H-001

深層学習を用いた混雑状況の判定
Determination of Crowded Situation using Deep Learning

加藤　直樹†　　瀬川　修†
Naoki Kato　　Osamu Segawa

1. はじめに

展示会場や施設などにおける混雑状況を自動的に把握し，来場者の誘導や運営者によるデータ活用（安全管理やマーケティングなど）のための判定技術のニーズが高まっている．これらを実現する要素技術として機械学習的な手法が盛んに研究されており，米司ら [1] は動画から抽出した細かい動き情報をクラスタリングすることで人物ごとの位置，移動方向，速度などを取得し，人流解析により混雑状況を可視化している．

このように従来手法では，時系列情報を持つ「動画」を用いたアプローチが主流となっている．一方，人間の感覚では 1 枚の「静止画」を見るだけで，明示的に人数をカウントすることなく混雑状況の判定が可能である．人間は混雑という境界が曖昧な概念に対する判断基準を持っていると考えられる．

そこで，本研究では「静止画」を対象とし，この人間の感覚（判断基準）を深層学習により模倣する判定モデルの獲得を試みる．さらに，ニューラルネットワークのモデルが画像中の何に注目して判断をしているかを可視化し，学習した判定モデルについて考察を行う．

2. 深層学習を用いた混雑状況の判定

2.1 画像セットの準備

判定モデルの学習および評価のため，以下に示す手順で画像セットを準備した．

(1) 画像収集と人手によるラベリング　筆者らが所属する中部電力株式会社で開催された「テクノフェア 2016」（技術開発本部を一般公開するイベント．2016/10/20-21 開催．来場者数約 3,200 人．以下「社内展示会」）において，実験設備やフロアなどの 9 会場を撮影した動画を，1 秒間隔でキャプチャした．これに 1 名の作業者が表 1 に従って混雑状況をラベリングし，Lv1（空いている）約 5,800 枚，Lv2（中程度）約 20,900 枚，Lv3（混雑している）約 2,200 枚の画像を収集した．混雑状況ごとの画像数に偏りがあるため，各混雑状況 2,000 枚ずつ 6,000 枚の画像をランダムに抽出し「画像セット 1」とした．

これとは別に，Google 画像検索*において「展示会，混雑」などの検索クエリを用いて，商業イベントなどの展示会（以下「社外展示会」）の画像を取得した．これに同様の条件で混雑状況をラベリングし，各混雑状況約 100 枚ずつ 300 枚の画像を収集した．この中から各混雑状況 40 枚ずつ 120 枚の画像をランダムに抽出し，画像セット 1 と合わせた 6,120 枚の画像を「画像セット 2」とした．残りの 180 枚の画像から各混雑状況 50 枚ずつ 150 枚の画像をランダムに抽出し，判定モデルの評価のための「汎化性能評価データ」とした．

表 1　混雑状況の判定基準

基準 1	画像中に来場者が 3 人以下の場合，Lv1（空いている）とする．
基準 2	画像中の全ての通路で，人にぶつからずに進めない場合は Lv3（混雑している）とする．
基準 3	基準 1 および基準 2 に該当しない場合は，Lv2（中程度）とする．

表 2　データ拡張のための画像変換の組合せ

画像変換	拡張 A	拡張 B	拡張 C	拡張 D
①　ランダムな値で明度変化	○	○	○	○
②　ランダムなサイズで切出し	○	○	○	○
③　ランダムな角度で回転	○		○	
④　左右反転	○	○		
④'　ランダムに左右反転			○	○
⑤　入力画像を含む	○	○		
備考	変換ごとに出力		全変換をして出力	

表 3　各画像セットの画像数

画像セット	画像数	画像セット	画像数
1	6,000	2	6,120
1A	30,000	2A	30,600
1B	24,000	2B	24,480
1C	30,000	2C	30,600
1D	24,000	2D	24,480

(2) データ拡張　画像セット 1 の各画像に対して表 2 に示す組合せで画像変換を行い，データ拡張を実施した．拡張 A および B では画像変換ごとに画像を出力し，C および D では全ての画像変換をした画像を出力した．同様に，画像セット 2 の各画像に対してもデータ拡張を実施した．ここで得られた画像セットの画像数を表 3 に示す．

