CF-002

画像と音声のマルチモーダル学習による混雑予測
Crowd Prediction by Multimodal Learning of Audiovisual Data

井手　伊織[1)]　豊坂　祐樹[1)]　成　凱[1)]
Iori Ide　Yuki Toyosaka　Kai Cheng

1　はじめに

新型コロナウイルスによるパンデミックが収束し，日常生活が徐々に元通りになっている中で観光やイベント開催が増え，人々の移動が活発化し，駅や観光地，商業施設などで発生するオーバーツーリズムや雑踏事故が問題視されている．例えば,2022 年に韓国で発生したソウル梨泰院雑踏事故で梨泰院通りで 10 万人以上が道幅は 3〜4 メートルと狭い坂道を訪れたことで圧迫や転倒などで 159 人が亡くなったとされている．こうした事故を防ぐために，人の移動によって混雑が発生しやすい場所において混雑検知を行い，人々の存在や行動を把握することが重要である．しかし，従来のシステムではカメラ画像内のみでの混雑検知を行っており，光や遮蔽物の影響により，精度が落ちることによって屋外への汎用性や緊急時における可用性の懸念点などが存在する [2].

本研究ではカメラ画像に加え，騒音や音響特徴量といった音声データを活用し，周辺の人数に合わせて変動するデータを元に空間内における混雑検知及び予測するシステムを提案する．音声データを用いることにより，カメラ画像だけでは捉えきれない特徴を補完することができ，画像に異常があったとしてもより正確に群衆カウントを行うことができると期待される．さらに, 混雑状況は時間とともに変化するものであり, 変化の傾向を把握することで混雑予測につながることもできる.

2　問題提起

従来の研究では，主に画像データに基づいた畳み込みニューラルネットワーク（CNN）を用いた手法が主流である．これらの手法は，画像情報を活用して群衆の密度や人数を推定するものであり，多くの研究によって高い精度が示されてきた [1]．しかし，画像データに依存する手法には特に，光の変化や遮蔽物の存在によって視覚情報が妨げられる環境では，認識精度が低下しやすいという問題が挙げられる [2]．例えば，屋外での夜間監視や，建物やその他の障害物によって視界が遮られる場合，CNN ベースの手法は十分な性能を発揮することが難しくなる.

こうした課題を解決するために，近年では音声データを補助情報として活用するためにマルチモーダル学習が研究されている [2]．音声データは，光や遮蔽物の影響を受けず，人の話し声や足音などの環境音を解析することで，視覚情報では捉えきれない情報から混雑度の推定することが可能である．また，マルチモーダル学習では，音声データと画像データを統合し，学習を行うことで，より高精度な混雑検知モデルを構築することが可能である．この手法により，夜間や遮蔽物が存在する環境でも，混雑状況を正確に把握することが可能となる.

しかし, 既存の研究では, 個別の混雑シーンを検出すること, いわゆる混雑検知を目的としおり, 混雑状況の発生前からの混雑予測を行うものではない. 混雑が発生する前から人の集まり具合の変化から混雑を予測することが, 事故発生の未然防止に重要である. 本研究では，図 1 に示すように, **カメラ画像**の解析による群衆カウントに加え, **環境音**の解析による集まり状況の変化を合わせて, **時間**の推移に伴いに変化する傾向を学習し, より高い精度の混雑予測を目指す. 光や遮蔽物の影響を受けない音声データを補助情報として用いることで人夜間や遮蔽物の影響があってもリアルタイムで混雑状況をより正確に捉えるモデルを作成し，そこから変動する人数推移から直近の人数を予測することで雑踏事故を未然に防ぐことに寄与する.

