*The 35th Annual Conference of the Japanese Society for Artificial Intelligence, 2021*

# イラスト認識精度向上のためのスタイル転移による形状バイアス学習

## Shape biased learning using style transfer for improving accuracy of illustration recognition

向後　ジェフリー[*1]
Jeffrey Kougo

渡邊　敬之[*1]
Takayuki Watanabe

大和　淳司[*1]
Junji Yamato

平　博順[*2]
Hirotoshi Taira

成松　宏美[*3]
Hiromi Narimatsu

杉山　弘晃[*3]
Hiroaki Sugiyama

[*1] 工学院大学 Kogakuin University
[*2] 大阪工業大学 Osaka Institute of Technology
[*3] NTT コミュニケーション科学基礎研究所 NTT Communication Science Laboratories

Illustration images, such as those used as options in English exam questions, tend to have lower accuracy than photographs in object recognition using CNN, etc. It has been pointed out that CNN learns more texture than shape in object recognition, which is presumably an obstacle to improving the recognition rate of illustration images. In this study, we tried a method that inhibits the learning of texture information by synthesizing various textures for object images of the same shape utilizing style transition, promoting the learning of shape information, and confirmed the improvement of the recognition rate of illustration images.

## 1. はじめに

### 1.1 研究背景

我々は，2011 年に開始した国立情報学研究所の人工知能プロジェクト「ロボットは東大に入れるか?」[新井 12]において，センター試験英語のイラスト問題に取り組んでいる．出題されるイラスト画像から描かれているオブジェクトを認識，図 1 左に示す本，バッグなどの様にクラス分類してその結果を解答に用いる解法で研究が行われてきたが，これまでの取り組みの中で，イラストで描かれたオブジェクトに対する認識精度が写真に対する認識精度よりも低いという課題が指摘されており，写真にはないイラスト特有の問題の影響が示唆されている[中村 18].

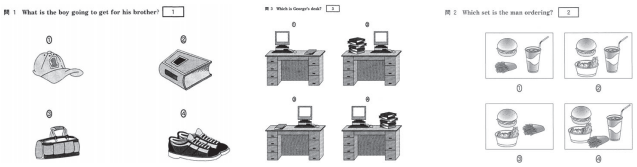

図 1　センター試験英語のイラスト問題の例
(左から 2006 年度試行，2006 年度本試，2008 年度追試)

### 1.2 研究目的

本研究ではイラスト画像に対する認識精度の向上を目的とするが，センター試験英語で出題されるイラスト問題の形式は様々あり，問題に応じて解答までの過程や処理が異なる．

そこで，センター試験英語で出題されるイラスト問題の形式について，「選択肢間の差異がどの様に設けられているか」に注目して図 1 の 3 種類に分類した．

問題の形式として，図 1 の左から，「単一のオブジェクトの違い」，「オブジェクト間の位置関係」，「特定のオブジェクトの有無」がそれぞれ選択肢間の差異となっている．この 3 つの内，単純な形式は「単一のオブジェクトの違い」が選択肢間の差異となる問題である[盛 18].

そこで，本研究では単一のオブジェクトが描かれたイラスト問題に絞って取り扱う．すなわち，本研究の目的は，イラストで描かれた単一のオブジェクトに対する認識精度を向上させる事である．その為，イラスト画像に対する認識精度が低い課題について原因を考察し，Robert ら[Geirhos 19]のアプローチを適用して認識精度の比較，検証を行った．

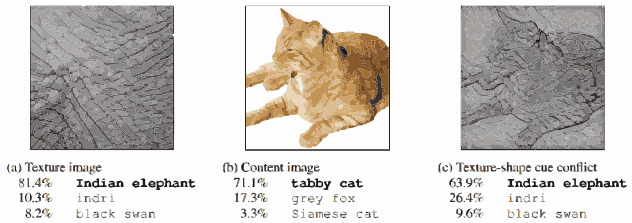


図 2　ImageNet で学習した ResNet-50 に 3 枚の画像を認識させた結果
(左から象，猫，猫の輪郭と象のテクスチャの合成画像)

## 2. 本研究のアプローチ

### 2.1 参考研究について

Robert らは，ImageNet で学習した ResNet-50 を用いて図 2 右に示す様な猫の輪郭に象のテクスチャを合成した画像を認識させた．結果，ResNet-50 はテクスチャ情報の元である象のクラスを認識した．更に，ImageNet を学習させた 4 つの CNN(ResNet-50，AlexNet，GoogLeNet，VGG-16)と被験者 97 名を対象に 5 つの画像形式ごとに画像の 16 クラス分類を行った．結果，画像にテクスチャ情報を持つ original と greyscale 画像は人間と同等の精度を示したが，テクスチャ情報を持たない silhouette，edges 画像に対しては人間以上に認識精度が低下した(図 3).

