2P-09

自然画像中のテキスト領域検出手法の課題抽出とその解決

櫻庭　天†　　青木　輝勝†
† 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

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．この技術は，文字認識技術と組み合わせることにより，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有用性の高い技術である．

本稿では，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，既存手法の問題点を明確化するとともにそれらの問題の解決方策について議論する．

図１　入力画像（左図）と
テキスト領域検出結果（右図）の例

2.関連研究

これまでにテキスト領域検出における既存手法として様々な手法が提案されている．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．

2020 年 に Ye らが提案した TextFuseNet[1] も物体検出に基づく手法であり，現在この分野で最高性能を示す手法の一つとして広く知られている．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 値をマークしている．また，当研究室の菅原が 2021 年に発表した結果[2]によると，テキスト領域検出＋文字認識という一連の処理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.予備実験

菅原の主張がテキスト領域検出にも当てはまるかを検証するために，予備実験を行った．

TextFuseNet の原著論文モデルに対し，ICDAR2015 テストデータ [1] の全画像に施した以下の３つの処理ごとに評価を行った（表 1）．

A)　一切の加工を施さない．
B)　5 ピクセル四方のカーネルで平滑化を施す．
C)　1 枚ずつ左右交互に 45° の射影変換を施す [2]．

表１　TextFuseNet の原著論文モデルにおける
テスト結果

テストデータ形式	P[%][3]	R[%][3]	F[%][3]
A	93.9	90.5	92.2
B	96.8	68.8	80.5
C	96.5	81.5	88.3

4.TextFuseNet の課題

表 1より，平滑化，射影変換を施したそれぞれのテストデータにおいてテキスト領域検出精度が低下したことがわかる．したがって，TextFuseNet にとってブラーや 3D 回転はテキスト領域検出の難易度が上がる要因であることが明らかとなった．

Clarification of Open Issues for State-of-the-art Scene Text Detection Methods and Its Solution
† Sakuraba Ten, Aoki Terumasa, Graduate School of Bionics, Computer and Media Science, Tokyo University of Technology

[1] 合計 500 枚の自然画像からなる．
[2] 射影変換の軸は画像の縦軸に平行な軸である．
[3] P, R, F はそれぞれ Precision, Recall, F 値を意味する．

5.提案手法

前章で述べた知見を踏まえ，本研究では既存手法と，ブラーまたは 3D 回転があるテキスト領域の検出に特化した手法を組み合わせたテキスト領域検出手法を提案する．提案手法では，以下の図 2で示すように，TextFuseNet とブラー特化モデルまたは射影変化特化モデルの 2 モデルから得た推論結果を統合し，最終的な推論結果とする．なお，TextFuseNet は原著論文に示されるモデルを使用する．

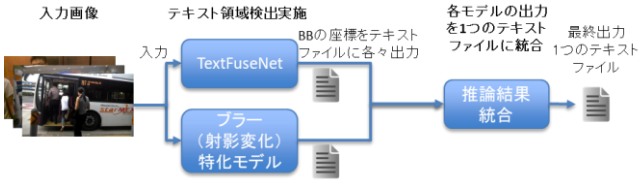

図 2　提案手法の推論時における構成

まず TextFuseNet から推論結果として出力された，テキスト領域を指し示すバウンディングボックス（以下 BB）座標はすべて提案手法の最終的な推論結果の一部となる．続いて，ブラー特化モデルないし射影変化特化モデルから推論結果として出力された BB 座標については，TextFuseNet から出力されたすべての BB との IoU (Intersection over Union) が 50%未満の BB 座標のみ提案手法の最終的な推論結果に送出される．IoU による推論結果のふるい分けを行う理由は，提案手法の評価時に同じ正解領域を重複して検出するのを防ぐためである．

提案手法の訓練方法は，同じテキスト領域検出フレームワーク (TextFuseNet) に異なる訓練画像を入力して訓練を行う点が特徴である．TextFuseNet は訓練画像を無加工のまま入力し，ブラー特化モデルはブラー処理を施した訓練画像を，射影変化特化モデルは射影変換を施した訓練画像を入力して訓練を行う．こうして，3 つの独立したモデルが得られる．

6.実験

本章では前章で述べた提案手法の性能評価実験の結果について述べる．まずブラー特化モデルについて，同じ元画像に対しカーネルサイズ 3, 5, 7 ピクセル四方の平滑化を個別に施した 3 つの訓練データを用いてそれぞれにモデルの訓練を行った．続いて射影変化特化モデルについて，同じ元画像に対し画像の縦軸に平行な軸で 30°，45°，60° での射影変換を個別に施した 3 つの訓練データでそれぞれにモデルの訓練を行った．図 3に射影変換を施した画像の例を示す．

図 3　射影変換前の画像（左図）と
60° の射影変換を施した画像（右図）の例

そして，前章の方法で TextFuseNet の推論結果とブラー特化モデルまたは射影変化特化モデルの推論結果を統合した結果を以下の表 2に示す．なお，TextFuseNet の推論結果は，表 1の”A”と同一である．以下の表 2に実験結果を示す．

表 2　提案手法の実験結果

モデル	T[%][1]	P[%]	R[%]	F[%]
ブラー, 3*3	90.0	92.8	91.1	92.0
ブラー, 5*5	99.0	93.2	91.0	92.1
ブラー, 7*7	99.0	91.7	91.3	91.5
射影変換, 30°	99.0	93.4	90.6	92.0
射影変換, 45°	99.0	93.5	90.6	92.0
射影変換, 60°	99.0	93.3	90.7	92.0

7.結論と今後の課題

表 1および表 2より，TextFuseNet の原著論文モデルに対して提案手法は高い Recall を記録した．これは TextFuseNet の原著論文モデルでは検出できなかった新たなテキスト領域を検出したことを意味する．一方で，F 値は TextFuseNet の原著論文モデルよりわずかながら低下してしまった．今後の課題として，既存手法の F 値を上回るには，Precision の低下量よりも Recall の上昇量を大きくする工夫が必要である．

参考文献

[1] Ye, Jian, Zhe Chen, Juhua Liu and Bo Du, “TextFuseNet: Scene Text Detection with Richer Fused Features,” International Joint Conference on Artificial Intelligence, pp. 516-522, July 2020.

[2] 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

[1] T は Threshold を意味し，確信度が T 以上である BB が推論結果として出力される．