Vol.2020-CVIM-220 No.20
2020/1/24

情景内文字情報を用いた情景認識

塩山 惇太郎[1,a)]　内田 誠一[,b)]

概要：身の回りの看板やポスター，そして商品パッケージ上の文字は，情景内文字と呼ばれ，物体や環境の正確な理解に必要な情報を付加していることが多い．例えば飲料ボトルのラベルに書かれた文字を読むことで，はじめてその飲料が何であるかを理解できる．また，会計という文字列があれば，その付近は何らかの商店であることもわかる．一方で，公園の写真にサッカーチーム名の書いた T シャツを着た人が映り込んでいれば，公園を競技場と誤認識する可能性もある．本研究では，このような文字情報の活用が情景認識の認識精度にどのように影響するのかについて実験的検討を行う．具体的には Place365-Standard データセットを用いて，撮影されている情景が 365 クラスのいづれかであるかを認識するタスクにおいて，映り込んだ文字情報の利用が有益なクラスや，逆に文字を使うことで性能が劣化するクラスを調査し，その原因を考察する．

1. はじめに

画像認識において，依然として困難とされるタスクに，情景認識がある．これは，撮影している情景がどこであるかを答えるタスクであり，例えば，商店正面や砂場が認識クラスの例となる．このタスクを困難にしている理由の一つは，クラスの曖昧さである．例えば，情景認識のタスクとして有名な Place データセット [1] においては，水が写っていそうなクラスとして，wave, lake, pond, ocean などがあるが，wave と ocean の区別は本質的に難しい．図 1 左も同様に，mountain か town なのか，人間にすら判断に迷うであろう．他に，商店に関係するものとして，Jewelry shop や gift shop, toy shop などがある．

クラス曖昧性を回避するための一つの方法として，我々人間は文字情報を積極的に利用している．例えば，この商店の例において，例えば商店の看板に書かれた toy という文字列や，画像に映り込んだ doll という文字列により，その商店が toy shop であると，極めて正確に認識できる．逆に言えば，文字がなくては判断に迷う場合において，我々は看板や店頭広告 (POP) などに必要な文字情報を加え，適宜情景認識が容易になるように支援しているとも言える．こうした曖昧性回避のための文字の利用は情景認識に限らず，例えばスポーツ選手の背番号や，商品パッケージに書かれた商品名，入口・出口のサインなど，至る所で有効活用されている．

[1] 九州大学
[a)] juntaro.shioyama@human.ait.kyushu-u.ac.jp
[b)] uchida@ait.kyushu-u.ac.jp

筆者らは，情景内の文字情報が，ラベルとメッセージの二種類に大別できると考えている．ラベルとはまさに曖昧性回避を目的としたものである．例えば，上述の店舗の看板のように，その文字が貼付された物体や周辺を説明するためのものである．図 1 右の例では，建物の外壁に店の名前が書かれることにより，建物に入らずとも，その建物がコーヒー店であることを認識できる．このようにラベルとしての文字は，情景の分類において大きな効果を発揮すると考えられる．特に外見特徴が非常に酷似している情景クラスにおいては，こうしたラベル情報を適切に使うことで，分類が容易になると期待できる．

一方のメッセージとは，本の本文ページやスマートフォン画面に書かれた文字情報である．それらを介して伝えられるものは，一般的な情報（ニュース，学術的知識，メールの伝言，広告，意見等）であり，その本やスマートフォンという物理的実体そのものを説明（非曖昧化）しているものではない．すなわちメッセージは，それが本のページに印刷されていようとスマートフォン画面に表示されていようと，その価値に違いはない．このように，同じ情景内の文字情報であっても，それがラベルとメッセージのどちらであるかによって，その機能は大きく異なると言える．

本報告では，情景内文字による情報が，情景認識にどのような影響を与えるかを，実験的に検証した結果について述べる．情景内の文字を検出し，正しく認識できれば，上述の toy shop の例のように，文字列により情景が非曖昧化されれば，文字情報はこの認識タスクに対してポジティブな効果を与えると期待できる．また間接的な効果ではあるが，情景内に文字が映り込んでいないことが有効である可

