The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

深層学習を用いた高解像度画像からの人物カウント

Counting People via High Resolution Images using Deep Learning

山田 佑亮 *1 Yusuke Yamada　河野 慎 *2 Makoto Kawano　米澤 拓郎 *2 Takuro Yonezawa　中澤 仁 *1 Jin Nakazawa

*1慶應義塾大学環境情報学部
Faculty of Environment and Information Studies, Keio University

*2慶應義塾大学大学院政策・メディア研究科
Graduate School of Media and Governance, Keio University

We propose the method which is counting people in high resolution and wide areas images via deep neural network. There are many areas where there are no people in the wide area image. We improve the imbalance of this dataset using sampling algorithm. Our experiments demonstrate the out method effectiveness.We think that this method contribute to tourist spot. Because we can measure crowdedness in tourist spot and analyze the ability to attract customers.

1. はじめに

第一次産業やスポーツなどへの IT の導入が進む一方で，公共施設や観光地への IT の導入が進んでおらず，様々な問題が生じている. その問題の一つに，観光地の混雑状況を計測することができないことが挙げられる．混雑状況を自動で計測することができれば，その観光地の集客力などを定量的に分析することが可能となる．現在の人物カウントのシステムは，交通量把握のために，単一の交差点や狭いエリアにおいて，物体検出手法を用いたモデルを用いて行われることが多い．だが，広域の観光地の混雑状況を計測する場合，混雑時に人が重なってしまったり，広域であるがゆえに人が小さく写ってしまうなど，検出をする上での大きな弊害が存在する．そのため，従来の物体検出手法で検出することができない．実際に，HOG 特徴量を用いた人物検出 [Dalal 05] を今回のデータセットにおいて適用させた結果が，図 1 であり，検出ができていないことが分かる．

本研究では，観光地の広域捉えた高解像度の画像から密度推定を用いて人物カウントを自動で行うシステムを提案する．Zhangh らは，密度推定を用いた Crowd Counting CNN において，画像内の一部をパッチとしてランダムにサンプリングして学習させている．しかし，広域の観光地の画像は，広域であるがゆえに，人が一人もいないエリアが多くを占めるデータである．そのため，ランダムにパッチをサンプリングして学習をしまうと，人がいるエリアの学習が安定しない．本研究においては，データの不均衡性に対処するために，サンプリングアルゴリズムによって高解像度の画像からパッチを抽出して学習させた．提案手法の有効性を検証するため，人物カウントの精度比較の実験を行った．データセットには，江ノ島の展望台で実際に撮影された海岸沿いの高解像度画像 (図 2) を用いたものであり，解像度は 4000×6000 である．実験の結果，提案手法が既存研究よりも高い精度を得ることができた．

連絡先: 山田佑亮，慶應義塾大学
〒 252-0882 神奈川県藤沢市遠藤 5322 デルタ棟 S213,
yamad@ht.sfc.keio.ac.jp

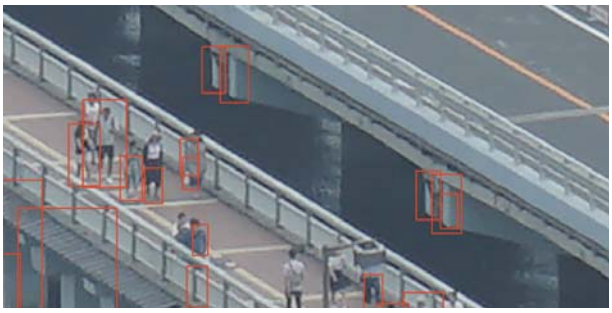

図 1: HOG 特徴量を用いた人物検出

図 2: データセットの例

2. 関連研究

人物カウント手法においては，Crowd Counting CNN[Zhang 15] という手法が挙げられる．画像を小さいパッチに分割し，そのパッチの中で，人物の密度マップを作成し，密度を積分すると画像内の人物の数になるように正規化する．それを CNN を用いた回帰モデルとして学習させる．予測時には，画像をパッチサイズに分割して，それぞれで得られた人物の数を合計する．このモデルの派生研究はいくつか存在する．Image Crowd Counting Using Convolutional Neural Network and Markov Random Field [Han 17] は，隣接するパッチは，密度に大きな差が存在しないという仮定を置き，パッチごとに密度マップを推定した後，その密度

