ISSN 2186-5647

9-7

空間画像判別におけるデータセット作成方法に関する研究
～空間用途と形容詞に着目したデータセット評価手法の提案～

日大生産工（院）　○鈴木　駿介
日大生産工　岩田　伸一郎

1. 研究背景・目的

画像認識技術を用いて、画像内の対象物が持つ固有の特徴を判別することが可能になっている。建築分野では、韓ら[1)]がSNS上で感覚的に評価された風景画像を対象に、畳み込みニューラルネットワーク[注1)]（以下、CNN）を適用し、いいね数に応じた判別を行うことで、明確な尺度を持たない風景画像の評価構造を分析している。そこで、同じ空間用途でも固有の特徴を持たず、人では特徴を見つけることが困難である空間画像に対してCNNを行い、判別の精度を検証することで、特徴が曖昧な空間用途の違いや人の感覚を可視化することができると考えた。

本稿では、CNNが得意とする対象物の正確な判別ではなく、感覚的な部分を人と同じように判別できるかどうかを段階的に検証する。そのため、一段階目として、空間の用途を正しく判別することができるかどうか検証する「名詞的判別」、二段階目として、感覚的に空間の説明を行う際に用いられやすい形容詞を正しく判別できるかどうか検証する「形容詞的判別」の二種類の判別を行うことで、空間画像の判別におけるCNNの精度やその有効性を検証することを目的とする。

2. 研究方法

2.1. 研究の概要

まず、空間の用途が明確に異なるが、使用しているインテリアが似ているため、空間の特徴のみで判別を行うことができると考えられる、「居住空間」と「執務空間」を対象の画像として収集し、判別を行う（以下、名詞的判別）。次に、建築空間を表す形容詞の中から、同じ空間画像を見ても評価が分かれやすい形容詞を対象とすることで、より曖昧な感覚を対象とすることができると考え、「現代的な」と「歴史的な」の二つの形容詞を判別する（以下、形容詞的判別）。形容詞的判別では、人の感覚を対象とするため、対象者によって画像を判別してもらい、CNNによって人と似た感覚の判別を行うことができるかを評価する。

2種類の判別結果をグラフ化し、その損失値[注2)]・正答率[注3)]や、学習と検証のグラフの乖離率から今回のモデルの精度や汎用性を評価・考察する。学習のフローを図1に示す。

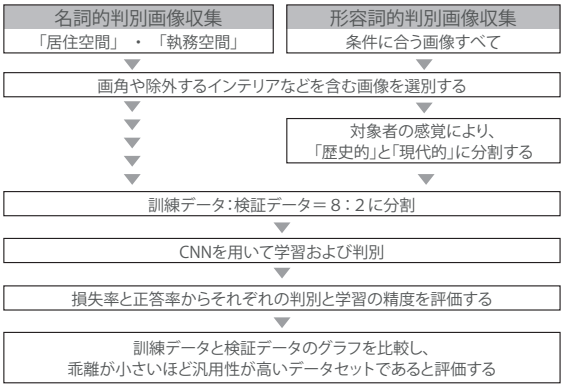


図1　研究のフロー図

2.2. 画像データセットの作成方法

本稿では、新建築データ[注4)]に掲載されている2024年9月号までのすべての建築作品を対象の中から、以下の条件に合う空間画像を検索し、選別したものを対象とする。

2.2.1. 名詞的判別対象画像の検索・収集条件

対象の空間用途の中でも、その用途に特徴のある画像を収集するため、居住空間の検索条件は、「新建築」「集合住宅」「住宅」「日本」「居住空間」とし、執務空間の検索条件は、「新建築」「事務所」「日本」「執務空間」とした。なお、用途の違いを明確にするため、店舗兼住宅やSOHOなど事務所兼住宅の場合は対象外とする。

