公園利用実態把握のための動画・音・BLEを用いたマルチモーダルセンシング

Multimodal Sensing Using Video, Audio, and BLE for Understanding Park Usage

寺岡 莉玖[1*] 細川 蓮[1] 松田 裕貴[2] 安本 慶一[1,3] 諏訪 博彦[1,3]

Riku Teraoka[1] Ren Hosokawa[1] Yuki Matsuda[2] Keichi Yasumoto[1,3] Hirohiko Suwa[1,3]

[1] 奈良先端科学技術大学院大学 [1] Nara Institute of Science and Technology
[2] 岡山大学 [2] Okayama University
[3] 理化学研究所 革新知能統合研究センター [3] RIKEN Center for Advanced Intelligence Project

要旨: 人口減少や人口流動の変化に伴い，地域に点在する公園の中には利用実態が十分に把握されていないものが多く存在する．このため，公園政策の立案や見直しにおいて，住民の満足度向上につながる意思決定を行うための科学的エビデンスが不足している．また，実環境における人手によるアノテーションは高コストであり，大量のラベル付きデータの収集は困難である．本研究では，動画・音・Bluetooth Low Energy（BLE）から得られる大量の未ラベルデータを活用した自己教師あり学習に基づくマルチモーダルセンシング手法を提案し，少量のラベル付きデータで公園利用者の有無，人数、行動（通過・滞留）を推定する．

1 はじめに

実社会における人流や人の行動を推定する技術は，都市計画や公共空間の管理，地域施策の検討において重要な役割を果たす．近年では，カメラを用いたビデオセンシングが広く利用されており，人物の有無や人数，行動といった詳細な情報を取得できる点で高い有効性を持つ．一方で，ビデオセンシングには画角の制約があり，遮蔽や設置位置，照度変化の影響を受けやすいという課題がある．また，映像データは情報量が大きく，長期間・広域での運用においては計算資源やストレージの観点から負担が大きい．

これまでの人流推定に関する研究では，主にビデオデータを用いた人数推定や行動認識が検討されてきた[1, 2]．しかし，実社会では環境変化（天候・照度・遮蔽）やセンサ設置制約により観測条件が変動しやすく，単一モダリティに依存した推定では頑健性に課題が残る．このため，複数のセンシングモダリティを組み合わせて弱点を補う，マルチモーダル人流推定が注目されている．

一方，Bluetooth Low Energy（BLE）や音センサは，視線方向に依存せず情報を取得でき，カメラの死角を補える可能性がある．しかし，BLE や音は得られる情報が間接的であり，単体で人流や行動を高精度に推定することは容易ではない．このため，マルチモーダル化により精度や頑健性の向上が期待されるが，実環境ではモダリティごとに観測の信頼性や情報量が大きく異なる．にもかかわらず，既存研究の多くは各モダリティを同等に扱う融合設計を採用しており，どのモダリティが，どのタスクに対して，どのように有効に働くかという役割分担の整理は十分ではない．言い換えると，モダリティの追加が常に一様な改善をもたらすとは限らず，タスクに応じて有効な使い方が異なる可能性がある．

そこで本研究では，実環境における役割分担に着目し，動画を主たる観測としつつ，音と BLE を補助情報として統合する枠組みを検討する．具体的には，5 分単位の時間窓ごとに，人の有無，総人数，滞留人数，通過人数を推定する枠組みを構築し，動画のみ，動画+音，動画+BLE，動画+音+BLE の条件比較を通じて，補助モダリティの寄与がタスクごとにどのように異なる形で現れるかを定量的に検証する．さらに，実環境での長期観測ではラベル付与が高コストであるため，本研究では未ラベルデータを活用できる自己教師あり学習を導入し，少量のラベル付きデータで下流タスクへ適用する構成を採用する．本稿では，公園環境で収集した動画・音・BLE データを用いた評価実験により，上

*連絡先：奈良先端科学技術大学院大学
〒 630-0192 奈良県生駒市高山町 8916 番地-5
E-mail: teraoka.riku.tt6@naist.ac.jp

記の枠組みの有効性と，モダリティごとに有効な役割が異なるという点を示す．

評価の結果，存在検知では動画+音が最良となり，人数推定では動画+BLE が最良となった．すなわち，音は検出漏れの抑制に，BLE は人数回帰の誤差低減に有効であり，補助モダリティの有効性はタスク依存であることが示された．

2 関連研究

2.1 人流推定における自己教師あり学習

人流推定（一定時間内の通過人数やユニーク人数，滞留数など）を実環境で高精度に行うには，長時間の動画・センサデータに対して詳細なアノテーションが必要となる．しかし，個体 ID 付与やフレーム単位の人数ラベル付けは高コストであり，学習データ不足が性能のボトルネックになりやすい．そこで近年，ラベルなしデータから汎用的な表現を獲得し，下流の推定タスクを少量のラベルで成立させる自己教師あり学習（SSL）が注目されている．

動画を用いた人流推定の文脈では，従来の静止画カウント（単一フレームの人数推定）に加えて，動画区間内で重複なくユニークな個体数を数える Video Individual Counting（VIC）が提案されている．Huang らは，VIC に対して自己教師あり学習を導入し，ラベルなし動画から群衆の動きに関する表現を学習する VIC-SSL を提案している [3]．人物領域の時系列的な移動を模擬するデータ拡張や，フレーム間の対応付けを促進する機構により，ID レベルの高価なラベルに依存せずに性能を向上させる点が特徴である．これは時間窓内の利用者数（延べ人数／ユニーク人数）」を推定する際に，ラベル削減と頑健性の両立を図る代表的アプローチと位置付けられる．

