを示しており，時間帯による利用者数変動が大きい都市公園での実用性が向上したといえる．

本研究で採用した基盤モデル CLIP-EBC に SAM2 による背景除去と密度閾値の設定を追加し，MAE 3.15 人の精度を実現した．従来の教師あり学習手法では新しい環境への適用時に大量の教師データが必要となるのに対し，提案手法は事前の背景マスク設定と閾値調整のみで高精度な計測を可能にした．

7. まとめ

本研究では，基盤モデル CLIP-EBC を用いた都市公園利用者数の自動計測手法を考案した．基礎実験において，都市公園環境では樹木や建造物による背景誤検知と，閑散時と混雑時の利用者密度変化による精度変化が確認された．これらに対し，CLIP-EBC に加えて，SAM2 による背景除去と密度閾値処理を組み合わせた 3 段階処理手法を提案した．

6 地点での実験の結果，CLIP-EBC は MAE 16.92 人であったのに対し，SAM2 による背景除去で MAE 3.65 人，密度閾値調整で MAE 4.78 人となり，両処理を組み合わせることで MAE 3.15 人を達成した．MAPE も CLIP-EBC の 99.99%から両処理を組み合わせることで 12.84%へと改善され，特に低密度環境での過大評価の抑制が示唆された．MaxE も CLIP-EBC の 80 人から両処理を組み合わせることで 10 人へと減少し，極端な誤差の発生が抑制されたことが確認された．

固定カメラ環境の特性を活かし，SAM2 による事前の背景マスク設定と画像ごとの密度閾値調整を組み合わせることで，追加学習なしで高精度な計測を実現した．

今後の課題として，密度閾値の自動調整手法の検討が挙げられる．現在は画像ごとに手動で閾値を設定しているが，画像特徴から自動的に最適な閾値を決定する手法により，実用性の向上が期待される．また，天候変化や夜間など，より多様な環境条件での検証を進め，継続的な利用者数計測システムの構築を目指す．

謝辞　本研究においては，山口県土木建築部山口きらら博記念公園交流拠点化推進室様に，山口きらら博記念公園内での調査にご協力を賜りました．厚く御礼申し上げ，ここに感謝の意を表します．

参考文献

[1] 国土交通省都市局公園緑地・景観課：都市公園の柔軟な管理運営のあり方に関する検討会提言. 令和 4 年 10 月

[2] 国土交通省：都市公園利用実態調査について． 令和 4 年 9 月．

[3] Li et al.: "CSRNet: Dilated convolutional neural networks for understanding the highly congested scenes," CVPR 2018.

[4] Shi et al.: "CLIP-Count: Towards text-guided zero-shot object counting," ACM Multimedia 2023.

[5] Yiming et al.: "CLIP-EBC: CLIP Can Count Accurately through Enhanced Blockwise Classification," ,arXiv preprint arXiv:2403.09281.

[6] Alec et al.: "Learning Transferable Visual Models From Natural Language Supervision," ,arXiv preprint arXiv:2103.00020.

[7] Ravi et al.:. "Sam 2: Segment anything in images and videos. ",arXiv preprint arXiv:2408.00714, 2024b.

著者情報

岡野将大：八千代エンジニヤリング（株）技術創発研究所 AI 解析研究室 研究員．主に土木分野への AI 適用の研究に従事．

武田陸：八千代エンジニヤリング（株）九州支店 都市デザイン部 技術第一課 シニアアソシエイト．主に都市計画・まちづくりに関する計画策定業務に従事．

藤井純一郎：八千代エンジニヤリング（株）技術創発研究所 AI 解析研究室 室長．主に土木分野への AI 適用の研究に従事．