OCR スコアを利用した情景画像内の文字列抽出

上野 将義（金沢大学 大学院自然科学研究科，ueno@blitz.ec.t.kanazawa-u.ac.jp）
南保 英孝（金沢大学 大学院自然科学研究科，nambo@ec.t.kanazawa-u.ac.jp）
木村 春彦（金沢大学 大学院自然科学研究科，kimura@ec.t.kanazawa-u.ac.jp）
上田 芳弘（石川県工業試験場，ueda@irii.go.jp）

Text extraction in natural image by using OCR score

Masayoshi Ueno (Graduate School of Natural Science and Technology, Kanazawa University, Japan)
Hidetaka Nambo (Graduate School of Natural Science and Technology, Kanazawa University, Japan)
Haruhiko Kimura (Graduate School of Natural Science and Technology, Kanazawa University, Japan)
Yoshihiro Ueda (Industrial Research Institute of Ishikawa, Japan)

要約
カメラの発達により、色々な場面での撮影が可能となった。さらにカメラの解像度も高くなったため、カメラで撮影した情景内の文字を認識することも可能となった。我々の周囲には数多くの文字が存在しており、それらは有益な情報をもたらしている。よって、もし情景内の文字情報を自動的に認識することが可能となれば、様々なシステムにおいて役に立つと考えられる。しかし、文字認識のためには文字列の位置を特定する必要があり、困難を伴う。本論文では、連結成分抽出法とCSERを用いて、背景を分離し文字候補を抽出する。さらに、文字候補を絞り込むためにOCRスコアとヒストグラムを用いた。また、OCRスコアを用いることで、既存研究では不可能であった1文字からなる文字領域を抽出することが可能となった。そして、提案手法を用いた実験では、抽出精度74.6 %という結果が得られた。

キーワード
MSER，CSER，文字列抽出，文字認識，OCRスコア

1. はじめに

現在、ビデオカメラ、スマートフォン端末等のカメラ付き携帯機器の普及に伴い、利用者は様々な画像を撮影することが可能である。さらに、カメラの高解像度化に伴い、情景内の文字を認識することが可能になりつつある。私たちの身の回りには文字情報が多く存在し、それらは私たちにとって有益な情報を提供してくれる。したがって、身の回りの情景内に存在する文字情報を認識することができれば、様々なシステムとの連携が可能になると考えられる。

例えば、店舗の名前や住所、地名などが記載されている看板上の文字や、経路情報などの道路交通情報が記載されている文字を読み取り、利便性や交通の安全性向上を目的としたシステムとの連携が可能である。

現在、光学文字認識（OCR）を用いて活字の文書画像をコンピュータが編集できる形式に変換が可能であるが、情景画像内の文字を認識することは困難である。文字を認識するにはまず、文字の画像内での位置を特定する必要があるが、一般的に情景画像内の文字の抽出（認識）を困難にする 要因として以下が挙げられる。

- 文字情報以外のオブジェクト（空、建物、車、人など）が存在する
- 文書画像と異なり文字の背景が複雑である
- 撮影時の状況によって外乱（影や光の反射）を受ける

したがって、情景画像内から文字を正確に取得するためには、上記の問題を克服しなければならず、本研究では、前述したサービスの実現のために、様々な種類の情景画像から文字列を抽出することを目的とする。

2. 既存研究

情景画像内の文字列の抽出に関する手法は大きく分けて2つに分けることができ、パッチベース処理、連結成分ベース（領域ベース）処理がある。パッチベース処理では、画像内で文字列である可能性が高いかどうかを矩形単位で機械学習を用いて判定し、文字列の抽出を行う（Chen and Yuille, 2004; Kim et al., 2003）。しかし、パッチベース処理で得られる抽出結果は背景と文字列の分離ができておらず、抽出した文字列を認識するためにはさらに処理を加える必要がある。

パッチベース処理に対して、連結成分ベース処理では、同一文字のピクセルは類似した特性を持つと仮定し、同一文字の連結成分を利用してピクセルを領域にグループ化することで各文字を抽出している。連結成分ベース処理の利点は、連結成分の濃淡が一般的に文字列の特性（スケール、方向、フォント）に依存しないことである、また、連結成分ベース処理の中でも、文字抽出の際にMaximally Stable Extremal Regions（MSER）をベースとした手法が有効である（Chen et al., 2011; Neumann and Matas, 2012; Yin et al., 2014）。しかし、MSERは多くの領域を検出してしまうことが問題である。そこで、MSERをベースとしたCSER（Class Specific Extremal Regions）では、簡易的に文字、非文字の分類を行っている（Neumann and Matas, 2012）。CSERでは、連結成分を2値化した際に適切なERsを抽出するために、文字認識にも有効に利用できる利点がある。しかし、文字、非文字の分類を行っていながらも、それでもなお抽出される文字候補の数は多く、その後の文字