2.2 判定モデルの学習

判定モデルとして畳込みニューラルネットワーク（CNN）の一形態である CaffeNet [2] および GoogLeNet [3] を用いた．どちらも出力層のクラス数を 3 とした．予備実験の結果を踏まえ，CaffeNet では画像セット 1，1B，2，2B を，GoogLeNet では画像セット 1，1D，2，2D を用いて学習を行った．

各画像セットの 75%を教師データ，25%をテストデータとして，バッチサイズ 25 で最大 100,000 イテレーション（iter）まで学習を行い，5,000 iter ごとにモデルを保存した．すなわち，CNN 2 種類，画像セット 10 種類，モデル 20 個の組合せにより，全 400 個の判定モデルを生成した．判定モデルは，CNN，画像セットおよび保存時の iter を組合せて，CaffeNet-1A-5,000 のように表記する．

なお，CNN の学習には深層学習のフレームワークである Caffe [2] を用いた．学習率は，CaffeNet では一定の値 0.0001 を，GoogLeNet では段階的に変化する値（初期値 0.01 に対して 5,000 iter ごとに 0.5 倍した値）を用いた．

*https://www.google.co.jp/imghp?hl=ja

†中部電力株式会社　エネルギー応用研究所
Chubu Electric Power Co., Inc.
Energy Applications Research and Development Center

表 4　CaffeNet による判定モデルの性能

	拡張 B		拡張なし	
実験①：社内展示会モデルによるクローズドメインの性能評価	89%	CaffeNet-1B-45,000	88%	CaffeNet-1-20,000
実験②：社内展示会モデルによるオープンドメインの性能評価	34%	CaffeNet-1B-55,000	39%	CaffeNet-1-25,000
実験③：社内＋社外展示会モデルによるオープンドメインの性能評価	63%	CaffeNet-2B-35,000	33%	CaffeNet-2-5,000

表 5　GoogLeNet による判定モデルの性能

	拡張 D		拡張なし	
実験①：社内展示会モデルによるクローズドメインの性能評価	87%	GoogLeNet-1D-35,000	87%	GoogLeNet-1-20,000
実験②：社内展示会モデルによるオープンドメインの性能評価	55%	GoogLeNet-1D-25,000	67%	GoogLeNet-1-15,000
実験③：社内＋社外展示会モデルによるオープンドメインの性能評価	81%	GoogLeNet-2D-15,000	57%	GoogLeNet-2-10,000

2.3 判定モデルの評価

まず CaffeNet による判定モデルの性能について述べる．社内展示会の画像で学習したモデルの，クローズドメインのテストデータに対する正解率（accuracy：混雑状況を正しく判定された画像の割合）を表 4 の実験①に示す．この結果から，クローズドメインについては，良好な性能が得られることがわかった．同じく，社外展示会のオープンドメインのテストデータ（2.1 節で準備した汎化性能評価データ）に対する正解率を表 4 の実験②に示す．この値は実験①と比べて低く，社内展示会の画像だけでは混雑状況の汎用的な判定基準は学習できていないと言える．一方，社内＋社外展示会の画像で学習したモデルの，社外展示会の画像に対する正解率を表 4 の実験③に示す．この値は実験②と比べて，拡張ありの場合に大きな向上が見られた．わずか 120 枚の追加画像とその拡張により，社内の 9 会場に限定されていた背景や混雑状況を示す特徴のバリエーションが増え，汎化性能の向上につながったと考えられる．

次に GoogLeNet による判定モデルの性能を表 5 に示す．実験①～③の結果は CaffeNet（表 4）と同様の傾向であったが，GoogLeNet の方が全般に性能が良いことが確認できた．

3. 判定根拠の考察

判定モデルが画像中の何に注目して判断を行っているか，Selvaraju らの 2 つの手法 [4] を用いて可視化することで，判定根拠について考察を行う．第一の手法である Grad-CAM では，画像を CNN に入力した際の，あるクラスの出力確率に寄与する畳込み層の特徴マップにおけるピクセルの位置情報を勾配のヒートマップとして可視化している．第二の手法である Guided Grad-CAM では，高解像度の可視化手法である guided backpropagation [5] による画像から Grad-CAM を用いてクラスに関連する箇所を特定している．