また、検索で該当した建築の中から、画角の違いによる学習の不都合を防ぐため、壁や窓が正面に存在する画角の画像のみ収集する。

居住空間　執務空間

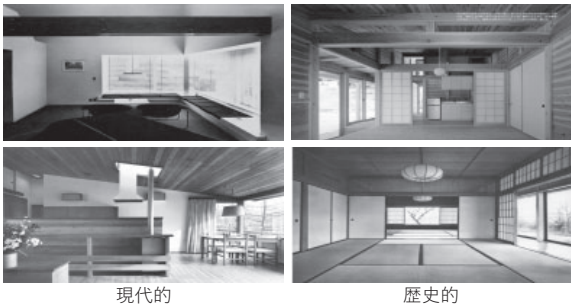
図2　名詞的判別における対象画像の例

A Study on Dataset Creation Methods in Spatial Image Discrimination
-Proposal of a dataset evaluation method focusing on spatial usage and adjectives-

Shunsuke SUZUKI and Shinichiro IWATA

2.2.2. 形容詞的判別対象画像の検索・収集条件

形容詞の違いのみを判別するため、建築用途を住宅に限定し、使用する部材の違いによって判別されることを防ぐため、構造形式は木造を対象とした。これを含む検索条件は「新建築」「木造」「集合住宅」「住宅」「日本」とする。なお、名詞的判別と同様に壁や窓が正面に存在する画角の画像のみを収集する。収集した画像の内、家具の新旧によって判別されないように、テレビやエアコンなどの電化製品が画像に含まれている場合は対象外とした。さらに、服装や髪型など人の持つ雰囲気が影響しないように、人が映っている場合も対象外とした。これらの条件によって収集した画像において、素材感や雰囲気以外の画質や色の有無によって判別されないように、すべての画像を白黒に加工した。このように収集した画像を、建築を学ぶ本学学生13名に「現代的」か「歴史的」かを感覚的に判別してもらい、各画像ごとにその割合を導く。それぞれ判別された割合が70%以上の画像を対象の形容詞が使用される空間としてデータセットを作成し、70%未満の画像については、人の意見が分かれるため評価が不可能と判断し、対象外とする。



現代的　　歴史的

図３　形容詞的判別における対象画像の例

2.2.3. 学習のためのデータセット分割方法

画像認識の学習のためにすべての画像を学習に用いる訓練データと学習の精度を検証するために用いる検証データに分割する必要がある。本研究では、訓練データ：検証データ＝8：2と設定し、それぞれの画像を分割する。以下に、対象建物数および対象画像枚数を示す。(表1)

表１　対象建物数および対象画像枚数

対象テキスト	建物数（棟）	総データ数（枚）	訓練データ数（枚）	検証データ数（枚）
居住空間	119	188	150	38
執務空間	72	105	84	21
現代的	282	239	191	48
歴史的		110	88	22

2.3. CNNの実装環境・設定方法

Google Drive上での画像を管理や、CNNを実装するために使用するディープラーニングモデルであるPytorchを利用することができるため、Google Claboratory[注5)]を使用してCNNを実装する。ネットワークモデルは既往論文より、感覚的な風景画像を判別することが可能であると分かっているため、本研究にも利用することができると考え、VGG-16[注6)]を使用する。入力画像サイズ[注7)]は判別に適しているファイルサイズと横長であることを考慮し、224px × 448px × 3とした。畳み込み層[注8)]と最大プーリング層[注9)]のカーネルサイズ[注10)]は今回の入力画像サイズに一般的に使用されやすいサイズを参照し、それぞれ3px × 3px、2px × 2pxとする。ストライド[注11)]も同様にそれぞれ1px、2pxとする。また、学習した結果を正確に検証するため、過学習[注12)]を起こさない範囲であった10回を学習回数とした。そして、バッチサイズ[注13)]はデータ枚数と使用されやすい数字を考慮し、訓練データで32、検証データで16とした。その他、使用するCNNモデルのパラメータはVGG-16を参照し、今回の入力画像サイズに合致するよう改良した。各パラメータの詳細を表2に示す。

表２　CNNモデルのパラメータ詳細

畳み込み層およびプーリング層のパラメータ詳細

	入力サイズ (px×px×チャネル数)	出力サイズ (px×px×チャネル数)		入力サイズ (px×px×チャネル数)	出力サイズ (px×px×チャネル数)
畳み込み層1	224×448×3	224×448×64	最大プーリング層3	56×112×256	28×56×256
畳み込み層2	224×448×64	224×448×64	畳み込み層8	28×56×256	28×56×512
最大プーリング層1	224×448×64	112×224×64	畳み込み層9,10	28×56×512	28×56×512
畳み込み層3	112×224×64	112×224×128			
畳み込み層4	112×224×128	112×224×128	最大プーリング層4	28×56×512	14×28×512
最大プーリング層2	112×224×128	56×112×128	畳み込み層11,12,13	14×28×512	14×28×512
畳み込み層5	56×112×128	56×112×256			
畳み込み層6,7	56×112×256	56×112×256	最大プーリング層5	14×28×512	7×14×512