また，群衆カウントの分野では，ラベル付きデータを一切用いない完全自己教師あり学習の試みも報告されている．Sam らは，未ラベル画像群から自己教師ありに密度推定器を事前学習した後，推定値の分布が事前に仮定した群衆分布に一致するよう最適化することで，アノテーションなしで群衆カウントモデルを学習する枠組み（CSS-CCNN）を示している [4]．これらの研究は，人手ラベルが得にくい現場（屋外空間，長期観測）においても，未ラベルデータを活用して推定性能を引き上げうることを示唆している．

2.2 単一モダリティの人流推定

単一モダリティによる人流推定は，導入コストや運用の容易さの観点から依然として重要である．動画単独では，群衆カウント（フレームごとの人数）や VIC（動画区間内のユニーク人数）が主なタスクとして確立されており，高精度な視覚特徴に基づき人数推定が可能である．一方で，低照度，遮蔽，逆光，カメラ死角などの実環境要因により性能が低下しやすいという課題がある．

音単独による推定は，プライバシー制約の強い環境で有効な選択肢となる．例えば Hossain は，発話区間を除外した非発話の環境音のみを入力とし，Transformer ベースのモデルで混雑度（占有）を推定する枠組みを提案している [5]．音は視覚が破綻する状況でも取得しやすい反面，人数そのものよりも混雑度や活動量といった指標に落とし込みやすい点が特徴であり，推定対象（人数回帰か，混雑カテゴリ分類か）の設計が重要となる．

無線（Wi-Fi/BLE）単独の人流推定は，カメラを用いずに人の存在や流量を把握できる点で注目される．BLE に関しては，Goto らが 2 台の BLE スキャナを道路沿いに設置し，検出時刻差や信号強度差の時系列から人数に加えて移動方向まで推定する BLESS を提案している [6]．また，BLE 信号センシングに基づく人流・混雑推定の実用性を検証する取り組みとして，複数の都市環境を対象とした実践的研究 [7] や，飲食店・公共施設などの屋内空間における混雑度推定 [8]，固定路線バスにおける客観・主観混雑度の推定 [9] が報告されている．さらに，交通機関のような閉空間を対象とした研究では，Kanamitsu らが固定路線バスにおける混雑推定 [10]，Taya らが列車車両内の混雑推定 [11] を示しており，BLE がカメラなしで混雑状況を推定し得ることが報告されている．無線単独の利点はプライバシーと設置自由度である一方，端末保有率や MAC ランダム化，電波環境変動により推定が不安定になり得るため，特徴量設計や補正，および他モダリティによる補完が課題として残る．

2.3 マルチモーダルデータを用いた人流推定

単一モダリティの限界を補うため，複数モダリティを統合した人流推定が活発に検討されている．動画＋音では，夜間や遮蔽など視覚が劣化する極端条件で音が有効な補助情報となることが報告されている．Hu らは音と映像を同時に用いた群衆カウントを提案し，DISCO データセットを構築した上で，視聴覚融合が様々な条件下で性能改善に寄与することを示した [12]．さらに Sajid らは，音と映像特徴を Transformer で統合する Audio-Visual Transformer により，モダリティ間の関連を注意機構で学習し，群衆人数推定の性能を向上させることを示している [13]．これらは視覚が弱いときに音で補うという役割分担を明確に示す先行研究である．