連絡先：向後ジェフリー，工学院大学大学院情報学専攻，
〒192-0015 東京都八王子市中野町 2665-1，
em20007@ns.kogakuin.ac.jp

以上の結果から，CNN が画像の学習を行う際，画像の形状情報よりもテクスチャ情報に重きをおいたテクスチャバイアスの学習を行っている事が明らかになった．

また，この論文では ImageNet の画像に様々なテクスチャを持つ画像を生成し Stylized-ImageNet というデータセットを作成して CNN に学習させる事で，画像のテクスチャ情報を中立化，形状バイアスな学習を行わせる事が可能である事が示された．

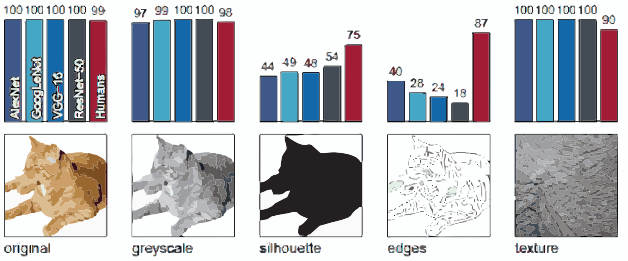


図 3　CNN と人間による 5 つの画像形式ごとの 16 クラス分類の正答率
(画像形式ごとに，左 4 本の棒グラフが各 CNN，右の 1 本が人間の正答率を示す)

## 2.2 イラスト画像の認識精度が低い原因について

参考研究より，CNN が画像を学習する際，クラス分類を行う上ではテクスチャバイアスの学習を行う事で解決してしまう特性が明らかになった．この事から，画像ごとのテクスチャ情報に変化が少ないイラスト画像では学習が難しく，この特性がイラスト画像に対する認識精度が低い原因になっていると考えられる．そこで，この課題に対する提案として，学習用のイラスト画像から様々なテクスチャを持つ画像を生成し，学習に利用することで同一クラス内における画像のテクスチャ情報を中立化させる．これにより，形状バイアスな学習を行わせることができ，認識精度を向上できると考えた．

図 4　[Geirhos 19]で用いられたスタイル転移の一例
(左が元画像，右の 10 枚が適用後の画像)

## 2.3 多数テクスチャを持つ画像の生成について

学習用のイラスト画像から多数のテクスチャを持つ画像を生成する方法について，本研究ではスタイル転移という手法を用いて画像の生成を行った．スタイル転移とは，図 4 の様に元の画像にエッジ検出を行って画像の形状情報を抽出し，別の画像をテクスチャとして合成する手法である．この手法を，複数のテクスチャ用画像を用意して適用する事で，同一形状で様々なテクスチャを持った画像をテクスチャ用画像の枚数分生成することが出来る．

本研究では，学習用の画像にエッジ検出とシルエット化を施してそれらを合成した画像(図 5)と，テクスチャ用画像として PASCAL VOC データセットから選定した 10 枚の画像 (図 6)を，Ostagram[Morugin]を用いて生成した(図 7)．

図 5　学習用のイラスト画像からエッジ検出とシルエットを合成した画像

図 6　本実験で用意したテクスチャ用画像の一部
(PASCAL VOC 2007 より 10 枚を選定)

図 7　Ostagram で合成した画像の一部

# 3. 実験

## 3.1 実験目的

本実験では，スタイル転移を学習用画像に適用した場合としなかった場合の学習における認識精度の変化を計測，比較する事を目的に，イラスト画像の 2 クラス分類(bag, book の 2 クラス)を行った．