列生成を困難にする要因となっている。既存研究では、与えられた画像からグレイスケールに変換し、画像1枚から抽出を行っているため、外乱（影や光）の影響を受けると文字列を正しく抽出できない（Yin et al., 2014）。さらに、情景内の単一の文字は既存研究では考慮されていない。

3. データセット

本研究で用いるデータセットは、ICDAR2003で利用されたデータセットの合計251枚である（Lucas et al., 2003）。データセットの画像の詳細は以下の通りである。

- 画像サイズ：307×93から1280×960
- 対象文字：英数字のみ
- カラー画像のみ
- 文字数：最低1文字以上画像内に存在

4. 提案手法

本研究では、背景と文字が分離できているという利点と簡易的に文字分類を行っているCSERを用いて文字候補を抽出する。そして、得られた領域にOCRを利用してOCRスコアを取得し、それを利用して文字候補の削減を行った後、文字列を抽出する手法を提案する。またOCRスコアを利用することで既存研究では行われていなかった単一文字の抽出も可能であると考えられる。本研究の流れは、大きく分けて1.文字候補の抽出、2.文字候補の削減、3.文字列の抽出である。図1に各段階の処理結果を示す。

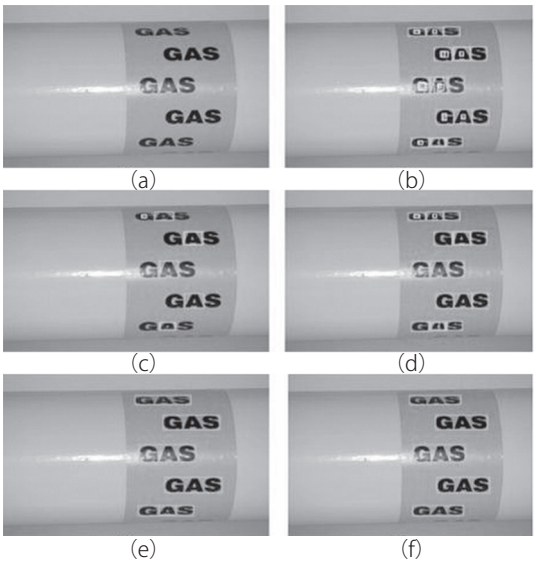


図1：各段階の処理結果

注：(a) 元画像、(b) 全文字候補、(c) OCRスコア適用後、(d) 類似画像抽出、(e) 文字列生成、(f) 文字列統合

4.1 文字列候補の抽出

文字候補の抽出は以下の手順で行う。

- カラー画像をRGBそれぞれの要素に分割する
- MSERを用いて、画像内から安定した領域を抽出する
- 抽出した領域に対してメディアンフィルタを用いて平滑化処理を行う
- CSERを用いて文字候補を抽出する

4.1.1 MSER

MSERは、Matas et al.（2004）で提案された領域分割の手法であり、画像中の輝度値が類似した画素を1つの領域にまとめていく手法である。抽出された領域は周りの画素値と比較して明るい、または暗い領域である。MSERはグレイスケール画像に対して閾値を徐々に変化させることで領域を抽出する。MSERは以下の手順で処理を行う。

- 濃淡画像から、閾値を徐々に変化させ連続する2値画像を生成する
- 各2値画像の連結領域（Extremal Regions）を求める
- 面積の変化が最も緩やか（Maximally Stable）な連結領域を特徴領域とする

4.1.2 CSER

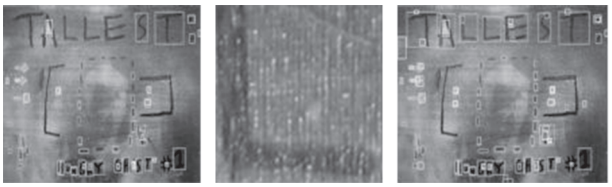
CSERはNeumannらによって提案された手法である。基本的な考え方は、適切なExtremal Regions（ERs）を画像の全コンポーネント木から選択する点でMSERに似ている。しかし、CSERでは文字検出の分類学習を利用することで適切なERsを選択する点でMSERと異なる。したがって、MSERによって抽出された安定した領域が必ずしも選択されるわけではない。CSERではグレイスケール画像を利用するが、本研究ではRGB各要素[1]とそれらを反転させた画像[2]（計6枚）を利用し、[1], [2]の抽出結果をまとめ、計2枚の抽出結果を取得した。なお、グループ化する際には文字候補同士が80％以上重なっていれば、同一領域とみなした。