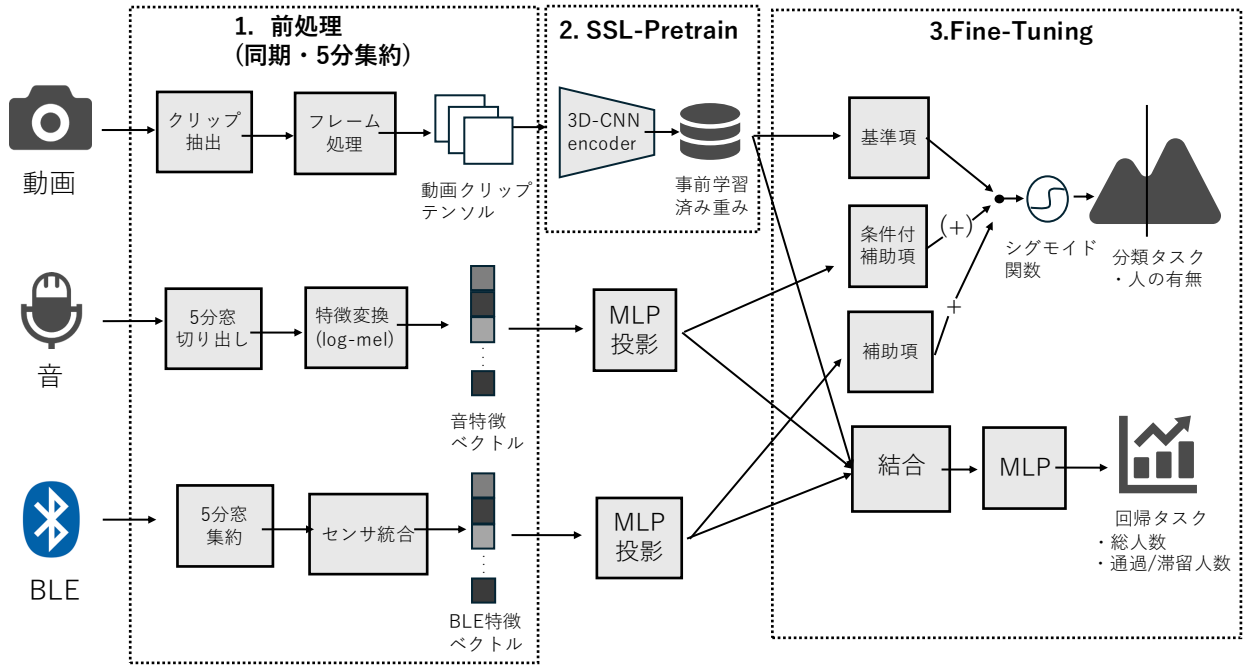


図 1: 提案手法の全体構成

動画＋無線（Wi-Fi/BLE）では，運用設計の観点で 2 つの方向性がある．第一に，短期間のみ動画で真値を取得して無線特徴との対応を学習し，その後は無線単独で長期運用する時間分離型のアプローチである．CountMeIn は，Wi-Fi プローブ等の統計量から人数を推定するために短期カメラ校正を行い，カメラ常設を避けつつ長期の人流推定を可能にする枠組みを示した [14]．第二に，同期取得した動画と Wi-Fi を同時に入力し，注意機構等で融合して精度を高める同時融合型であり，Hao らは Wi-Fi と映像を統合する群衆カウント手法を提案している [15]．前者は実運用（プライバシー・コスト）に強く，後者は推定精度を重視した設計として位置付けられる．

以上を踏まえると，公園のように環境変動が大きく，かつ長期観測が必要な場面では，(i) ラベル付けコストを抑える自己教師あり学習の導入，(ii) 動画・音・無線の相補性を活かしたマルチモーダル統合，(iii) 運用を見据えた校正・推定の設計，が重要な論点となる．

3 提案手法

本研究では，公園環境で同時計測した動画，音，BLE を用い，5 分単位の時間窓ごとに人の有無，総人数，滞留人数，通過人数を推定する．図 1 に全体構成を示す．処理は大きく，(i) 時間窓の整形と特徴化（前処理），(ii) 未ラベル動画を用いた動画エンコーダの自己教師あり事前学習，(iii) ラベル付き窓を用いた下流タスクへのファインチューニング，から構成される．

本研究では音・BLE を動画と同等に単純統合するのではなく，モダリティ特性に合わせて統合方法を設計した．予備検討として，全モダリティ埋め込みの単純結合を，人の有無推定に適用したところ学習が不安定となり，特に動画+音の条件で誤検出が増大した．そこで人の有無推定では，動画の判別が不確実な時間窓に限って音の寄与を大きくし，BLE は常時小さく寄与させる設計とした．本稿の比較はモダリティの単純な優劣ではなく，各モダリティをどのように用いると有効かという利用戦略の違いを評価するものである．

3.1 前処理

観測期間を 5 分単位の時間窓に分割し，各窓に対してラベル（人の有無，総人数，滞留人数，移動人数）を対応付ける．モダリティ間の対応付けはタイムスタンプに基づいて行い，実環境では欠損が生じ得ることを前提とする．欠損は後述のように，入力をゼロ埋めし，必要に応じて取得有無フラグをモデルへ与えることで扱う．

3.1.1 動画

各 5 分区間に対応する動画から短いクリップを抽出し，3D-CNN に入力可能なテンソルへ変換する．本実

装では，1 クリップあたりのフレーム数を 8 とし，空間解像度を 128 に統一する．フレーム抽出は，5 分区間内のある時刻を起点として，5 秒間隔で 8 枚を取得する（すなわち約 35 秒幅を 1 クリップとして表現する）．抽出したフレーム列はリサイズ等の変換を行い，$\mathbf{x}^{(v)} \in \mathbb{R}^{C\times T\times H\times W}$（$C=3, T=8, H=W=128$）の形式に整形する．また，計算効率のため抽出済みクリップをキャッシュし，再学習時の動画デコードを省略できるようにする．

3.1.2 音

各 5 分窓に対応する音声波形からログメルスペクトログラムを算出し，窓単位の音特徴ベクトルへ変換する．実装では，事前に窓ごとの音特徴ベクトル（ログメルから得た埋め込み）を バイナリファイルとして保存しておき，学習時に $\mathbf{x}^{(a)} \in \mathbb{R}^{d_a}$（$d_a = 64$）として読み込む．音が欠損している窓では $\mathbf{x}^{(a)} = \mathbf{0}$ とし，あわせて音の取得有無を表すフラグ $m^{(a)} \in \{0, 1\}$ を入力に付与する（$\tilde{\mathbf{x}}^{(a)} = [\mathbf{x}^{(a)}; m^{(a)}]$）．

3.1.3 BLE

BLE は複数センサで受信した広告パケットから，5 分窓ごとの特徴量を構成する．各センサについて，受信数，ユニーク端末数，RSSI 分布の統計量，時間変動量などを算出し，必要に応じて平滑化や正規化，および複数時間幅のローリング統計（15 分，30 分）を付加する．最終的に，窓特徴 $\mathbf{x}^{(b)} \in \mathbb{R}^{d_b}$ はセンサごとの特徴を結合して構成する．BLE が欠損する窓では $\mathbf{x}^{(b)} = \mathbf{0}$ として扱う．

