自然画像におけるテキスト領域検出の高精度化に関する研究

The research about improvement of detecting texts in natural scenes

櫻庭 天[1)]
指導教員　青木 輝勝[1)]

1) 東京工科大学大学院 バイオ・情報メディア研究科 コンピュータサイエンス専攻
コンピュータビジョン研究室

キーワード：Text Detection, 深層学習, 物体検出, OCR

1.背景

自然画像におけるテキスト領域検出は,人工的でない現実世界の風景を撮影した画像中にある文字領域の位置を特定する技術である.この技術を光学文字認識（OCR）システムと連結させることにより,街中の外国語を日本語に翻訳したり,盲目者に対して音声ガイダンスを行ったりすることが可能となる.この分野にはベンチマークとなるデータセットがいくつかあり,ICDAR(International Conference on Document Analyis and Recognition)2015 もその一つである.ICDAR2015 は図 1 のように偶発的な場面で撮影された画像からなるデータセットである.本研究の目的は,自然画像におけるテキスト領域検出というタスクにおいて,precision,recall,F_1などの数値指標を用いて評価を行ったとき,提案手法が既存手法を上回る値をマークすることである.

図 1 データセット ICDAR2015 のある画像について Ground Truth を重畳した例

2．関連研究

本節では,テキスト領域検出における既存手法や本研究に関連する関連研究について述べる.

かつては 1 文字ずつ検出する方式や単語を直接検出する方式などが主流であったが,現在では任意形状のテキスト領域を検出するために物体検出手法をベースラインとした手法などが主流である.現在最高性能を誇るテキスト領域検出モデルである 2020 年に Ye らにより提案された TextFuseNet[1]もこの方式に属する.TextFuseNet は文字,単語,マスク領域の 3 特徴を用いてテキスト領域検出を行い,その検出結果を融合して最終出力を得る点が特徴である.TextFuseNet はデータセット ICDAR2015 のテストデータに対し 92.1%のF_1をマークしている.

また,当研究室の菅原が 2021 年に発表した結果[2]によると,文字認識において現在主流なモデルに対し文字が正しく検出できない主要因は,3D 回転（射影変化）とブラー（焦点ぼけ,移動ぼけ）であることが判明している.

3．既存手法の課題と提案

2.より TextFuseNet は現在最高性能を誇るモデルの一つであるが,射影変化やブラーを含む文字領域の検出が難しいという課題があり,依然精度向上の余地がある.TextFuseNet の原著論文では 1 つの TextFuseNet のみを使用してテキスト領域検

出を行っているが，本研究では複数の TextFuseNet を使用してテキスト領域検出を行う手法を提案する．その理由は，検出が難しいテキスト領域特有の特徴（射影変化やぼけ）に特化させたモデルを組み合わせることで，既存の TextFuseNet では検出の難しいテキスト領域も検出が可能になり，既存手法を上回る精度の達成が期待できるからである．菅原の研究やこれまでの調査で，ぼけや 2 次元，3 次元的な回転がテキスト領域の検出難易度を上げる特徴であることが明らかになっている．したがって，本研究では TextFuseNet の原著論文と同様のオリジナルモデルに加え，ぼやけたテキスト領域に特化したブラー TextFuseNet と 2 次元，3 次元的に回転したテキスト領域の検出に特化した射影変換 TextFuseNet の 3 つを組み合わせてテキスト領域検出を行う手法を提案する．提案手法の構成と訓練プロセスについてそれぞれ図 2，3 に示す．

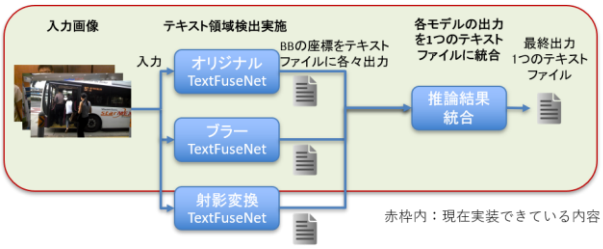

図 2 提案手法推論時の構成

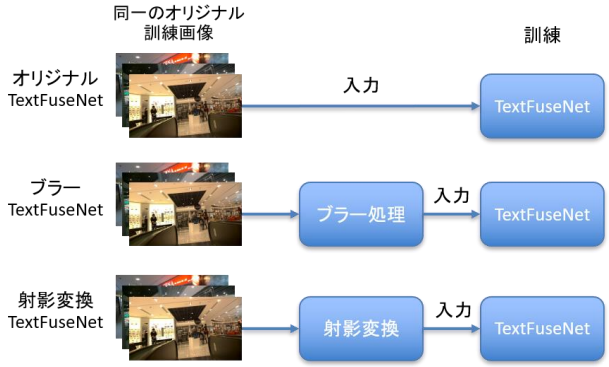

図 3 提案手法の訓練プロセス

図 2 内の「推論結果統合」では統合した BB(Bounding Boxes)の全てが互いに計算した IoU(Intersection over Union)≦0.3 を満たすように 3 つのテキストファイルを 1 つにまとめる．IoU によるしきい値を設ける理由は，提案手法の評価時に同じ正解領域を重複して検出してしまうのを防ぐためである．

4．予備実験

3.で提案した内容の一部である，ぼやけたテキスト領域の検出に特化したモデルと従来のモデルを組み合わせてテキスト領域検出を行うことの有用性を検証するために予備実験を実施した．

4．1　方法

まず，ICDAR2015 にある 1000 枚の訓練データでそのまま訓練した TextFuseNet（以下，オリジナル TextFuseNet）と，カーネルサイズ 5×5 のガウシアンブラーを施した訓練データで訓練した TextFuseNet（以下，ブラーTextFuseNet）を用意する．次に，オリジナル TextFuseNet とブラー TextFuseNet の両方に同一のテスト画像を入力して推論を行った後，両方の推論結果を一つにまとめる．

4．2　実験結果と考察

オリジナル TextFuseNet の推論結果とブラー TextFuseNet の推論結果，そして前者 2 つを融合したものの 3 つを比較したものを以下の表 1 に示す．

表 1 3 モデルの比較

評価指標	original	blur	original+blur
precision	0.912	0.912	0.898
recall	0.390	0.373	0.415
F_1	0.546	0.530	0.568

表 1 より，2 つの TextFuseNet の推論結果を融合したとき，recall が最も高い．これは検出できなかったテキスト領域が最も少ないことを意味する．またF_1も最も高い．したがって，通常のモデルとブラーに特化したモデルを統合することには一定の有効性があると考えられる．

5．参考文献

1. 菅原 啓史，『自然画像中の文字認識技術の定量的性能評価に関する研究』，東京工科大学コンピュータサイエンス学部，令和 3 年度卒業論文．
2. Ye, Jian, Chen, Zhe, Liu, Juhua, Du, Bo.(2020). TextFuseNet: Scene Text Detection with Richer Fused Features.516-522.