4.2 問題点

CSERを利用して文字領域の抽出を行った結果を図2(a)に示す。なお、青枠で囲んだ結果はグレイスケール画像をそのままCSERで検出した領域で、黄色枠で囲んだ結果は、それらのグレイスケールの画素値を反転させた画像から得られた領域である。以降に記載する画像に関しても同様である。図2(a)の"TALLE"の文字はMSERでは検出されていたが、CSERの結果では抽出できなくなっている。

図2(a)の抽出できなかった文字を切り取り拡大すると図2(b)のような画像が得られた。図2(b)より、黒字の上に光や文字が書かれている材質の影響で白い部分が点のように見える部分が存在する事が分かる。この部分が影響し、2値化した際に一つの領域としてとらえることができずに、CSERで非文字に分類されたと考えられる。

そこで、画像全体にメディアンフィルタを適用して輝度値を平滑化することにより、CSERで先ほど抽出できなかった領域が抽出できた（図2(c)）。しかし、メディアンフィルタで安定した領域を作った分、CSER後の文字抽出数が多くなってしまうという問題が出てきた。そこで、MSERで抽出された領域に対してのみメディアンフィルタを用いて平滑化すること



(a)　(b)　(c)

図2：メディアンフィルタによる平滑化
注：(a) メディアンフィルタ不使用、(b) 拡大画像、(c) メディアンフィルタ使用

により、画像全体ではなく部分的に平滑化を行い文字抽出数の増加を抑制させた。

4.3 文字列候補の削減

CSERを用いて抽出された文字候補では候補数が多いために文字列の生成を困難にする。したがって、より信頼性の高い文字候補のみを残すことで文字列の生成を容易にすることを試みる。文字候補の削減はOCRスコアと2 つの画像間のヒストグラムを比較することで行う。

以下にその流れを示す。

- 抽出した画像に対してOCRスコアを計算する
- OCRスコアを閾値として、閾値以上の文字候補のみを抽出する
- 抽出された文字候補を基準として、水平方向に存在した閾値以下の文字候補との類似度を計算する
- 類似性があると判断されれば文字候補として抽出する

4.3.1 OCR スコアの利用

オープンソースのソフトウェアであるTesseract-OCRを利用した（Tesseract-OCR, 2015）。また、事前の実験により同じ画像でも特に画像サイズが大きい場合に認識精度が悪くなるため、120 × 80のサイズに正規化を行った。また、正規化するサイズより小さい画像は拡大することで逆に誤認識したため、処理を行わなかった。そして、本来文字領域であっても閾値未満となる可能性があるので、水平方向に存在した文字候補とのヒストグラムの類似度を算出し、類似していれば文字とした。また、全ての文字候補が閾値未満の場合は、画像内から最低1文字は抽出されるように、スコア値が最も高い候補を抽出した。

4.3.1 類似画像の抽出

ヒストグラムの類似度の算出には、バタチャリヤ距離を用いた。バタチャリヤ距離とは、二つの分布を独立事象とみなした時のそれらの同時確率に対する自己情報量として定義される。OCRスコアが閾値以上であった文字候補と水平方向に存在する文字候補との垂直方向の割合rを求め、そのrに応じてヒストグラムの類似度sの閾値（$\alpha 1 \sim \alpha 4$）を変更し、以下の条件を満たせば文字候補とした。

- $r \geq 0.9$ かつ $s < \alpha 1$
- $0.7 \leq r < 0.9$ かつ $s \leq \alpha 2$
- $0.6 \leq r < 0.7$ かつ $s \leq \alpha 3$
- $0.5 \leq r < 0.6$ かつ $s \leq \alpha 4$