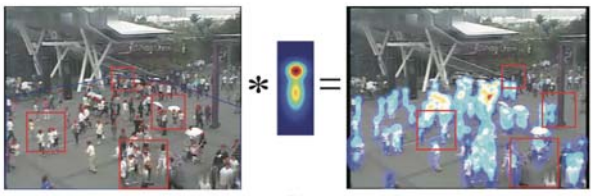

図 3: Crowd Counting CNN の概要図 [Zhang 15]

マップに対してマルコフ確率場により隣接するパッチの平滑化を行うことで，パッチ間の関係性を踏まえた人物カウントを行うことができるようになる．もう一つの派生研究として，Switching Convolutional Neural Network for Crowd Counting[Sam 17] がある．これは，分割したパッチを 3 種類に分類する Switching Network を学習させ，分類した後にそれぞれ違うアーキテクチャのネットワークで密度推定を行うというモデルである．これらの手法は密度推定による人物カウントにおいて，高い精度を出している．一方で，これらの研究で扱われている画像は，撮影しているエリアが狭域であるが，今回の研究におけるデータセットは高解像度で広域な画像であり，人がいないエリアと人がいるエリアのデータ量が不均衡である．そのため，既存手法をそのまま使用しても，高い精度を期待できない．

3. アプローチ

本研究で扱うデータは，既存研究のデータセットが捉えているエリアより広域であり，人がいない領域が大きい．そのため，今回扱う画像は不均衡データであると言え，学習時に，データサンプリング手法を用いる必要がある．不均衡データにおけるサンプリングの手法として，under-sampling[Kubat 97] が挙げられる．これは，クラス間でデータ数が不均衡である場合，データ数が多いクラスが，学習に使うデータ数を減らし，データ数が少ない方に数を合わせるという手法である．over-sampling[Chawla 02] は，データ数が少ない方のデータを増やし，データ数が多いクラスと同じ数にすることで，不均衡性に対処するという手法である．今回は，これらを組み合わせたサンプリング手法を用いる．

3.1 Crowd Counting CNN

本研究は，Crowd Counting CNN と同じく，密度推定を用いて人物カウントを行う．まず，学習用のデータを用意する．学習用のデータは，72×72 のパッチごとに作成する．学習用の画像全てにおいて，人物の頭の座標点のみアノテーションを行い，この座標点を基にガウス密度マップを生成する．任意のピクセル p の密度は，次の式によって算出される．

$$D(p) = \sum_{P \in P_i} \frac{1}{\|Z\|} (\mathcal{N}_h(p; P_h, \sigma_h) + \mathcal{N}_b(p; P_b, \Sigma)) \quad (1)$$

ここで，教師データにおける画像内の人の数と，密度マップの合計を等しい値にするために，正規化定数である $\frac{1}{\|Z\|}$ を使用する．P_h はアノテーションされた頭の座標点であり，P_b は体の座標点を示す．体の座標点は，人の身長を 1．75m，その画像内における 1 ピクセルあたりの長さ (m) を M とすると

$$P_b(x, y) = (P_h(x), P_h(y) + 1.75(M/2)) \quad (2)$$

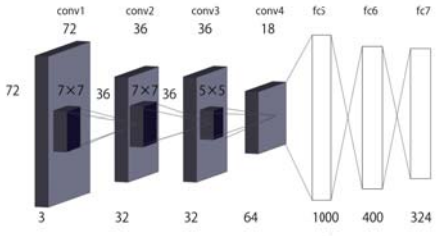

図 4: 提案手法のネットワーク構造

として計算される．この計算式は，頭の座標点から作成されるガウスカーネルの項と，体の座標点から作成されるガウスカーネルの項で構成され，結果として，図 3 で示されたヒートマップができる．また，σ の値は，σ_h を $0.2M$ として計算する．Σ の値は，σ_x を $0.2M$ として計算し，σ_y を $0.5M$ として計算する．