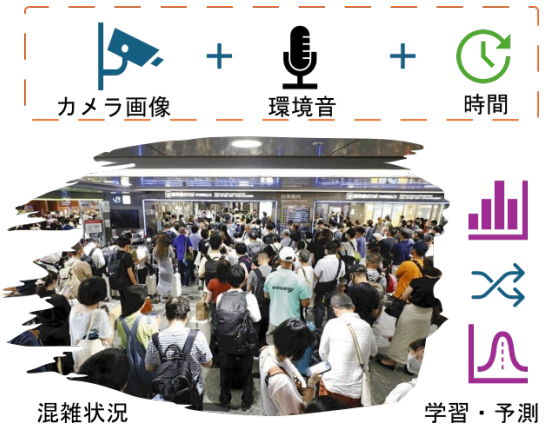

図 1: 提案システムのイメージ

3　混雑検知の関連技術

3.1　画像による混雑検知

混雑検知に関する研究は, 最初から主に画像を用いた人数カウントや密度推定を行うものであった. 主な手法としては, 物体検出に基づく群衆カウント (検出ベース,Detection-based Crowd Counting) と, 密度推定による群衆カウント (密度推定モデル,Density-based Crowd Estimation) に大別できる. 初期のアプローチでは HOG（Histogram of Oriented Gradients）などの手作業で特徴量を設定し，人の頭や顔を検出する手法が取られていた. 近年では，CNN ベースでの深層学習を用いた高精度な手法があり，YOLO（You Only Look Once）や SSD（Single Shot MultiBox Detector）などの検出ベースによる物体検出アルゴリズムは，リアルタイムでの検出に優れている. また，CSRNet [1] などの密度推定モデルは, 群集の画像から密集度をヒートマップとして表現した群集密度マップ（density map）を作成し, ピクセル単位で数の計測を行うモデルである.CSRNet は特徴量抽出に特化したフロントエンドの CNN と, プーリング演算の役割を担うバックエンドの CNN を組み合わせたネットワーク構造が特徴であり, 検出精度を大幅に改善できると知られている.

1)　九州産業大学, Kyushu Sangyo University

3.2 音声による混雑検知

音声データを用いた混雑検知は道路交通における車の騒音と交通量との関係性を測定する際に用いられている [4] [5]. 騒音の特性上，不規則かつ大幅に dB が変動するため瞬間の dB に左右されないように等価騒音レベル（E quivalent continuous A weighted sound pressure L evel: LAeq）の計算式によって評価されることが多い. 等価騒音レベルとは，一定の時間内に区切ることでその中での変動する騒音レベルのエネルギーに着目して時間平均値を算出する手法である. この計算式を用いることで瞬間の dB のピークに左右されずに平均的な交通量などを測定することが可能である.

[6] ではスマートフォンのマイクを使用して周囲の音声データを収集し，その音声データを解析することで，特定のエリアの混雑度を推定する手法が研究されており，1 分間の環境音データに対して FFT（高速フーリエ変換）を適用し，周波数ごとの振幅スペクトルを求め，0〜2000Hz の周波数帯域のスペクトル値の総和を特徴量を K 近傍法によって低・中・高に定義した混雑度に分類を行っている．また，スペクトログラムと呼ばれる縦軸が周波数 (Hz)，横軸が時間，色の濃度が音圧 (dB) を可視化した図形による分析では 0〜2000Hz 帯が群衆密度によって振幅スペクトルが増加しており，混雑度によって大きな差が現れていることが示されている．このことから音声データを用いた混雑検知は有効性があると考える.

3.3 マルチモーダル学習による混雑検知

マルチモーダル学習とは，画像や音声，センサーデータなどの異なる単一のモーダリティのデータを統合して学習する手法で，異なるモーダリティ同士を用いて学習することで今まで単一のモーダリティから捉えることができなかった複雑な特徴を見つけ出すことが可能である．画像と音声によるマルチモーダル学習の研究はいくつかあり，[3] では騒音環境下でも唇の動きの画像を用いることで音声認識の精度を向上させた．また，[2] では AudioVisualCrowdCounting モデルと呼ばれる密度ベースモデルで生成した密度マップに音声データを可視化したものを CNN を用いて特徴量化し，両方の特徴量を融合するモデルを用いて，画像のみのモデルよりも精度の高い混雑検知で遮蔽物や光の影響を軽減することが可能となった．このことから，画像と音声を用いることで有効性があることから応用として人数推定から人数予測が行えるモデルの作成を本研究では目指す.

4 提案手法

本研究では, カメラ画像に加え，音声データも活用し, 騒音や音響特徴量により, 周辺の人数に合わせて変動するデータを元に空間内における混雑検知及び予測するシステムを提案する．カメラ画像だけでは捉えきれない特徴を環境音の変化といった音声特徴により補完し，画像に異常があったとしてもより正確に群衆カウントを行う．さらに, 混雑状況の変化傾向を把握することで混雑予測する.