全結合層のパラメータ詳細

入力サイズ	全結合層1,2	出力クラス数
7×14×512=50,176	4,096	2（判別した結果数）

2.4. 分析方法・考察の手順

各データセットに対して、名詞的判別と形容詞的判別の2つの判別を行い、それぞれの損失値・正答率を導く。それぞれの結果について、学習したモデルの精度を視覚的に確認するため、訓練データおよび検証データの損失値・正答率をグラフ化する。損失値のグラフは、右下がりであり、最小値が小さいほど精度が高く、正答率のグラフは右上がりであり、最大値が大きいほど精度が高いと評価する。また、損失値・正答率のどちらも、訓練データと検証データの乖離率が小さいほど、学習した結果の汎用性が高いCNNモデルであると評価する。図4、図5にCNN学習の概念図およびフローを示す。

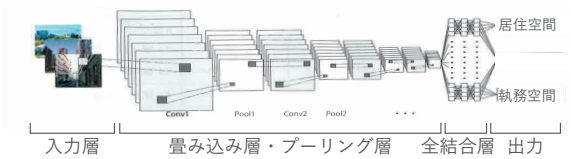


図４　CNN学習の概念図

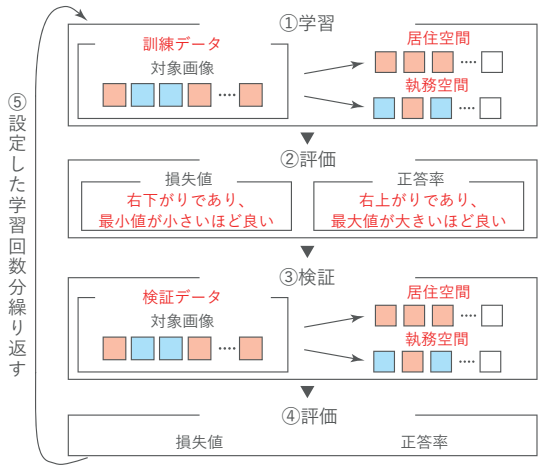


図５　CNN 学習のフロー

3. 名詞的判別の結果と考察

名詞的判別における、訓練データおよび検証データの損失値・正答率を学習回数ごとにまとめたグラフを下図に示す。（図６，図７）

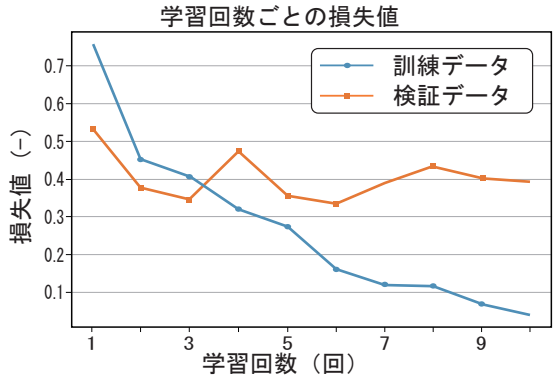


図６　名詞的判別における損失値の推移

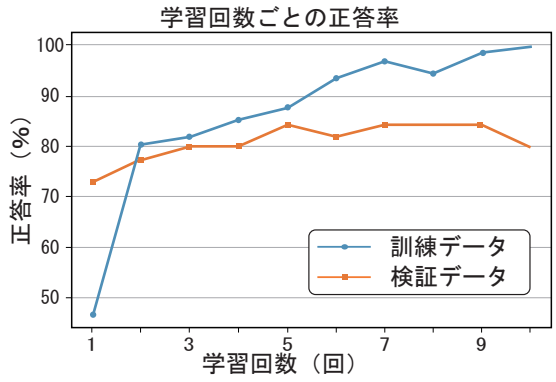
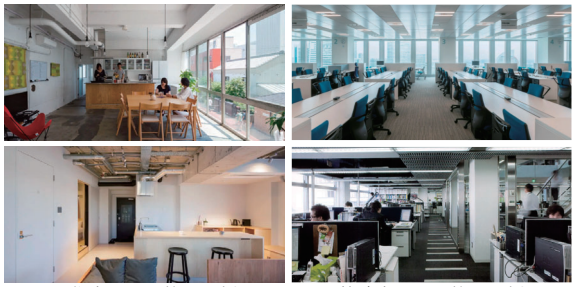