M の値は，画像内の特定の一人が y 方向に占めているピクセル数をアノテーションし，それを p として，

$$M = p/1.75 \quad (3)$$

として計算する．

この計算を，パッチの各ピクセルごとに行い，密度マップを生成する．これを教師データとして，CNN を学習させる．

3.2 ネットワーク

CNN は，72×72 の画像を受け取り，18×18 のサイズの密度マップを出力する構造である．具体的な構造は図 4 に示す．各層には，活性化関数として ReLU 関数を用いる．また，畳み込み層のあとには，ストライドが 2 の MaxPooling 層を用いている．ネットワークの最適化には，密度マップの各ピクセルの二乗誤差 L_D と，密度マップを合計して算出される人数の二乗誤差 L_C を使用し，

$$L = NL_D + L_C \quad (4)$$

で算出される．ここで，N は二つの誤差関数を調整するためのハイパーパラメータである．

3.3 サンプリング

先行研究においては，72×72 のパッチサイズに分割し，その中で人物検出を行なっていた．だが，今回取り扱う画像は，前述した通り，人がいない領域が広い．そのため，ランダムにパッチを作成した場合，ほとんどの場合に人がいないエリアがサプリングされるため，人がいるデータに対する学習がうまくいかない可能性がある．そこで今回は人がいるエリアといないエリアのデータ量を制御する確率 λ を導入し，λ をもとにデータをサンプリングする．画像はそれぞれ 72×72 の大きさでサンプリングされる．パッチが生成される際に 0 ～ 1 の間で乱数を発生させ，それが 0．7 以上であった場合，人がいる場所の中でランダムに一箇所選び，72×72 の大きさでサンプリングする．0．7 未満であった場合，画像内からランダムに 72×72 の大きさでサンプリングする．このとき，人がいる場所がクロップされる場合もある．このようにしてサンプリングされる箇所を制御して学習を行う．

4. 実験

今回の実験においては，サンプリングアルゴリズムを用いない Crowd Counting CNN と，サンプリングアルゴリズム

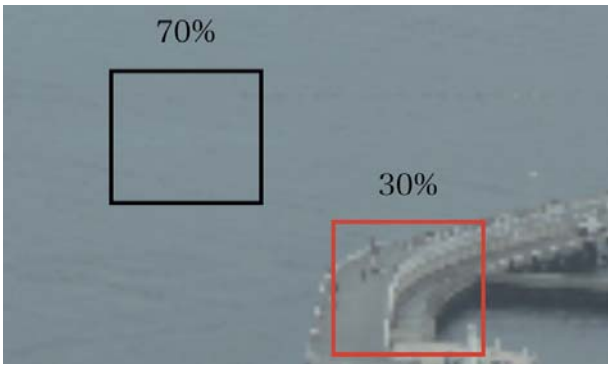

図 5: サンプリングの例

を導入して学習させた提案手法を比較し，精度の変化を見た．最適化手法は SGD を使用し，学習率は 0．01，エポック数は 7000 回とした．また今回は，バッチ処理で学習を行った．誤差関数において，$N = 10$ とし，サンプリングの確率において，$\lambda = 0.7$ を設定し，18 枚の 6000×4000 のデータセットに対して 6 枚のテストデータで K 分割交差検証法を行った．データセット一枚あたりの人物の数の平均は，132 人であった．

4.1 実験結果

表 1 に実験結果を示す．評価は，人物の数の正解値と，予測値の絶対二乗誤差の平均と絶対値誤差の平均の二種類で行う．結果，サンプリングアルゴリズムを導入したモデルの方が，既存のモデルの精度を大きく上回った．6000×4000 の画像を予測させるためには，CPU で 39.5 秒の時間がかかった．また，K 分割法における，人数の誤差の分散は 384.9 であり，標準偏差に直すと 19.6 であった．