図 1 情景認識の困難性を示す例 (左) および情景内文字の有用性 (右)

能性もある．例えば海や空といった自然界の画像には文字が写ることは稀であるから，文字が存在しないことがそうした情景クラスの prior となり得る．

一方，以下二つの理由により，すべてのケースで文字情報が情景認識に有効とは限らない．そうした限界を生じる第一の理由として，情景内の文字情報がラベルであるかどうかを区別する方法が存在しないことが挙げられる．例えば，画像内に写った建物の壁にポスターが貼られていて，そこに “Concert” という単語があったとしても，その建物がコンサートホールであるとは限らない．この場合のポスターはむしろメッセージのコンテナであり，すなわちそのポスターが貼付されたものとは直接関係ない情報を提供している．このため，その文字情報をラベルとして利用したことで，建物がコンサートホールであると誤認識してしまうリスクがある．

第二の理由として，クラス分類に必要な文字情報がそもそも映り込んでいない画像も相当多いことが挙げられる．文字情報が得られなければ，そもそもポジティブな効果を期待することすら不可能である．また，もし文字列が映り込んでいたとしても，それがメッセージだったり，情景のラベルとしては不十分なものだったりすれば，やはり精度改善は不可能である．なお，海や空を例として文字情報の不在自体の効果を上で述べたが，その効果は実際のところむしろ限定的であると予想される．

以上の理由のため，本報告においては，情景内文字情報の利用による情景認識率の向上を目指すというよりは，どのようなケースにおいて情景内情報が有益なのか，もしくは無益，さらには悪影響するのかを観察することを主目的としている．具体的には，Place データセットに存在する 365 クラスにおいて，どのクラスで改善・改悪が見られるかを観察し，さらにそれらのクラスにおける情景内文字の状況について観察および考察を行う．

2. 関連手法

既存の情景認識手法の多くは，特に文字情報を意識せずに，画像全体を一括した情報源として扱って，その情景クラスを予測する．実際，Places365 [1] を公開した論文 [2] では CNN を使用して 50%以上の精度を達成しているが，文字情報を別途抽出して利用するといった試みはなされていない．

その一方で，情景内文字情報を使って fine-grained image classification を行う試みも皆無ではない．例えば文献 [3], [4], [5] では，画像内の文字情報を用いることで ImageNet の分類タスクを高精度化できることを示している．特に文献 [5] では，ImageNet 中の商店に関する 28 クラスを対象とした認識実験も行っており，本研究との関連は非常に深い．文献 [6] も同様のタスクにおいて，本論文で提案するのと類似した方法により，情景内文字情報と画像情報をの統合を行っている．文献 [7] では，商店正面画像 (storefronts) を用いた商店分類の際に，情景内文字情報（特に看板等に印字された情報）が有効であることを示している．

文献 [8] では，情景認識ではないが，情景内文字情報の興味深い利用形態を提案している．同文献は，Visual question answering のタスクを拡張し，情景内の文字を読まなくては回答できない問題を扱っている．例えば，「青いバスの行き先はどこか」や，「炭酸水の値段はいくらか」といった問題に回答するタスクである．前者の場合は，画像認識により青いバスを見つけた上で，その正面に書いている行き先情報を検出・認識する必要がある．後者においては，まずは画像認識によりボトルを見つけた上で，ボトルに添付されたラベルを文字認識することで炭酸水を見つけ，さらにその付近の値札を読み取る必要がある．実際にはこうした段階的な方法ではなく end-to-end な構成で実現できているが，いずれにせよ，こうした複雑な問題が扱えるようになったのは特筆すべきであろう．なおこのタスクに関しては，データが公開されており，ICDAR2019 においてコンペティションが開催されている [9]．