3.2 自己教師あり事前学習

長期観測データの多くは未ラベルであるため，本研究では動画の 3D-CNN エンコーダを自己教師あり学習で事前学習する．実装では，同一クリップに対するデータ拡張（左右反転）から得られる 2 つのビューの表現が近くなるように最適化する対比学習を用いる．事前学習後，エンコーダ重みを保存し，下流タスクのファインチューニング時に初期値として読み込む．

3.3 ファインチューニング

各モダリティ入力から埋め込み表現を得た後，人の有無分類と人数回帰（総人数，滞留人数，移動人数）を同時に学習する．動画エンコーダは 3D-CNN（R3D-18）であり，クリップ $\mathbf{x}^{(v)}$ を入力して埋め込み $\mathbf{z}^{(v)} \in \mathbb{R}^{d_v}$（$d_v = 512$）を得る．BLE と音はそれぞれ MLP により埋め込みへ写像し，$\mathbf{z}^{(b)} \in \mathbb{R}^{d_{b'}}$，$\mathbf{z}^{(a)} \in \mathbb{R}^{d_{a'}}$ を得る．

3.3.1 人の有無分類

人の有無分類では，まず動画埋め込み $\mathbf{z}^{(v)}$ から，存在クラスに対する確率変換前の判別スコアを算出する：

$$s_v = g_v(\mathbf{z}^{(v)}).$$

次に，BLE と音を補助情報として判別スコアに加算するが，両者の寄与の与え方を分ける．

BLE は人の有無に関連する特徴から補助スコアを生成し，スケール係数 α_b を介して常時加算する．具体的には，BLE 埋め込み $\mathbf{z}^{(b)}_{\mathrm{pres}i}$ を MLP $g_b(\cdot)$ で写像し，

$$\Delta_b = \alpha_b \, g_b(\mathbf{z}^{(b)}_{\mathrm{pres}})$$

を得て，動画由来の判別スコアへ加える．

一方で音は，動画のみでは判断が不安定になりやすい時間窓において補助として働かせるため，動画の確率変換前の判別スコアの大きさ $|s_v|$ を，判断の確信度を表す代理量として用いる．すなわち，$|s_v|$ が小さいほど，すなわち境界付近で不確実なほど音の寄与を大きくし，$|s_v|$ が大きいほど，すなわち動画で十分に確信が高いほど音の寄与を抑えるように，重み

$$w_a = \sigma(k(\tau - |s_v|))$$

を導入する．ここで τ は不確実とみなす閾値，k は重みの遷移の鋭さを調整する係数である．さらに実装上，音の補助スコアが負方向に過度に作用しないよう，音埋め込み $\mathbf{z}^{(a)}$ を MLP $g_a(\cdot)$ で写像した出力に ReLU を適用し，

$$\Delta_a = \beta_a \, w_a \, \mathrm{ReLU}\Big(g_a(\mathbf{z}^{(a)})\Big)$$

として加算量を定義する（β_a はスケール係数）．

最終的な 人の有無の確率変換前の判別スコアは

$$s = s_v + \Delta_b + \Delta_a$$

で与えられ，s を確率へ写像した値（シグモイド関数 $\sigma(s)$）を人の有無の推定確率として用いる．この設計により，音は動画の判断が曖昧な時間窓でのみ補助情報として寄与しやすく，動画が明確な時間窓では動画の判断を維持したまま推定できる．

3.3.2 人数回帰

人数回帰では，各モダリティから得た埋め込み表現を結合（concat）し，回帰ヘッド（MLP）に入力する．動画のみの場合は $\mathbf{h} = \mathbf{z}^{(v)}$ とし，動画+音の場合は

$\mathbf{h} = [\mathbf{z}^{(v)}; \mathbf{z}^{(a)}]$，動画+BLE の場合は $\mathbf{h} = [\mathbf{z}^{(v)}; \mathbf{z}^{(b)}]$，動画+音+BLE の場合は $\mathbf{h} = [\mathbf{z}^{(v)}; \mathbf{z}^{(b)}; \mathbf{z}^{(a)}]$ として，

$$\hat{y}_{\text{total}} = f_{\text{total}}(\mathbf{h}), \quad \hat{y}_{\text{stay}} = f_{\text{stay}}(\mathbf{h}), \quad \hat{y}_{\text{pass}} = f_{\text{pass}}(\mathbf{h})$$

を出力する．

3.4 学習目的と評価指標

学習では，人の有無を二値分類（BCEWithLogits），人数（total，stay，pass）を回帰（L1）として同時最適化する．さらに本研究では，総人数と内訳の関係として

$$y_{\text{total}} \approx y_{\text{stay}} + y_{\text{pass}}$$

が成り立つことを踏まえ，整合性を促す補助項として

$$L_{\text{cons}} = |\hat{y}_{\text{total}} - (\hat{y}_{\text{stay}} + \hat{y}_{\text{pass}})|$$

を導入し，最終損失を

$$L = L_{\text{cls}} + L_{\text{total}} + L_{\text{stay}} + L_{\text{pass}} + \lambda L_{\text{cons}}$$

とする（λ は重み係数）．

評価では，人の有無に Precision，Recall，F1，Accuracy を用い，人数回帰に MAE，RMSE を用いる．加えて，複数出力間の整合性を確認する指標として，$|\hat{y}_{\text{total}} - (\hat{y}_{\text{stay}} + \hat{y}_{\text{pass}})|$ の平均（total–(stay+pass) 誤差）を併記する．