## 3.2 実験手順

学習用画像に各種実験条件を適用してデータセットを作成，それぞれ学習を行い，認識精度を示す accuracy の値を比較した．学習器には SSD(Single Shot MultiBox Detector)[Wei 16]を使用した．実験の流れを以下に示す．

1. 学習用のイラスト画像を収集，実験条件に合わせて加工．
2. 加工した画像に crop と shift を行い Data Augmentation を実施，データセットを作成．
3. データセットを学習用 : 検証用 = 4:1 に分けて 5-fold cross validation を行い，accuracy を 5 回の平均で算出．
4. 1~3 を実験条件ごとに行い，算出した accuracy を比較．

表 1　各実験条件

| | 条件1 | 条件2 | 条件3 | 条件4 | 条件5 |
|---|---|---|---|---|---|
| 枚数 | 160 | 1600 | 1600 | 1600 | 16000 |
| サイズ | 300pixel * 300pixel | | | | |
| スタイル転移の有無 | 無 | 有 | 無 | 有 (grey) | 有 |
| Data Augmentation | 80 段階 | 80 段階 | 800 段階 | 80 段階 | 800 段階 |

### 3.3 実験条件

表 1 に，各実験条件の内容を示す．

Data Augmentation には， crop，shift を用いてデータセットの拡張を行った．crop は生成した画像を切り抜かない場合と画像の四隅から 180pixel * 180pixel に切り抜いたものを 300pixel * 300pixel に拡大した場合の 5 段階で適用した．shift は，画像全体を上下左右にそれぞれ最大 4pixel 分移動させた．

実験条件 1 では，従来手法として学習用のイラスト画像へのスタイル転移を行わず，80 段階の Data Augmentation を適用した合計 160 枚のデータセットを用いた．実験条件 2 では，提案手法として 10 枚のテクスチャ用画像を用いてスタイル転移を行い，80 段階の Data Augmentation を適用した 1600 枚の画像をデータセットとした．しかし，実験条件 1 と 2 ではデータセットの枚数が異なる為，実験条件 3 では，学習用画像へのスタイル転移は行わず，shift の刻み幅を細かくして Data Augmentation を 800 段階に拡張した 1600 枚のデータセットを用いた．実験条件 4 では，用意したテクスチャ用画像 10 枚をグレースケール化した状態でスタイル転移を適用し，Data Augmentation で枚数を拡張した 1600 枚の画像を用いた．実験条件 5 では，スタイル転移を適用し，Data Augmentation を 800 段階で拡張した合計 16000 枚の画像をデータセットとした．

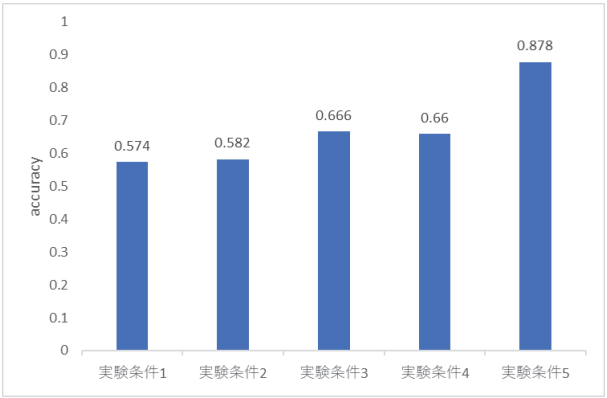


図 8 実験条件ごとの認識精度について

### 3.4 実験結果

図 8 に，実験条件ごとの認識精度を示す．それぞれ，条件 1 は 0.574，条件 2 は 0.582，条件 3 は 0.666，条件 4 は 0.660，条件 5 は 0.878 であった．

## 4. 考察

### 4.1 形状バイアス学習の検証

各実験条件のデータセットを学習した SSD を用いて，図 1 左の問題で選択肢に設けられている bag，book の画像を認識させた．結果，従来手法である条件 1，3 のデータセットで学習したモデルでは提案手法である条件 2，4，5 に比べ，一枚のイラストに対して同一のクラスを複数認識する場合(図 9)が多く見られた．この原因について，従来手法ではテクスチャバイアスでの学習が行われ，画像内でテクスチャ情報が酷似している箇所をオブジェクトとして認識していることが原因として考えられる．また，提案手法では一枚の画像に対し，同一のクラスが複数認識されることが少なかった為，形状バイアス学習が行われたと考えられる．