上記の試みを含め，情景内の文字情報を扱うには，情景内文字を検出し，さらに認識する必要がある．過去には困難な課題とされてきたが，いずれもラベル付きデータの公開と深層学習技術により高精度化の一途を辿っており，もはや実用化レベルと言っても差し支えない程度の性能を達成している．本実験では，検出には EAST [10]，認識には CRNN [11] という，すでに古典的ではあるが，安定した性能を持つ手法を利用する．当然ながら，未だ続々と提案されている最新の方法を用いて実験してもよいが，本報告の目的としている「情景認識における文字情報の効果の観察」であれば，これら古典的な方法でも十分であると考えた．

さらに，本報告では，いわゆる単語分散表現 (word embedding) により，情景内文字情報を単語ごとに意味ベクトル化する必要があった．これについてももはや古典的な word2vec [12] を用いることとした．最近ではフレーズや文を単位としてベクトル化が可能な方法も次々に提案されているが，情景内文字情報はビジネス文書のような一定したレイアウトに従って配列されておらず，そのためか検出・認識が単語単位で行われることが多い．また上述の通り，

本研究の目的が「効果の観察」であることから，現時点では単語単位のベクトル化を用いたほうがよいと判断した．

3. 情景内文字情報を活用した情景認識

本節では，文字情報を活用した情景認識手法について述べる．画像情報[*1]と情景文字情報の併用については，二つの手法を試みた．第一は画像情報と文字情報を特徴レベルで統合する手法，第二は文字情報法による重み付けに基づく手法である．以下にそれぞれについて述べる．

3.1 特徴レベルで画像情報と文字情報を統合した手法

図 2 に，特徴レベルで画像情報と文字情報を統合した手法を示す．一般的な CNN との相違点は全結合層への入力特徴である．一般的な手法では画像を CNN の畳み込み層に入力して得られた画像特徴 (25,088 次元) を全結合層に入力する．一方，文字情報を用いた本手法では，VGG16 モデルの CNN 層を画像特徴抽出器とし，入力画像から抽出した情景内文字（単語）を word2vec による分散表現で意味ベクトル化したもの (200 次元) と画像ベクトルと連結した 25,288 次元ベクトルを，全結合層への入力とする．これにより，情景の判断材料として文字情報のみを付加したこととなり，文字情報を加味したことによる情景認識結果の変化を調べることができる．なお，情景内単語が複数存在する場合，単語 1 語と画像をペアとして単語の数だけデータ数を増やして学習を行った．その際，文字の含まれていない画像については，200 次元の零ベクトルを連結している

3.2 単語とクラスの関連性により画像情報によるクラス尤度を重み付けする手法

図 3 に，単語とクラスの関連性により画像情報によるクラス尤度を重み付けする手法を示す．この重みの作成には，一般的な tf-idf を使用している．tf-idf とは，文章中に含まれる単語の重要度を評価する手法である．本研究では，Places365- standard データセットに含まれる 365 個のクラスを 365 種類の文章と捉え，各クラスに対する情景内にある各単語の重要度を 0 から 1 で表現している．

図 4 に，tf-idf を用いた各単語の重要度の求め方を示す．データセットの訓練用画像において検出・認識された単語集合のなかから，stop words ではなく，かつ高頻出単語リスト GLS [*2]に掲載された単語 2,098 語を抽出した．抽出した単語全てを Word2vec を用いてベクトル化し，k-means を用いてクラスタリングした．クラスタ数 $k = 1000$ とした．そして 1 クラスタを 1 単語のように見なして，すなわち語彙 $k = 1000$ として，各情景クラスについて tf-idf を