4 評価実験

4.1 実験環境

実験は，実際の公園環境において収集したデータを用いて行った．対象とする公園は，大阪府豊能郡豊能町にある光風台中央公園（光風台二丁目公園）であり，屋外環境として照度変化，植栽や遊具による遮蔽，利用状況の時間変動（時間帯・曜日・天候）といった要因が複合的に生じる．このような実環境条件下では，動画単独では見落としや誤検出が生じ得る一方，BLE や音は視覚的制約を受けにくいが情報が間接的であるため，モダリティ統合による補完効果の検証に適した環境である．

公園内には，カメラ，ボイスレコーダ，および BLE スキャナを設置し，長期間にわたってデータを取得した．センサの設置位置を図 2 に，設置した各センサの外観と構成を図 3 に示す．カメラ映像を保存したファイルのメタデータ上の解像度は 800×600，フレームレートは 25 fps，動画長は約 5 分である．音声はボイスレコーダにより MP3（192 kbps）で収録し，サンプリング周波数 44.1 kHz，低域カット（〜220 Hz）を有効化した．また，収録時の設定として入力レベル 90，出力レベル 4 メモリを用いた．BLE スキャナは Raspberry Pi 4 をベースに動作し，Bluetooth 4.0+EDR/LE Class1 対応 USB アダプタを使用して広告パケットを受信する．BLE は 15 秒間隔でスキャンし，取得したログを公園内の Wi-Fi 経由でクラウドに蓄積する．

本実験では，映像・音声の保存に対して追加の匿名化処理（顔ぼかしや音声の変換等）は実施していないが，取得・解析の設計において個人特定リスクを低減するよう留意した．具体的には，映像は 800×600 の解像度であり，個人識別に必要な細部情報が得られにくい条件で運用した．音声については，処理段階では波形そのものを用いず，ログメルスペクトログラム等の集約特徴量を用いて推定を行うことで，会話内容などの直接的情報を解析対象から外した．BLE については，カメラに比べて個人の外見情報を含まないという特性に加え，本研究では 5 分窓・複数センサの統計量（受信数や RSSI 分布等）として集約した特徴量を用い，個々の端末を追跡することを目的としない形で利用した．

本研究では，公園内でも利用が集中しやすいグラウンド領域を対象とし，グラウンド周辺に設置したセンサ 7，8，9，10 の 4 台のみを用いて特徴量を構成した．これは対象領域を明確化することで，アノテーションおよび推定結果との対応付けを容易にし，実運用を想定した評価を行うためである．

取得データの大部分は未ラベルであり，一部の時間区間についてのみ人手でアノテーションを行った．アノテーションはグラウンドを撮影するカメラ映像を観察し，5 分ごとの時間窓に対して総人数および行動内訳として通過人数と滞留人数を付与した．ここで滞留は，当該 5 分窓のうちおおよそ 2 分以上映像内に存在した人物を滞留として扱い，それ以外を通過として扱う．

本研究では，少量のアノテーションデータとして 2 日分のデータを用意し，うち 1 日分をファインチューニングに，もう 1 日分を評価に用いた．

4.2 比較手法および評価指標

提案手法の有効性を検証するため，評価実験では，動画を主軸とした補完効果が明確に比較できるよう，動画のみをベースラインとし，動画+音，動画+BLE，動画+音+BLE の 4 条件で性能比較を行った．これにより，動画に対する各補助モダリティの上乗せ効果と，両者を併用した場合の相補性を定量的に評価する．

評価指標として，人の有無分類には Precision，Recall，F1，Accuracy を用い，人数には MAE および RMSE を用いる．さらに本研究では，総人数 (total) と滞留人数 (stay)，通過人数 (pass) の間に total ≈ stay+pass という関係が成り立つことを踏まえ，複数出力間の整合性を確認する指標として「total–(stay+pass) 誤差」を併記する．

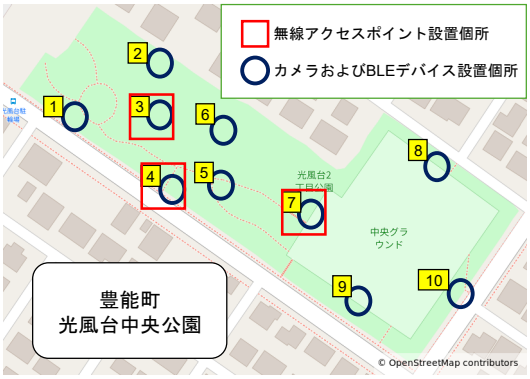
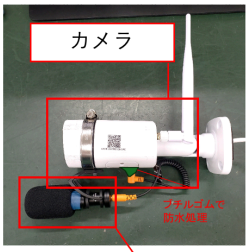
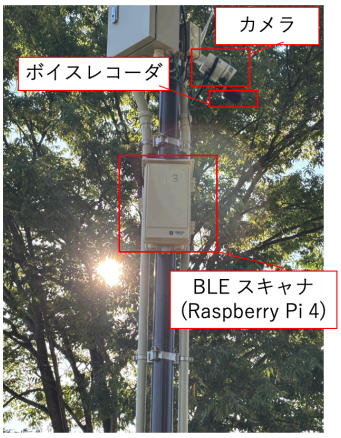


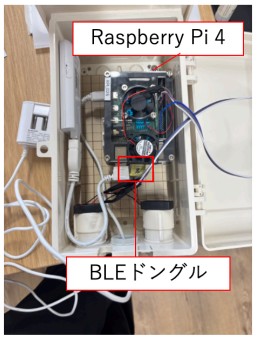
図 2: 光風台公園内に設置したカメラ・ボイスレコーダ・BLE スキャナの設置位置と撮影方向．赤枠内はグラウンドを表している．