4.1 混雑検知に利用可能な特徴量

本研究では，3.3 で紹介した手法を元に画像特徴量として密度ベースモデルから得た特徴量と，音声データの有効性を示すために人数と周辺の音響の相対的な関係性の分析を行う．分析する際は騒音が時間とともに不規則かつ大幅に変化しているため，ある時間内で変動する騒音レベルのエネルギーに着目して時間平均値を算出する等価騒音レベル (LAeq) を用いることで極端に発生するピーク値に左右されずに評価することができ，以下に数式を示す.

$$L_{Aeq} = 10\log_{10}\left(\frac{1}{N}\sum_{i=1}^{N} 10^{\frac{SPL_i}{10}}\right)$$

ここで，$SPLi$ は各フレームの音圧レベル，N はフレーム数，p は測定された音圧 (パスカル)，$p0$ は基準音圧（20 μ Pa）を表す．また，SPL は以下の式で表される.

$$SPL = 20\log_{10}\left(\frac{p}{p_0}\right)$$

さらに時系列データを付与することで RNN による学習により，予測が行えるように進めていく．画像特徴量では入力画像である群衆の写ったものをビデオカメラなどから収集を行う．収集の際は駅や祭りなど様々なシュチュエーションから収集することで画像特徴量と音響特徴量による有効性を示す．有効性の検証が示された場合は時刻データに基づき予測を行う.

4.2 画像特徴量と音声特徴量の融合

画像特徴量と音声特徴量を融合には様々方法があり，初期の融合方法として特徴を初期段階で結合し，一つの統合された特徴ベクトルとして扱い，学習を行う方法である初期融合（Early Fusion）や CNN などの学習モデルを通した後に融合を行う後期融合 (Late Fusion) などがある．最近の手法ではニューラルネットワークを通し，特徴量を抽出した後に融合を行う中間融合 (Intermediate Fusion) などが取られており，[2] で紹介した研究ではこの手法がとられている．特徴量の融合を行う場合の計算としては，ベクトルの乗法や加法を行うことである.

4.3 時間経過に伴う人数推移に基づく混雑予測

時間経過に伴う人数の推移から人数予測を行うために，リカレントニューラルネットワーク (RNN) や長短期記憶 (LSTM）ネットワークによる学習を行う手法を用いて予測を行っていく．RNN は，系列データを扱うためのニューラルネットワークアーキテクチャであり，入力データの時間的依存関係を捉えることができ，過去の情報を保持しつつ，それを用いて現在から次へと入力を処理を行い，学習を行なっていく手法である．しかし，RNN は短期的な依存関係は捉えやすいものの，学習途中に発生する勾配消失問題によって長期的な依存関係を捉えるのが難しいという問題がある．この問題を解決するために提案されたのが，LSTM である．LSTM は，情報を長期間にわたって保持し，必要に応じて情報を更新・削除することができる．LSTM の基本構造は，セル状態（Cell State）と input gate，forget gate，output gate の３つのゲートからなる．これにより，勾配消失を抑えることができ，関係性のある情報を長期間保持し，不必要な情報を削除する学習が可能となる．これらの学習を画

像と音声の特徴量融合を行なった後に LSTM を通して時系列データを用いて学習を行うことで予測を行うことを提案する．

5　予備実験

5.1　実験データ

実験を行うにあたって九州産業大学 1 号館の公共空間で収集を行った．使用したデバイスとして一般的なビデオカメラである SONY FDR-AX45: 1 台を使用し，音声は 2 チャンネルステレオ，解像度は 1280 × 720 で撮影を行い，動画の保存で広く用いられる mp4 ファイル形式での保存を行なった．このファイル形式からさらに画像および音声データに分け，時系列情報を付与しながらデータセットの作成を行う．

データの内容として，授業実施日で 11:35 から 12:50 分程の時間帯で昼食時のピーク時に混雑状況が最大になる期間に収集を行った．このデータを元に特徴量の抽出及び評価を行う．

- **撮影機材**: SONY FDR-AX45 1 台
- **撮影場所**: 九州産業大学 1 号館 1 階フロア
- **撮影時間**: 授業実施日正午の時間帯 1 時間 12 分
- **解像度**: 1280 × 720
- **マイク**: ECM-XYST1M 2 チャンネルステレオ
- **ファイル形式**: mp4 ファイル