図７　名詞的判別における正答率の推移

図６より、訓練データの損失値は右下がりであり、最小で 0.235 であった。図７より、正答率は右上がりであり、最大値は 99.57%であったため高い精度といえる。検証データの損失値は約 0.3～0.5 の範囲で推移しており、最小で 0.279 であった。訓練データとの乖離は 10 回目で最大となった。正答率は、85%～90%の範囲で推移しており、最大で 88.14%となった。このように、損失値では訓練データとの乖離があり、正答率は 90%に満たない結果となった。これは、総データ数および判別クラス数が少なかったことに原因があると考えられる。空間用途などの空間の特徴を判別するためには、多様な空間用途を学習させることで、検証時に特徴を反映しやすく汎用性の高いデータセットになるのではないかと考える。

また、正答した画像および誤答した画像の例（図８，図９）より、椅子などのインテリアが複数存在し、天井高が低い居住空間は執務空間と誤答しやすく、インテリアの密度が小さい執務空間は居住空間と誤答しやすいといった印象を受けた。そのため、CNN を用いて画像から空間の用途を判別する際には、インテリアや壁や天井の関係から判別しているのではないかと考えられる。



居住空間で正答した例　　執務空間で正答した例

図８　名詞的判別において正答した画像例

居住空間と誤答した例　　執務空間と誤答した例

図９　名詞的判別において誤答した画像例

4. 形容詞的判別の結果と考察

形容詞的判別における、訓練データおよび検証データの損失値・正答率を学習回数ごとにまとめたグラフを下図に示す。（図１０，図１１）

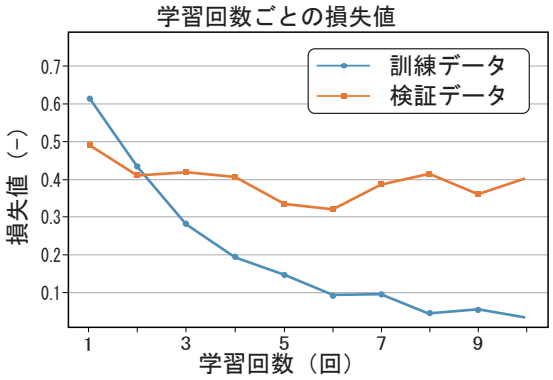


図１０　形容詞的判別における損失値の推移

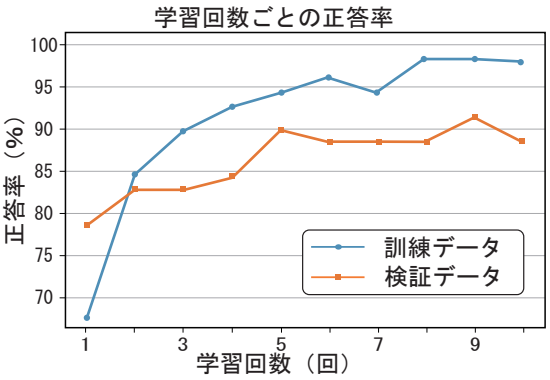


図１１　形容詞的判別における正答率の推移

図１０より、訓練データの損失値は右下がりであり、最小で 0.041 であった。図１１より、正答率は右上がりであり、最大で 98.57%であったため高い精度と言える。検証データの損失値は 0.5 ～ 0.3 の範囲で推移しており、最大 0.35 程度訓練データとの乖離が見られる。正答率は 80%～90%の範囲で推移しており、最大で 91.43%であった。また、訓練データとの乖離率は 8%程度であった。人の感覚における選別時に 70%以上の対象者が選択した画像であったため、検証データで 90%以上の正答率を得たことから、CNN を用いて人の感覚に近い判別をすることが可能であったと考えられる。しかし、損失値・正答率のどちらも訓練データと検証データとの乖離が見られたことや、グラフが一定の範囲で横這いになったことから、データセット作成時における画像の選別と CNN 設定値において改良の余地はあると考える。また、今回は感覚的に画像を分ける対象者が 13 名であったが、人数を増やすことで、より客観的なデータとすることが可能なのではないかと考える。さらに、選別時の評価方法を「現代的」と「歴史的」に加えて、曖昧な感覚の選択肢を増やすことで、明確に分かれた画像と曖昧だが多数の人が選別した画像との区別をつけることができ、学習の損失値を下げることができ、より汎用性の高いデータセットを作成することが可能だと考えられる。

