H-004

畳み込みニューラルネットワークを用いた一人称視点画像による自己位置推定

Estimating Self-location from First Person View using Convolutional Neural Network

杉中 出帆‡　飯塚 博幸‡　山本 雅人‡

Izuho Suginaka　Hiroyuki Iizuka　Masahito Yamamoto

1. はじめに

ロボットが自律的な行動を行うには，ロボットが環境を正確に把握する必要がある．環境を把握することに関して重要なことの一つにロボットの自己位置推定があり，現在，掃除ロボットや自動車などに適用されたり，その精度向上のための改良が進められたりしている．

自己位置推定の手法としては，GPS による位置情報，レーザによる距離の推定，360 度カメラによる三角法を用いた部屋の中心の推定がある．GPS 情報による位置情報は誤差が大きいことや屋内での使用が困難であり，レーザによる推定は高コストであり，三角法を用いた部屋の中心位置の推定は部屋という特定の場でのみでしか効果がないというデメリットがある．本研究では，人が一人称視点の景色から自己位置推定が行えることに着目して，人の脳を模倣した機械学習モデルであるニューラルネットワーク，特に近年画像認識で高い性能をあげている畳み込みニューラルネットワーク(以下，CNN)を用いる．外部装置を用いずに，低コストで，汎用的な自己位置推定を行うため，一人称視点の画像による自己位置推定を獲得したニューラルネットワークを構成する．

CNN を用いた自己位置推定に関連した研究として，石伏らは，自己位置推定の手法である MCL(Monte-Carlo Localization)における制御情報や観測情報の計測誤差の大域的な誤差の修正をするため，CNN の物体認識結果を統計的に統合して用いる手法を提案している [1]．ここでの CNN は 1000 個の物体を分類する学習済みのモデルを使用しており，認識した物体の系列と物体の位置に関する事前知識から自己位置を推定する．ある事前に決められたランドマークとしての物体のクラス分類を利用する場合には，そのランドマークが偶然なにかの影になってしまった場合には，全く異なる位置として認識されてしまう．本研究では，CNN を物体認識のツールとして用いるのではなく，空間を直接把握するニューラルネットワークを構成し，特にランドマークを定めることなく得られた一人称視点画像から自己位置推定をする．

2. 提案手法

本研究では，数多くの一人称視点からの画像と正確な位置を教師信号とするため， 容易にたくさんのデータを取得することのできる仮想環境である Minecraft [2,3]を用いて，CNN による自己位置推定を行った．

‡北海道大学, Hokkaido University

(a) 一人称視点画像例　(b) 実験で用いるフィールド

(c) 雨　(d) 夜　(e) 村人

図 1　実験に用いるフィールドと一人称視点画像

2.1 仮想空間としての Minecraft

Minecraft とは，Mojang AB 社の設立者 Notch により開発されたサンドボックスゲームである．Minecraft の世界は 1 ㎥立方体のブロックで構成されており，プレイヤは世界を自由に動き回り，素材の収集，武器の作成，建築などをして楽しむことができる．本研究で使用する一人称視点画像は Minecraft プレイヤの一人称視点画像である．仮想空間内の位置，方向，目の角度を定め，プレイヤを配置することで一人称視点画像を取得する．

2.2 畳み込みニューラルネットワークによる位置推定

CNN は，畳み込み層，プーリング層を交互に接続した構造を持ち，ネットワーク内で特徴抽出フィルタを学習する [4]．一人称視点画像による自己位置推定では，画像内の物体の形状，大きさ，種類と自己位置の関係が多岐にわたるため，CNN を用いた自己位置推定を行う．

$$y'_{ijkl} = \sum_{c=1}^{N}\sum_{p=1}^{m}\sum_{q=1}^{m} x_{(i+p)(j+q)c}\, W_{pqcl} + b_l \qquad (1)$$

$$y_{ijkl} = ReL\,U\left(y'_{ijkl}\right) \qquad (2)$$

式（1）は x を入力画素，W を重み， b をバイアスとした際に $m \times m$ サイズのフィルタを用い，N チャネルの画像を入力した際の出力 y である． y を式（2）の活性化関数で次の層に与える出力に変換する．プーリング層では，畳み込み層で出力された画像の解像度を下げ，入力画像間の微小な差に対する普遍性を実現する．本研究では，マックスプーリングを用いる．CNN の出力層付近には全結合層を配置する．全結合層では，畳み込み層とプーリング層で得ら

れた特徴をもとに，位置推定を行う．出力は位置の x 座標と y 座標に対応する 2 つのユニットを配置する．

本研究で使用するネットワークは 7 層で構成する．畳み込み層を C1，C2，プーリング層を P1，P2，全結合層 N1，N2 とし，C1，P1，C2，P2，N1，N2 の順に配置する．C1，C2 のフィルタサイズを 5×5，P1，P2 のフィルタサイズを 2×2 とし，特徴マップの縦×横×チャネルの数を入力，C1，P1，C2，P2，N1，N2 の順に 48×48×3，48×48×32，24×24×32，24×24×64，12×12×64，1×1×1024，1×1×2 とする．バッチサイズを 50，学習率を 0.0001，学習回数を 300 と設定する．

3. 数値計算実験

図 1(b)に赤枠で示した 40×40 ㎡内のフィールドにプレイヤの位置，方向をランダムに決定し，その場所からの一人称視点画像を取得する．このときのプレイヤの位置を教師信号として保存しておく．得られた 480×480×3 画素の画像を 48×48×3 画像に縮小して入力画像とする．