判定モデル CaffeNet-2B-35,000 を用いて，社内展示会の画像を判定した際の判定根拠を可視化した（図 1）．Lv3 の判定では，混雑した人々の中心付近が判定の根拠となっており，人の頭部による粒状の模様が密集した様子が可視化されている（図 1(a)）．同様に Lv2 の判定には，画像中のまばらな来場者が寄与しており，ある程度距離を置いた人のシルエットを判定の根拠にしている（図 1(b)）．対照的に，Lv1 の判定は，人工物や構造物のような背景を判定根拠にしている（図 1(c)）．一方で，誤判定の例を見ると，人物がいる画像に対して壁などを誤認する場合（図 1(d)）や，人物がまばらな画像に対して細かな模様を誤認する場合（図 1(e)）があった．

このように，学習したモデルの判定根拠は，人間の感覚と類似していることがわかった．混雑のような境界が曖昧な概念も，ビッグデータを用いて人の判断基準を学習させることにより，モデル化が可能であることがわかった．

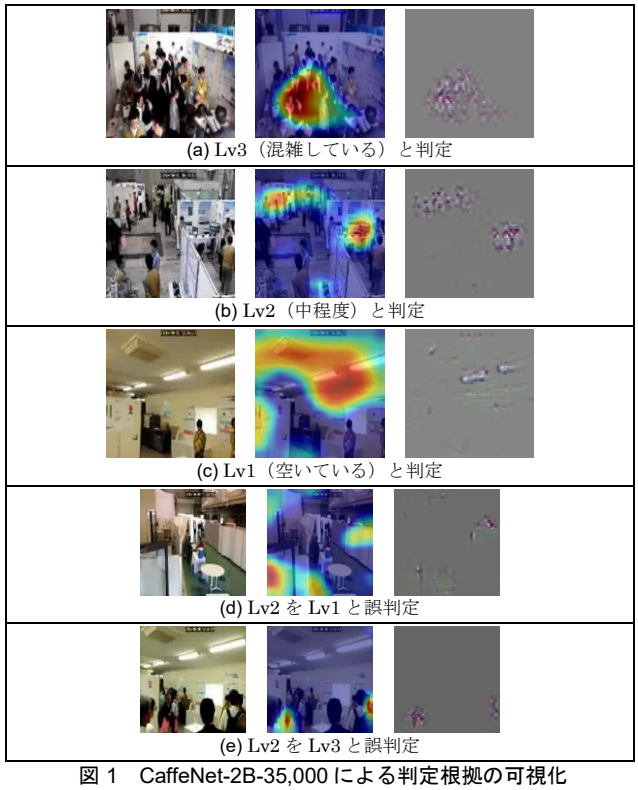
(a) Lv3（混雑している）と判定
(b) Lv2（中程度）と判定
(c) Lv1（空いている）と判定
(d) Lv2 を Lv1 と誤判定
(e) Lv2 を Lv3 と誤判定

図 1　CaffeNet-2B-35,000 による判定根拠の可視化
（左：入力画像，中：Grad-CAM，右：Guided Grad-CAM）

4. おわりに

本研究では静止画を対象とし，深層学習を用いて 3 段階の混雑状況の判定を試み，高精度な判定モデルを生成可能なことを示した．また，判定根拠を可視化することにより，ビッグデータを用いて人間の感覚を模倣した判定モデルが学習できていることがわかった．

参考文献

[1] 米司健一 他. “駅構内モニタカメラを用いた混雑度可視化技術”. 情報処理学会デジタルプラクティス. 2017, vol. 8, No.2, pp.152-159.
[2] Yangqing Jia et al. “Caffe: Convolutional Architecture for Fast Feature Embedding”. In The 22nd Association for Computing Machinery International Conference on Multimedia. 2014, pp.675-678.
[3] Christian Szegedy et al. “Going Deeper with Convolutions”. In The IEEE Conference on Computer Vision and Pattern Recognition. 2015, pp.1-9.
[4] Ramprasaath R. Selvaraju et al. “Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization”. In The IEEE Conference on Computer Vision. 2017, pp.618-626.
[5] Jost Tobias Springenberg et al. “Striving for Simplicity: The All Convolutional Net”. In The International Conference on Learning Representations. 2015. CoRR, vol. abs/1412.6806.