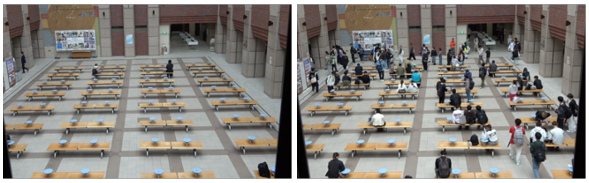

図 2: 学生休憩スペースにおける混雑状況変化

5.2　音響特徴量と人数推移の関係性評価

音響特徴量と実際の人数推移の関係性を評価するにあたって等価騒音レベルでの評価を行った．手順は以下の通りである

1. 動画データから mp3 形式のファイルに変換
2. 5 分間毎の騒音レベルの平均を計算する

表 1: 5 分間の等価騒音レベルと平均人数

時間 (分)	LAeq	平均人数	備考
0-5	69.58	4	2 限目授業実施中
5-10	66.97	2	2 限目授業実施中
10-15	70.64	2	2 限目授業実施中
15-20	70.89	7	2 限目授業実施中
20-25	65.41	4	2 限目授業実施中
25-30	66.76	13	2 限目授業実施中
30-35	70.27	8	2 限目授業実施中
35-40	71.10	9	2 限目授業実施中
40-45	68.25	10	2 限目授業実施中
45-50	69.20	9	2 限目授業終了
50-55	71.84	27	昼休み
55-60	74.64	57	昼休み
60-65	74.91	75	昼休み
65-70	74.03	94	昼休み

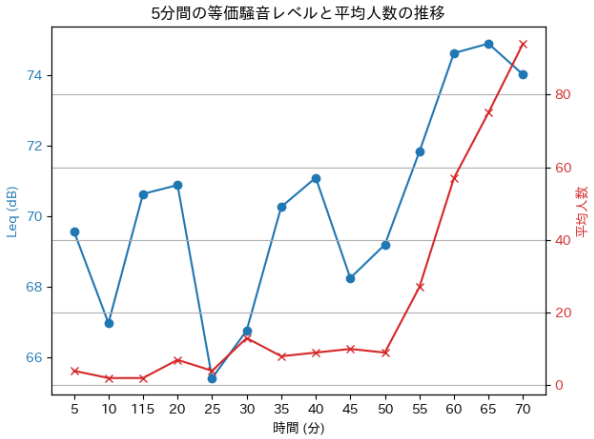

図 3: 等価騒音レベルと平均人数の変化の推移

表 1 の結果として等価騒音レベルと平均人数の変化からある程度の相関性があることがわかる．図 3 のグラフからも読み取れるように後半の人数の増加に比例して等価騒音レベルも増加していることが読み取れる．しかし，途中で人数が少ないにも関わらず増加している部分として，後ろの音声やノイズを拾ってしまった可能性があるためさらに分析を行う必要性がある．

また，音声データを FFT で周波数に分解し，スペクトログラムで可視化を行なった．Log-Mel スペクトログラムとは，高周波の音は鈍く，低周波の音は明確に聞き取ることができる人間の聴覚特性に基づいたメル尺度をスペクトログラム反映したものである．今回は Python の librosa ライブラリを用いて加工を行なった．縦軸は周波数，横軸は時間，色の濃淡派各周波数成分の dB の強さを表す．

1. オーディオファイルを読み込む
2. 音声データを 1 秒間毎に区切る
3. オーディオデータをフレームに分割する
4. 各フレームに対して高速フーリエ変換（FFT）を実行
5. パワースペクトルを計算する
6. メルフィルタバンクを適用する
7. 対数を取ってメルスペクトログラムを生成する

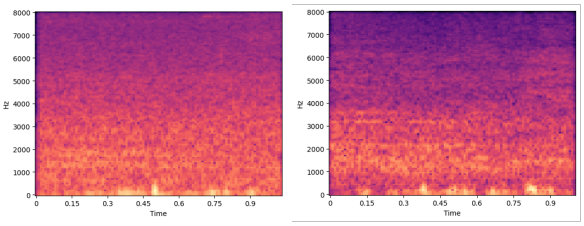

図 4: Log-Mel スペクトログラム
左: 人数: 小 右: 人数: 大

結果として人数が少ない時は 4000〜8000Hz 帯の色の濃淡は薄くグラデーションが緩やかであるが，人数が多い時は低周波数帯と高周波数帯の色がはっきりと分かれていることが確認できる．これはメル尺度によって低周波数帯の特徴を強調していることがわかることから特徴量として有効性があると考察する．また，これらのスペ