正答した画像および誤答した画像の例を図１２，図１３に示す。なお、現代的と誤答した例はなかったため、歴史的と誤答した例のみ示している。正答した例を見ると、現代的な空間は壁材が白く、明るい印象を受けるのに対し、歴史的な空間は木を使用している面積が多かった。また、掛け軸やいろりなどのインテリアが画像内に存在している画像で正答していることから、このようなインテリアによって判別されている可能性も考えられる。このことから、CNN は素材感やインテリアなどの影響を受けていることが考えられる。

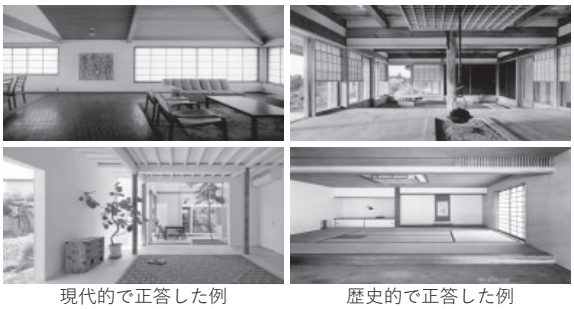


図１２　形容詞的判別において正答した画像例

図１３　形容詞的判別において誤答した画像例

注釈

注 1)　画像認識におけるディープラーニングのためのネットワークアーキテクチャ。画像の中からパターンを検出し、クラス化・カテゴリ化をする際に適している。

注 2)　モデルが予測した結果と実際の正解との間の誤差を表し、この値が減少しないまたは増加すると過学習を起こしていると判断できる。この値が小さいほどモデルの予測性能が高いとされる。

注 3)　正しく判別したデータの割合を示す。この割合が大きいほど、モデルの予測性能が高いとされる。

注 4)　株式会社新建築データが運営する、建築雑誌新建築などを Web 上で閲覧することができるサービス。

注 5)　Google 社が提供する、ブラウザ上で Python を記述、実行することができるサービス

注 6)　「ImageNet」と呼ばれる莫大な画像データセットで学習されたニューラルネットワークモデル。13 層の畳み込み層と 5 層の最大プーリング層、2 層の全結合層からなる。入力された画像を 1000 のクラスに分類が可能。

注 7)　ネットワークモデルに入力する画像のサイズ。縦画素数×横画素数×色チャネル数で表す。

注 8)　フィルタをスライドさせながら演算を行い、元の画像の特徴がどこに存在しているのかという情報を取得する層を表す。

注 9)　特徴を残しつつ画像を圧縮することで、ズレに対して頑健性を高く保つための層を表す。

注 10)　畳み込み演算を行う際に、入力画像の局所的な特徴を抽出するため適用するフィルタの大きさを表す。

注 11)　フィルタを適用する間隔を表す。

注 12)　学習した結果、特定の特徴のみを抽出してしまうことなどを原因として、学習の損失値が高くなりモデルの汎用性が低いと評価される。

注 13)　機械学習時に最適なパラメータを発見するために行う、学習するデータセットをいくつかのグループに分ける操作におけるグループ内のデータ数を表す。

参考文献

1)　韓天桐，鈴木俊策，岩田伸一郎：畳み込みニューラルネットワークを用いた SNS 上の風景画像の評価構造に関する研究；日本建築学会大会学術講演梗概集， pp. 93-94, 2022