(a) カメラとボイスレコーダの外観と構成



(c) 設置例



(b) BLE スキャナの外観と構成

図 3: 設置したセンサ概要

5　結果

本研究では，動画を主軸とした補完効果を明確に示すため，動画のみをベースラインとし，動画+音，動画+BLE，動画+音+BLE の 4 条件で比較を行った．表 1 に 人の有無推定の結果を，表 2，3，4，5 に人数推定（回帰）の結果を示す．

5.1　人の有無推定

人の有無推定では，動画+音が再現率（Recall）を高め，F1 および Accuracy において最良の性能を示した（表 1）．一方で，動画+BLE は適合率（Precision）が最も高く，誤検出を抑える方向に寄与する傾向が見られた．また，動画+音+BLE は Recall を高く保ちつつ F1 を改善するものの，動画+音を上回るには至らなかった．本研究の人の有無推定では，動画を基準としつつ補助モダリティの寄与を制御する統合を採用しているため，音と BLE は同一の役割で加算されるのではなく，タスクに応じて寄与の現れ方が異なる可能性がある．以上より，音と BLE はいずれも補助モダリティとして有効である一方，人の有無に対しては，音は主に検出漏れの抑制（Recall の改善）に，BLE は主に誤検出の抑制（Precision の改善）に寄与しやすいことが示唆される．

5.2　総人数推定

表 2 に，総人数の推定での各条件における MAE および RMSE を示す．

表 2 より，動画+BLE が MAE = 1.47，RMSE = 2.36 と最良の性能を示し，動画単独（MAE =2.11，RMSE =3.20）と比較して推定誤差が低減した．これは，グラウンド周辺に設置した BLE センサ（7，8，9，10）由来の端末検出数や RSSI 分布などの特徴が，動画由来の視覚特徴を補完し，人数推定の不確かさを低減した可能性を示す．一方で，動画+音は MAE =3.19，RMSE =5.07 と悪化しており，音の導入は人の有無推定には有効である一方，総人数の回帰に対しては必ずしも有効とは限らない．回帰では，各モダリティの埋め込み表現を結合して回帰ヘッドに入力する統合を採用しており，公園の屋外環境では風や交通騒音，遠方の活動音など，人数と直接対応しない要因により音が変動し得るため，単純な統合では回帰モデルにノイズとして作用する可能性がある．

5.3　通過・滞留人数推定

本研究では，5 分窓ごとに通過人数および滞留人数を推定する．ここで滞留は，当該 5 分窓のうちおおよそ 2 分以上対象領域に存在した人物として定義している．表 3,4 に，各条件における通過・滞留人数推定の MAE および RMSE を示す．

表 1: 人の有無推定の性能比較

Condition	Precision	Recall	F1-score	Accuracy
動画	0.737	0.767	0.752	0.745
動画+音	0.753	**0.836**	**0.792**	**0.779**
動画+BLE	**0.788**	0.712	0.748	0.759
動画+音+BLE	0.744	**0.836**	0.787	0.772

表 2: 総人数推定（回帰）の性能比較

Condition	MAE	RMSE
動画	2.11	3.20
動画+音	3.19	5.07
動画+BLE	**1.47**	**2.36**
動画+音+BLE	2.07	3.04

表 3: 滞留人数推定の性能比較

Condition	MAE	RMSE
動画	2.09	3.17
動画+音	2.91	4.73
動画+BLE	**1.72**	**2.59**
動画+音+BLE	2.34	3.38

表 4: 通過人数推定の性能比較

Condition	MAE	RMSE
動画	**0.55**	**0.92**
動画+音	0.63	0.94
動画+BLE	0.57	0.91
動画+音+BLE	0.62	0.97

滞留人数推定では，動画+BLE が MAE = 1.72，RMSE = 2.59 と最良となり，動画単独（MAE=2.09，RMSE=3.17）に対して誤差が低減した．これは，滞留のように窓内に留まる傾向を含む推定において，BLE が端末検出の継続性や RSSI 分布の変動として間接的な手掛かりを与え，動画を補完した可能性を示す．一方で，動画+音は MAE = 2.91，RMSE = 4.73 と悪化しており，屋外環境における環境雑音などの影響により，音情報がノイズとして作用し得ることが示唆される．

通過人数推定では，動画単独が MAE=0.55 と最良であり，RMSE においても 0.92 と最小であった．音および BLE の追加による改善は限定的であり，通過人数は滞留人数に比べて値域が小さいため，短時間の出入りが多い窓では観測・アノテーションのばらつきの影響も受けやすい．このため，補助モダリティを単純に追加しても，必ずしも誤差低減に繋がらない可能性がある．

さらに，本研究では総人数 (total) と内訳（滞留/通過 (stay/pass) 人数）の間に total ≈ stay+pass が成り立つことを踏まえ，複数出力間の整合性を確認する指標として total –(stay + pass) 誤差を併記した．表 5 より，動画+音+BLE は total - (stay + pass) 誤差が 0.13 と最小であり，内訳推定と総人数推定を同時に整合させる観点では有利となる可能性が示唆される．ただし，総人数や滞留人数の絶対誤差では動画+BLE が最良であるため，誤差最小化と整合性の確保は必ずしも一致しないことが分かる．