4.2 考察

今回の実験では，サンプリングアルゴリズムを導入したモデルが，既存手法の精度を大きく上回った．実際，サンプリングアルゴリズムを導入していないモデルの予測結果の平均は，16．8 人，サンプリングアルゴリズムを導入したモデル (λ=0.7) の予測結果の平均は 163.1 人であった．これに対し，正解データの平均は 132 人であった．この結果より，サンプリングによってデータを操作することで，この高解像度のデータに対する予測値が大きく変化することがわかる．今回は，サンプリングアルゴリズムを導入したことによって不均衡なデータに対して学習をすることができたため，既存手法を大きく上回ることができたと考えられる．しかし，今回は平均で 31 人の誤検出をしてしまっている．今回のデータセットは，人であるかどうかわからないようなものが写り込むこともあるほど見分けがつかないものが多く，アノテーションの段階ではそのようなデータを全て抜いているので，正解データより多くカウントしてしまっている可能性も考えられる．

一方，予測にかかる時間は長い．これは，モデルの学習を 72×72 のパッチサイズで行ったため，6000×4000 の画像を分割してそれぞれを計算するという処理に時間がかかったものと考えられる。

5. 今後の課題

5.1 λ の探索

今回は，$\lambda = 0.7$ においてのみ学習をしたのだが，この λ はハイパーパラメータであり，この λ の値をグリッドサーチで探索することによって，より高い精度が得られる可能性もある．

表 1: 実験結果

手法	MAE	MSE
提案手法	30．5 ± 19.6	1315.1 ± 384.9
Crowd Counting CNN	115.8 ± 7.8	14403.8 ± 61.4

5.2 マルコフ確率場の導入

先行研究でも取り上げた，Crowd Counting CNN への Markov Random Field の適用 [Han 17] は，今回のデータセットにおいても有効な手法であると考えられる．今回のデータセットにおいてはビルや海のエリアなどが存在するため，これらのエリアの一つのパッチにおいて大きな誤検出が起きた場合に，隣接するピクセルが誤検出をしなければ，誤検出のパッチは平滑化される．つまり，人がいないエリアが非常に大きい今回の画像において，マルコフ確率場による平滑化は誤検出を抑える手法として非常に有効である．

6. まとめ

観光地で混雑状況を計測し，集客力などを定量的に分析するために，広域の人物カウントシステムは非常に価値のあるものである．本研究では，Crowd Counting CNN にサンプリングアルゴリズムを加えることによって，高解像度の観光地の画像から人物をカウントするための手法を提案した．この手法を用いることで，既存研究を大きく上回る精度を出すことができ，実際にこのシステムを観光地に導入することで，混雑状況を正確に知ることが期待できる．

参考文献

[Chawla 02] Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P.: SMOTE: synthetic minority over-sampling technique, *Journal of artificial intelligence research*, Vol. 16, pp. 321–357 (2002)

[Dalal 05] Dalal, N. and Triggs, B.: Histograms of oriented gradients for human detection, in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, Vol. 1, pp. 886–893IEEE (2005)

[Han 17] Han, K., Wan, W., Yao, H., and Hou, L.: Image Crowd Counting Using Convolutional Neural Network and Markov Random Field, *arXiv preprint arXiv:1706.03686* (2017)

[Kubat 97] Kubat, M., Matwin, S., et al.: Addressing the curse of imbalanced training sets: one-sided selection, in *ICML*, Vol. 97, pp. 179–186Nashville, USA (1997)

[Sam 17] Sam, D. B., Surya, S., and Babu, R. V.: Switching convolutional neural network for crowd counting, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Vol. 1, p. 6 (2017)

[Zhang 15] Zhang, C., Li, H., Wang, X., and Yang, X.: Cross-scene crowd counting via deep convolutional neural networks, in *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pp. 833–841IEEE (2015)