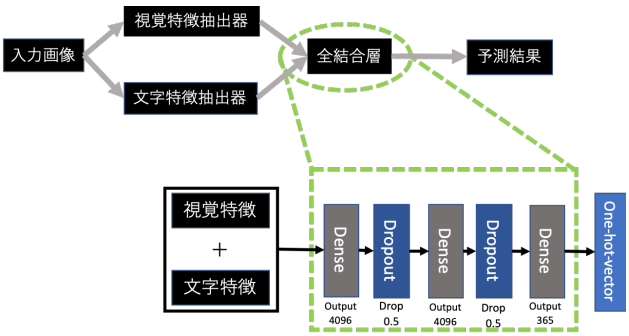

図 2 特徴レベルで画像情報と文字情報を統合した手法

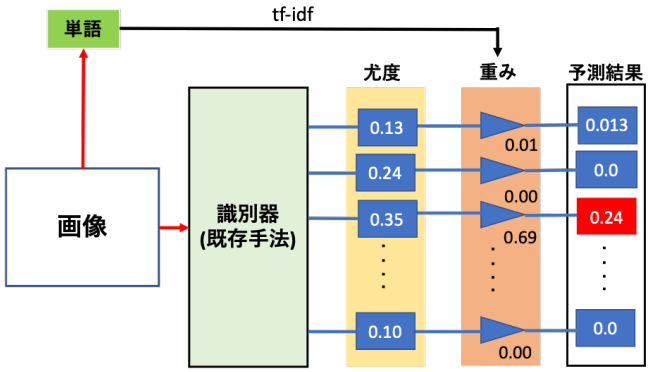

図 3 単語とクラスの関連性により画像情報によるクラス尤度を重み付けする手法

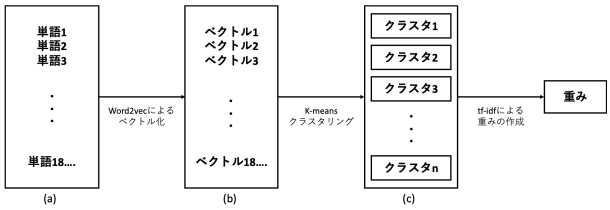

図 4 クラスタリングを伴った tf-idf による単語の重要度の導出

求めた．極力大規模な情景画像セットを用いたとしても，情景内文字として得られる単語数は限定的であり，従って訓練用画像中に含まれていない単語がテスト用画像中に含まれる可能性が高い．このため，本報告ではこのように word2vec とクラスタリングを用いて単語の量子化を行い，擬似的に語彙を縮小している．

4. 情景認識実験

4.1 データセット

本実験では，Places365-Standard データセットを使用した [1]．このデータセットは 365 種類の情景クラスから成り，訓練画像 1,803,460 枚，検証画像 36,500 枚，テスト画像 328,500 枚で構成されている．しかし，テスト画像についてはコンペティション用途のためにアノテーションがなされていない．よって検証画像の 7 割 25,550 枚を検証画像，3 割 10,950 枚をテスト画像とした．また，文字情報を

[*1] 情景内文字情報も元々は画像情報から抽出したものであるため，一種の画像情報であるが，本稿での画像情報とは画像全体を一括したビットマップとして見た場合を指す．
[*2] http://www.eapfoundation.com/vocab/general/gsl/frequency/

活用するために既存の文字認識器を用いて文字情報の抽出を行なった．

4.2 情景内文字検出

Places365 standard データセットに含まれる文字は，文字検出器 EAST（An Efficient and Accurate Scene Text Detector) [10] で検出を行なった．文字検出器 EAST は，FCN (Fully Convolutional Networks) を使用した情景内文字検出器である．EAST は任意の角度にある文字の検出に対応しているため，様々な文字が存在する情景内においても多くの文字を抽出することができる．既存の学習済みモデルは低解像度文字の検出が難しいため追加学習を行なった．追加学習には ICDAR2015 Robust Reading Competition データセットの train 画像 1000 枚を 1/1, 1/2, 1/3, 1/4 の解像度にリサイズしたものを使用している．図 5 に EAST による文字検出例を示す．

4.3 情景内文字認識

EAST によって得られた文字領域は，文字認識器 CRNN(Convolutional Recurrent Neural Networks)[11] で認識を行う．文字認識器 CRNN は CNN と RNN と組み合わせた構造をしており，単語領域が正しく与えられれば精度良く認識を行うことができ，複雑背景下でも，多くの文字を認識できる．CMN ならば単語辞書に依らず認識できるという利点の一方で，意味を成さない文字列が認識結果としとて与えられることがある．そこで本実験では stop words と使用頻度の低い単語を除いた GLS 単語 2,098 語を調査対象とした．図 5 に CRNN による文字認識例を示す．