クトログラムの画像を CNN を用いて特徴量の抽出を行うことで特徴量化することが可能である．

5.3 群衆カウントモデルの比較

提案手法の有効性を示すために，検出ベースモデルと密度ベースモデルで人数推定精度の比較を行った．検出ベースモデルでは比較的最新モデルである Yolov8x，密度ベースモデルでは特に高密度の群衆シーンにおいて優れたカウント精度である CSRNet を用いた．結果として正解として 461 人であり，目視でカウントした際も 460 人程度であった．

図 5: Yolov8x を用いた検出ベースの人数カウント
左: 元画像 右: 検出結果

YOLOv8x はパラメータの調整を行い，信頼度 (バウンディングボックスで囲った部分が，「背景」か「物体」かを表す数字) は 0.001，基準の枠とその他の枠の重なり具合 (IoU) は 0.3，classes の選択は人のみ，最大検出数は 1000 件になるように調整を行なった. 調整理由としてデフォルト値での検出はされなかったため手作業である程度検出が可能になるように調整を行なった．

検出ベースモデルは 554 人，CSRNet は生成した密度マップから 484 人と推定された．この結果から検出ベースモデルは重なった群衆に対して精度が低いことと随時パラメータの調整が必要であることから運用が難しい．また，今回は人の体全体での検出を行ったが，対象を頭部に絞った検出などもあり，今後の検証が必要である．

一方, 密度ベースモデルでは群衆画像に対して高い精度であることがわかった．これは，密度ベースモデルが人が重なった場面でも正確に人をカウントすることができ，頭部や体全体の検出とは違ってピクセル内に人のいる割合を学習を元に割り当てることで群衆に対して高い精度で推定することが可能であると考察する．

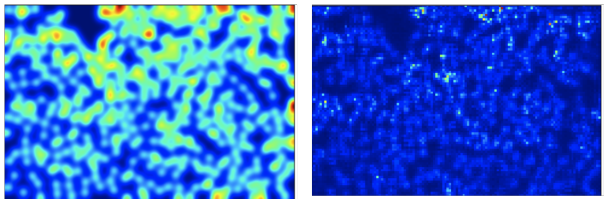

図 6: CSRNet を用いた密度ベースの人数推定
左: 正解データ 右: 予測データ

6 終わりに

本研究では，画像と音声のマルチモーダル学習を用いた混雑予測手法を提案を行なった．予備実験の結果，動画データの収集方法の説明を行い，収集した動画データから人数推移と音響特徴量の相関関係の有効性を示した．また，群衆カウントモデルの比較を行い，検出ベースモデルと密度ベースモデルの精度の比較を行なった．結果として密度ベースモデルの精度がよく，今後の画像特徴量の抽出のために有効であることがわかった．今後の研究では，夜間の祭りや遮蔽物のある公共空間でのデータ収集およびデータセットの作成を行い，画像と音声のマルチモーダル学習から人数予測が行えるように進めていく予定である．

参考文献

[1] Yuhong Li, Xiaofan Zhang, Deming Chen,”CSRNet: Dilated Convolutional Neural Networks for Understanding the Highly Congested Scenes”, arXiv:1802.10062v4 [cs.CV] 11 Apr 2018, 2018.

[2] Di Hu, Lichao Mou, Qingzhong Wang, Junyu Gao, Yuansheng Hua, Dejing Dou, Xiao Xiang Zhu,”Ambient Sound Helps: Audiovisual Crowd Counting in Extreme Conditions”,arXiv:2005.07097v2 [cs.CV] 16 May 2020, 2020.

[3] Weijiang Feng,Naiyang Guan,Zhigang Luo,”Audio visual speech recognition with multimodal recurrent neural networks”, 2017 International Joint Conference on Neural Networks (IJCNN)pp. 681-688, 2017.

[4] 渡辺 義則, 寺町 賢一, 牧田 晃介, ”騒音環境基準を遵守可能な幹線道路の交通量の簡易計算法について.”, 土木計画学研究・論文集 Vol.26 no.5 pp. 837-846, 2009.

[5] 仲 功, 野呂 雄一, 久野 和宏，交通量と時間率騒音レベルの関係 : Q-V 曲線を用いた検討，日本音響学会誌 56 巻 (2000) 5 号 pp. 301-307， 2000.

[6] 西村 友洋，樋口 雄大，山口 弘純，東野 輝夫．”スマートフォンを活用した屋内環境における混雑センシング”， マルチメディア, 分散, 強調とモバイル (DICOMO2013) シンポジウム pp. 1310-1321, 2013.