異なる環境設定での学習性能を評価するために，晴れ，雨，夜，村人 60 人，村人 100 人の 5 つの設定で各 12000 枚の画像を取得した（図 1(c)，(d)，(e)）．これに加え，各データを 2400 枚ずつ含む合計 12000 枚のデータセットを用意した（MIX）．雨，夜，の環境で学習するのは，ノイズや明度の違いに左右されずに位置推定を学習するためである．村人 60 人や 100 人の環境は，フィールド内に 60 人，または，100 人の村人をランダムに配置する環境である．この環境で学習するのは，画像の一部が隠れてしまった場合にも位置推定を学習できるようにするためである．

各環境設定で独立に学習した CNN モデルで訓練データ，テストデータに対する誤差値を算出することにより自己位置推定の性能評価をする．誤差値とは，CNN による推定位置と実際の位置とのユークリッド距離のテストデータ全体に対する平均 [m]である．各データセットを 6 分割し，訓練データ 10000 枚，テストデータ 2000 枚として訓練データ，テストデータを変更してそれぞれ 6 回実験を行った．

4. 実験結果

学習に用いた訓練データの種類と学習したモデルの各テスト画像における推定結果を表 1 に示す．訓練データに対しては 1 [m]程度の精度で位置推定ができ，同種類のテストデータに対してはおおよそ 3〜4 [m]程度の誤差であった．村人で学習したときは，60 人，100 人のいずれも雨の環境のテストデータに対する誤差が，晴れで学習したモデルのそれよりも小さい値を示した．すなわち，村人によって画像が部分的に遮蔽される環境で学習することで固定した特徴点を抽出するのではなく，画像から広く特徴を抽出しているためノイズの変換にロバストになった可能性がある．

空間内の各位置での自己位置推定と空間内の推定の全体的な歪みを明らかにするために，40×40 ㎡内に 5m 間隔のグリッド線を引き，格子点の位置から 8 方位の方向を向いたときに得られる画像から推定された位置を格子の関係を維持したままプロットした（図 2）．フィールドの外側に位置する点の推定結果が中心に歪んでいる．これは，推定結果をフィールドの中心から遠い位置に推定した場合，誤差値が大きくなりやすいためである．しかし，図 2 の上段中（上方向を見ているとき）の図の右上の格子の推定結果は，

表 1　各環境設定で学習したときの誤差値 [m]

テストデータの種類 / 訓練データの種類	訓練データ	晴れ	雨	夜	村人 60 人	村人 100 人	MIX
晴れ	1.09	2.98	8.02	11.7	3.42	3.82	6.03
雨	1.18	5.33	3.46	9.50	5.54	5.78	5.92
夜	0.95	11.4	10.8	2.81	11.3	11.5	9.51
村人 60 人	0.91	2.86	7.86	11.7	3.09	3.35	5.83
村人 100 人	1.03	3.04	7.95	11.8	3.23	3.43	5.93
MIX	1.18	3.56	4.72	4.55	3.91	4.17	4.21

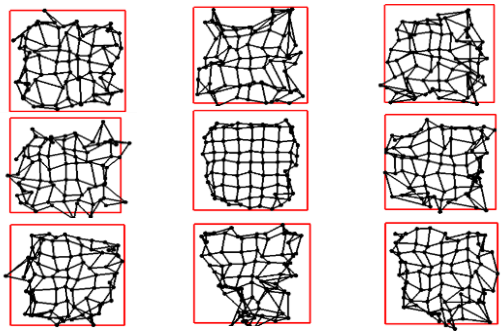

図 2　格子の位置推定（晴れ環境で学習したときの晴れ環境のテストデータに対する位置推定）：図の位置はプレイヤの向きに対応している．例えば，左下図は格子点から左下を見たときの推定位置を示す．中心の図は，各点の 8 方位の画像の推定位置の重心を示す．

特徴的な建物（右上のフィールド近くにある建物）が見られるため，そこに近い位置，つまり中心から離れた位置に推定している．一方，図 2 の下段中（下方向を見ているとき）の図の下部分の左右の格子が大きく歪んでいる．特徴の無い，対称性がある物体（フィールド下方向にある白い建物）の画像による位置推定は困難であることを示している．このとき，誤差がなるべく小さくなるように，下の領域の真中付近に推定するという合理的な結果が得られた．各点において，8 方向の画像からの推定位置から重心を求めると，空間の端の方を除いて元の格子を少ない歪みで再現することができていることがわかる．

5. おわりに

本研究では，CNN により，一人称視点画像による自己位置推定を行う実験を行った．これにより，CNN から直接自己位置を推定できることを示したが，建物に近く，壁が均一なパターンのときに誤差が大きくなった．これは見回して得られる画像から各推定位置の重心を求めることによって改善できた．

参考文献

[1] 石伏智,谷口彰,高野敏明,谷口忠大:Convolutional Neural Network による物体認識の自己位置推定への統計的活用,The 29th Annual Conference of the Japanese Society for Artificial Intelligence, 2015

[2] minecraft.net,https://minecraft.net

[3] PLANET MINECRAFT Imperial City Minecraft Project, http://www.planetminecraft.com/project/monumental-imperial-city

[4] 岡谷貴之,画像認識のための深層学習,人工知能学会誌,28 巻 6 号,pp.962-974(2013)