4.4 単語ベクトルの作成

本研究では，Word2vec を使用し，情景内に存在する単語の数値化を行なった．Word2vec はディープラーニングを使用した単語埋め込み手法の一つであり，文章中に出現した単語とその前後に現れる単語の間にある関係を学習している．学習により，単語同士の意味が似ているか否かを類似度として数値化することができる．Word2vec によって得られた単語ベクトルは，単語の意味が類似したものは類似度が高く，意味が異なるものは類似度が低くなる．実験では，word2vec 学習用の基本コーパスである text8 を用いて学習したモデルを使用し，ベクトルの次元数を 200 次元とした．

4.5 画像情報を用いる CNN の学習

画像情報のみを用いた場合の情景認識のために，VGG16 アーキテクチャの学習を行なった．VGG16 とは畳み込み層 13 層，全結合層 3 層から成る CNN である．データセットには Places365-Standard の訓練画像 1,803,460 枚，作成した検証画像 25,550 枚を使用した．この場合の入力は 3

図 5 文字の検出と認識の例

表 1 情景認識精度 (%)

画像情報のみ	47.21
文字特徴レベルでの統合	47.61
クラス尤度の単語重要度による重み付け	47.20

チャネルの画像であり，文字情報は付加されていない．学習に際し，ネットワークの入力サイズである 224 × 224 にリサイズを行なった．こうして得られた CNN の畳み込み層は，「特徴レベルで画像情報と文字情報を統合した手法」での画像情報抽出器としても使用する．

4.6 情景認識精度

情景認識精度は，Places365-Standard データセットの検証画像を分割し作成した 10,950 枚を使用して評価を行った．画像によっては情景単語が複数個存在する場合があるため，図 7 のような操作を行った．まず，単語 1 つと画像をペアとし，ペアごとに尤度を求める．情景内の単語の数だけ得られた尤度は，クラスごとに平均することにより最終的な予測結果を得た．

各手法 (画像情報のみ，画像情報と文字情報を特徴レベルで統合，文字情報で重み付け) の認識精度を表 1 に示す．画像情報のみの場合に比べて精度改善が 0.4%と非常に小さいものの，全体の精度としては画像・文字情報を特徴レベルで統合した手法の精度が最も高いことがわかる．一方，文字情報による重みを付け加えた場合においては，わずかではあるが 0.01%の精度が低下していることもわかる．このように，情景認識の際に文字情報を利用しても，全クラスを一括してみる限り大きな精度改善は無かったことは，本報告による一つの知見である．

4.7 クラスごとの詳細解析

画像情報のみを用いた場合と，特徴レベルで画像情報と文字情報を統合した手法について，クラスごと正答数の比較を行った．表 2 は正答率が改善した上位 5 クラスと，改

表 2　正答数の向上率ランキング，降順 (左)，昇順 (右)

rank	class	rate(%)	rank	class	rate(%)
1	staircase	300	1	bookstore	68
2	lobby	300	2	archelogical excation	71
3	stadium football	233	3	medina	75
4	cottage	225	4	soccer field	78
5	artists loft	180	5	berth	80

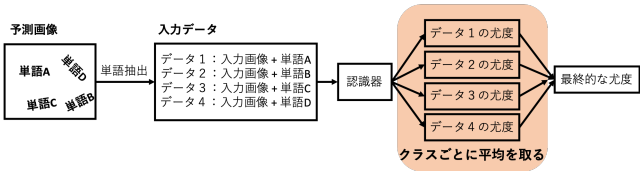

図 6　情景内に単語が複数存在する場合の対処方法

悪した 5 クラスである．改善クラスには，訓練画像そしてテスト画像ともに画像中に含まれる文字が少ないという特徴が見られる．訓練画像には 111,733 単語，テスト画像には 655 単語が含まれているが，ランキング上位に含まれる画像には，文字全体の 0.5%以下しか含まれていない．また，クラスによっては誤認識が増加しているものもあった．図 7 に結果の例を示す．ランキング下位に含まれるクラスには，単語が多く含まれるクラスが存在するという特徴が見られた．