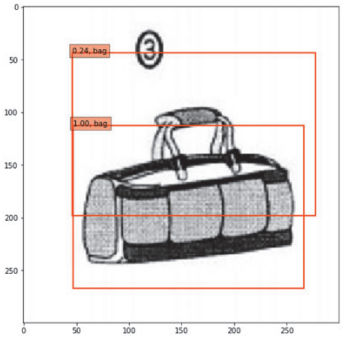


図 9 画像内に同一クラスを複数認識した一例
(実験条件 3 のデータセットを学習したモデルより)

### 4.2 提案手法による精度の検証

二点目に，条件 1 と 2 より，提案手法である条件 2 の認識精度が従来手法の条件 1 と比較して約 1%精度の向上が見られた．同様に，条件 3 と 5 では提案手法である条件 5 の認識精度が従来手法である条件 3 と比べて約 20%向上した．しかし，3.3 節でも述べた通り，提案手法ではスタイル転移によってデータセットの枚数が 10 倍に増加する為，向上した原因がテクスチャ情報の中立化によるものかは不明である．

また，枚数を揃えた条件 2 と 3 を比較すると，提案手法である条件 2 の認識精度が従来手法である条件 3 に比べ約 8%低下した．この原因について，以下の二点が考えられる．

- テクスチャ用画像の枚数が足りず，テクスチャ情報の中立化が不足していた．
- 本実験で用いたテクスチャ用画像 10 枚の特徴量に偏りが発生し，テクスチャ情報の共通性を学習してしまった．

### 4.3 テクスチャ情報を中立化させる画像特徴量の考察

最後に，テクスチャ情報の中立化に必要な画像の特徴量として画像の色情報の必要性を検討した．そこで，条件 2 と 4 を比較すると，テクスチャ用画像をグレースケール化させた条件 4 が条件 2 に比べて認識精度が約 8%向上した．この原因について，以下の二点が考えられる．

- テクスチャ用画像 10 枚の色情報に偏りが発生していた為，グレースケール化によって画像の色情報が無くなり認識精度が向上した．
- イラスト画像の認識に色情報が不要であり，グレースケール化によって認識に不必要な特徴量が削除された．

## 5. まとめ

本研究では，イラストで描かれた単一のオブジェクトに対する認識精度を向上させることを目的に，学習用画像に様々なテクスチャを持つ画像を用いて画像の形状情報に重きを置いた形状バイアスの学習を行い精度が向上するかを検証した．結果，形状バイアスの学習が行われ，提案手法の認識精度が従来手法に比べて最大で約20%向上したが，同じ枚数で比較すると認識精度は約 8%低下した．

本稿で示した実験では学習データ量がまだ十分に多いとは言えない．従って学習データ量の増加による精度向上が大きく影響する為，テクスチャ中立化の効果を正しく見極めるには不十分であった．そこで学習データ増加による影響を排除した評価を行うため，データセットの枚数を更に増加させ，認識精度が飽和した状態でテクスチャ情報の中立化を行う必要がある．今後こうした条件下で認識精度を計測し，より適切な評価を行う予定である．

## 参考文献

[新井 12] 新井紀子，松崎拓也，“ロボットは東大に入れるか？”，人工知能学会誌 27 巻 5 号 463-469，2012.

[中村 18] 中村未知，内田零，大和淳司，杉山弘晃，“センター試験英語イラスト問題の画像認識 -イラストとグレースケール画像の角度変化に対する汎化性能の比較- ”，第 21 回画像の認識・理解シンポジウム，2018.

[盛 18] 盛陽，南泰浩，“画像と文章知識の融合による英語テスト自動解答手法の研究”，情報処理学会第 80 回全国大会，2018.

[Geirhos 19] Robert Geirhos，et．al．“ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness”，ICLR，2019.

[Morugin]“Ostagram” https://www.ostagram.me/lenta?locale=en.

[Wei 16] Wei Liu，et．al．“SSD: Single Shot MultiBox Detector”，arXiv:1512.02325v5 [cs.CV] 29 Dec 2016.