6 考察

6.1 人の存在検知における音・BLE の寄与

表 1 より，人の有無分類では動画+音が Recall = 0.836，F1 = 0.792，Accuracy = 0.779 と最良であり，音情報が検出漏れの抑制に寄与する傾向が確認された．本研究の人の有無推定では，音は動画の判別が不確実になりやすい時間窓において寄与が強まるように設計しており，動画単独で見落としが生じやすい状況を補う役割を担う．公園環境では，植栽や遊具による遮蔽，画角外への移動，逆光や照度変化などにより，動画単独では人物の存在を見落としやすい時間窓が生じ得る．このとき音は視線方向に依存せず，人の活動に伴う音響的変化を捉えられるため，存在の手掛かりとして働きやすいと考えられる．

一方で，動画+BLE は Precision=0.788 と最良であり，誤検出を抑える方向に寄与した．人の有無における BLE は，常時小さく寄与させる設計としており，動画が誤って人がいると判断しやすい窓に対して，過度に推定を上振れさせずに補助的な根拠を与える役割を担う．BLE は端末保有率や電波環境の影響を受けるものの，特定の時間窓で端末検出が継続している場合には，人が存在しないにもかかわらず動画が誤検出する状況を否定する補助信号となり得る．以上より，人の有無においては，音は主に Recall 改善（検出漏れ低減）に，BLE は主に Precision 改善（誤検出低減）に寄与しやすく，補助モダリティの有効性は同一タスク内でもどのように寄与させるかによって異なる形で現れることが示唆される．

表 5: 総人数と内訳の整合性（誤差）

Condition	total–(stay+pass) 誤差
動画	0.26
動画+音	0.41
動画+BLE	0.20
動画+音+BLE	**0.13**

6.2 総人数推定における音・BLE の寄与

表 2 より，総人数の回帰では動画+BLE が MAE=1.47，RMSE=2.36 と最良であり，動画単独（MAE = 2.11，RMSE = 3.20）から明確に改善した．本研究の人数回帰では，各モダリティの埋め込み表現を結合して回帰ヘッドに入力する統合を採用しており，BLE は人数と相関しやすい統計量（検出数や RSSI 分布等）を通じて動画特徴の不確かさを補完しやすい．また，本研究ではグラウンド周辺のセンサ 7，8，9，10 に限定して特徴量を構成しているため，推定対象領域と BLE 観測の対応が比較的取りやすく，回帰誤差の低減に繋がった可能性がある．

一方で，動画+音は MAE=3.19，RMSE=5.07 と悪化しており，総人数の回帰に対して音の導入が常に有効でないことが示された．公園の屋外環境では，風や交通騒音，遠方の活動音など，人数と直接対応しない要因が時間的に変動し得る．さらに複数方向からの音源が混在しやすく，グラウンド領域の人数と観測音の対応が弱い場合，音特徴が回帰モデルにとってノイズとして作用する可能性がある．したがって，音を人数回帰に活用するには，領域内に由来する音の抽出や，寄与を状況に応じて制御する統合設計（人の有無推定と同様の適応的寄与制御など）の導入が課題となる．

6.3 通過・滞留人数推定における音・BLE の寄与

表 3 より，滞留人数推定では動画+BLE が MAE=1.72，RMSE=2.59 と最良であり，動画単独（MAE = 2.09，RMSE = 3.17）を改善した．滞留は一定時間，同一領域に存在する性質を持つため，BLE のように観測が時間的に継続する信号は，滞留量の推定に対して補助情報として働きやすいと考えられる．一方で，動画+音は滞留で悪化（MAE = 2.91，RMSE = 4.73）しており，屋外雑音や領域外音源の影響により，滞留推定では音特徴が不利になった可能性がある．この点は，音が人の有無のような検出課題では有効でも，人数回帰（特に屋外）ではノイズとして作用し得るという，タスク依存性を示す結果と解釈できる．

通過人数推定では，動画単独が MAE = 0.55，RMSE = 0.92 と最良であり，音や BLE の追加による改善は限定的であった．通過人数は滞留人数に比べて値域が小さく，短時間の出入りが多い窓ではアノテーション境界や観測タイミングのずれの影響を受けやすい．このため，補助モダリティを単純に追加しても誤差が下がりにくく，むしろ特徴の増加が学習の難しさにつながる場合がある．したがって通過人数については，窓幅やラベル設計の見直し，あるいは時系列モデルによる滑らかな推定などが改善方向として考えられる．

加えて，本研究では総人数 (total) と内訳（滞留/通過 (stay/pass) 人数）の間に total ≈ stay+pass が成り立つことを踏まえ，total–(stay+pass) 誤差を併記した．表 5 より，動画+音+BLE はこの誤差が 0.13 と最小であり，総人数推定と内訳推定の整合性という観点では効果が強く現れた．一方で，総人数や滞留人数の絶対誤差では動画+BLE に及ばないため，単純な誤差最小化と内訳を含む整合性の確保は必ずしも一致しない．この結果は，複数出力を同時に扱う設定では，整合性をどの程度重視するかに応じて，統合設計や学習目標（損失の重み付け等）を分離して検討する必要があることを示している．

7 まとめ

本研究では，実環境の公園（光風台中央公園）グラウンド領域を対象に，動画を主軸として音と BLE を補助的に統合し，5 分窓ごとに人の有無，総人数，滞留人数，通過人数を推定した．

4 条件で比較した結果，人の有無推定では動画+音が最良となり，音情報が検出漏れ低減に寄与することが確認された．一方，人数回帰では動画+BLE が総人数および滞留人数で最良となり，BLE が人数推定誤差の低減に有効であることが示された．移動人数推定は動画単独が最良であり，短時間事象の推定は補助モダリティ追加のみでは改善しにくいことが分かった．また，total–(stay+pass) 誤差は 動画+音+BLE が最小となり，出力間の整合性という観点では複数モダリティ併用が有利となる可能性が示唆された．

以上より，各モダリティの有効性は一様ではなく，タスク特性に応じてどのモダリティを，どのように寄与させるかが重要である．今後は，通過人数に対する窓幅・ラベル設計の見直しに加え，人数回帰に対しても寄与を状況に応じて制御する統合設計や，総人数と内訳の整合性を学習に組み込む設計の検討が課題である．

謝辞

本研究は，JST さきがけ（JPMJPR2465）および JST 共創の場形成支援プログラム JPMJPF2115 の支援を受けたものです.