4.8　考察

認識結果の詳細を調べることにより，正答数が増えたクラスに含まれる単語数は少なく，正答数が減ったクラスには単語数が多いことが確認できた．これにより，情景内に含まれる単語数と予測の正答数は比例しないとわかる．この結果と画像情報に文字の重みを加えた手法による結果から，文字情報が情景認識に与える影響は非常に小さく，文字がないということが分類精度を高める手がかりの一つとなっていると予想できる．

このような結果となった理由として，文字情報に対する価値の比重が小さ過ぎた可能性がある．なぜなら，Word2vec を用いて求めた 200 次元の文字特徴は画像特徴の約 1/125 のサイズであるため，画像情報がもたらす情報量は文字情報に比べ非常に大きい．よって，文字情報に大きな価値をおくことで異なる結果が得られる可能性がある．本手法では，我々人間のような文字情報を用いた認識精度の向上，という結果は得られなかったが，文字の有無によって分類性能が変化することが確認できた．適切な形で文字情報を量子化し，視覚情報と組み合わせることで，より人間に近い認識が実現できる可能性がある．

5.　おわりに

本論文では，情景認識に情景画像に写り込んだ文字情報を活用する手法について提案し，その特徴を調査した．そのために，Places365-Standard データセットに含まれる 180 万枚の画像から，情景文字検出器 EAST と情景文字認識器 CRNN を用いて単語の抽出を行なった．得られた単語は，Word2vec を用いて数値化を行い画像情報と連結し学習を行う手法，そして既存手法に情景画像内に現れる単語の重要度を tf-idf を用いて求めることにより重みづけを行う手法として既存手法と比較を行なった．

図 7　改善が見られたクラス (左) および改悪が見られたクラス (右)

謝辞

本研究の一部は，科研費（JP17H06100）に依った．

参考文献

[1] http://places2.csail.mit.edu
[2] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, “Places: A 10 million Image Database for Scene Recognition,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 40, no. 6, pp. 1452–1464, 2018.
[3] S. Karaoglu, J. C. Van Gemert, and T. Gevers, “Object reading: Text recognition for object recognition,”, LNCS, vol. 7585 (ECCV), 2012.
[4] S. Karaoglu, R. Tao, T. Gevers, and A. W. Smeulders, “Words matter: Scene text for image classification and retrieval,” IEEE Trans. Multimedia, vol. 19, no. 5, pp. 1063–1076, 2017.
[5] S. Karaoglu, R. Tao, J. C. Van Gemert, and T. Gevers, “Con-Text : Text Detection for Fine-Grained Object Classification,” IEEE Trans. Image Process., vol. 26, no. 8, pp. 3965–3980, 2017.
[6] X. Bai, M. Yang, P. Lyu, Y. Xu, and J. Luo, “Integrating scene text and visual appearance for fine-grained image classification,” IEEE Access, vol. 6, pp. 66322–66335, 2018.
[7] Y. Movshovitz-Attias, Q. Yu, M. C. Stumpe, V. Shet, S. Arnoud, and L. Yatziv, “Ontological supervision for fine grained classification of Street View storefronts,” Proc. CVPR. 2015.
[8] A. F. Biten et al., “Scene Text Visual Question Answering,” Proc. ICCV, 2019.
[9] A. F. Biten et al., “ICDAR 2019 Competition on Scene Text Visual Question Answering,” Proc. ICDAR, 2019.
[10] Xinyu Zhou, et, al., ”EAST: An Efficient and Accurate Scene Text Detector,” CVPR, pp. 5551-5560, 2017.
[11] Baoguang Shi, et al., ”An End-to-End Trainable Neural Network for Image-based Sequence Recognition and Its Application to Scene Text Recognition,” TPAMI, vol. 39, no. 11, pp. 2298-2304, 2017.
[12] T. Mikolov, et al., “Distributed Representations of Wordsand Phrases and Their Compositionality,” NIPS, pp.3111-3119, 2013.