参考文献

[1] Guangshuai Gao, Junyu Gao, Qingjie Liu, Qi Wang, and Yunhong Wang. Cnn-based density estimation and crowd counting: A survey. *arXiv*, 2020.

[2] Preksha Pareek and Ankit Thakkar. A survey on video-based human action recognition: recent updates, datasets, challenges, and applications. *Artificial Intelligence Review*, Vol. 54, pp. 2259–2322, 2021. Published: 25 Sep 2020.

[3] Feng-Kai Huang, Bo-Lun Huang, Li-Wu Tsao, Jiun-Cheng Wu, Hong-Han Shuai, and Wen-Huang Cheng. Flowing crowd to count flows: A self-supervised framework for video individual counting. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25)*, pp. 8234–8243, 2025.

[4] Deepak Babu Sam, Abhinav Agarwalla, Jimmy Joseph, Vishwanath A. Sindagi, R. Venkatesh Babu, and Vishal M. Patel. Completely self-supervised crowd counting via distribution matching. In *Computer Vision – ECCV 2022 (LNCS 13691)*, pp. 186–204. Springer, 2022.

[5] Forsad Al Hossain, M. Tanjid Hasan Tonmoy, Andrew A. Lover, George A. Corey, Mohammad Arif Ul Alam, and Tauhidur Rahman. Crowdotic: A privacy-preserving hospital waiting room crowd density estimation with non-speech audio. In *Proceedings of the 25th International Workshop on Mobile Computing Systems and Applications (HotMobile '24)*, pp. 79–85. ACM, 2024.

[6] Ippei Goto, Kentaro Ueda, Yuki Matsuda, Hirohiko Suwa, and Keiichi Yasumoto. Bless: Ble based street sensing for people counting and flow direction estimation. In *IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops 2024)*, pp. 76–81, 2024.

[7] Yuki Matsuda, Hirohiko Suwa, Kotaro Hayashi, Taito Yoshimura, Arata Yoshihara, and Ismail Arai. Estimating people flow and crowdedness for various urban environments based on ble signal sensing: Practical studies. *IEICE Transactions on Communications*, pp. 1–11, 2025.

[8] Yuki Matsuda, Kentaro Ueda, Eigo Taya, Hirohiko Suwa, and Keiichi Yasumoto. BLECE: BLE-based crowdedness estimation method for restaurants and public facilities. In *Fourteenth International Conference on Mobile Computing and Ubiquitous Network, ICMU 2023, Kyoto, Japan, November 29 - Dec. 1, 2023*, pp. 1–6. IEEE, 2023.

[9] Takumi Ikenaga, Yuki Matsuda, Ippei Goto, Kentaro Ueda, Hirohiko Suwa, and Keiichi Yasumoto. Using BLE signals to estimate objective and subjective crowdedness levels on fixed-route buses. *IEEE Access*, Vol. 13, pp. 67488–67499, 2025.

[10] Yuji Kanamitsu, Eigo Taya, Koki Tachibana, Yugo Nakamura, Yuki Matsuda, Hirohiko Suwa, and Keiichi Yasumoto. Estimating congestion in a fixed-route bus by using ble signals. *Sensors*, Vol. 22, No. 3, pp. 1–15, 2022.

[11] Eigo Taya, Yuji Kanamitsu, Koki Tachibana, Yugo Nakamura, Yuki Matsuda, Hirohiko Suwa, and Keiichi Yasumoto. Estimating congestion in train cars by using ble signals. In *The 2nd Workshop on Data-Driven and Intelligent Cyber-Physical Systems for Smart Cities (DI-CPS '22)*, pp. 1–7, 2022.

[12] Di Hu, Lichao Mou, Qingzhong Wang, Junyu Gao, Yuansheng Hua, Dejing Dou, and Xiao Xiang Zhu. Ambient sound helps: Audiovisual crowd counting in extreme conditions, 2020. Introduces the DISCO (auDiovISual Crowd cOunting) dataset.

[13] Usman Sajid, Xiaoxu Chen, Hassan Sajid, Tae Kim, and Guanghui Wang. Audio-visual transformer based crowd counting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 2249–2259, 2021.

[14] Gürkan Solmaz, Pankaj Baranwal, and Flavio Cirillo. Countmein: Adaptive crowd estimation with wi-fi in smart cities. In *IEEE International Conference on Pervasive Computing and Communications (PerCom 2022)*, pp. 187–196. IEEE, 2022.

[15] Lifei Hao, Baoqi Huang, Bing Jia, and Guoqiang Mao. Heterogeneous dual-attentional network for wifi and video-fused multi-modal crowd counting.

IEEE Transactions on Mobile Computing, Vol. 23, No. 12, pp. 14233–14247, 2024.