また、$\alpha 1$から$\alpha 4$になるにしたがって閾値の値は小さく設定した。

4.4 文字列の抽出

4.4.1 文字列の作成

本研究では、英数字文字のみを対象としているため、文字列は水平方向に存在する。また、同一文字列内の背景色bcまたは文字色ccは同色であると考えられる。したがって、ある文字候補から水平方向の文字候補を探索し、見つかった文字候補の背景色または文字色が類似していれば同一文字列とした。k平均法（k = 2）を利用し、あらかじめ抽出候補の画素値を取得し、それぞれの重心と比較することで背景色と文字色を決定した。k平均法では初期値設定によって異なる結果が得られることがわかっているが、本研究ではk-means++法を利用した（Arthur, 2007）。なお、文字色と背景色を算出した際に重心の距離を算出し、ほぼ一致する文字候補に関しては同一の文字候補であると判断し、候補から削除した。文字列の作成条件を以下に示す。（$\beta 1 \sim \beta 4$）は閾値である。文字列の作成では、あらかじめ水平方向の文字との垂直方向の重なり具合（割合r）を求め、それを基に以下の条件を適用した。また、文字列内に文字が2文字存在する場合は、以下の条件に加え文字同士の距離を計算し、文字の高さより小さければ文字列とした。

- $r \geq 0.9$ かつ $bc < \beta 1$
- $0.5 \leq r < 0.6$ かつ $bc < \beta 2$
- $0.5 \leq r < 0.6$ かつ $cc \leq \beta 3$ かつ $bc \leq \beta 3$
- $0.5 \leq r < 0.6$ かつ $cc \leq \beta 4$ かつ $bc \leq \beta 4$

$\beta 4$になるにしたがって値は小さくなるように設定した。これは文字候補が多くなった場合に、縦方向の重なる割合が低くなると選択される文字候補も多くなり、誤ってグループ化することを防ぐためである。

4.4.2 文字列の統合または削除

文字列として作成した文字列内にさらに文字列が作成されていたり、文字列内に文字が抽出されたりする。これはAやRなどの文字ではAの中にある三角の領域や、Rの中にあるDのように見える領域が抽出されるためである。そこで、重なっている文字列または文字を統合または削除するようにした。以下にその処理の流れ、統合条件を示す。ある文字列Tの高さ、幅をそれぞれheight (T)、width (T)と考える。

- 文字列Aと重なっている文字列BのAにおける位置を計算する
- Aを3等分したときの中間にBが存在（または共有）しない場合、Bは独立しているとみなす
- B が A の中間を共有する場合、Bを削除またはAとBを統

合する

- 統合または削除の条件
  - height（B）≤ height（A）の場合、統合した場合にwidth（A）の拡大する長さを計算し、Aに存在する文字3個分の長さ以下ならば統合、そうでなければBを削除
  - height（B）≥ height（A）の場合は統合
  - BがAの領域の中に完全に存在する場合はBを削除

4.4.3 単一文字の抽出または削除

文字列を生成しなかった文字候補は単一文字と考えることができる。しかし、それらをすべて残した場合、非文字が文字として抽出される（特に文字の場合誤認識）ことが多かった。したがって、今回は数字と認識された文字候補のみを単一文字として抽出した。

5. 実験

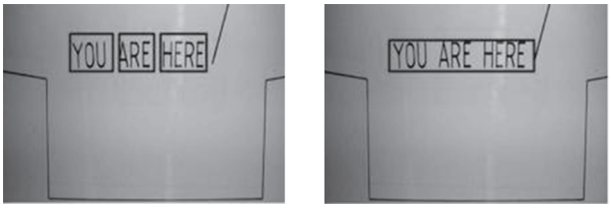
本研究では、OpenCV3.0（OpenCV, 2015）を利用した。また、CPUがCore i5 1.80 GHzのPCを用いた。

5.1 評価方法

データセット内に正解領域の座標を示したファイルが存在する。正解データでは単語ごとに文字領域が記されており、図示すると図3(a)のようになる。しかし、本手法で対象とする領域は単語ごとではなく、同一直線上にある場合は1 つの領域として抽出を行うため、図 3(b) のように正解データの修正を行った。評価方法には、再現率、適合率、F 値を用いた。

- 再現率（Recall）
  画像中の正解データの領域を画素数A、抽出した領域でかつ正しい領域の画素数をBとすると、以下の式で定義される。

$$\text{Recall} = A / B$$



（a）　　（b）

図3：正解データの修正
注：（a）修正前、（b）修正後

図4：文字抽出結果

- 適合率（Precision）
  本実験で抽出した文字列領域の画素数をCとすると、以下の式で定義される。

$$\text{Precision} = B / C$$

- F値
  再現率と適合率の調和平均であり、再現率をR、適合率をPとすると、以下の式で定義される。

$$\text{F-Measure} = 2RP / (R + P)$$

5.2 実験結果

評価結果を表1に示す。表中のYun et al.（2014）の結果は、本研究と同じ評価方法で行ったものである。その他の手法（Pan et al., 2011; Lee et al., 2011; Epshtein et al., 2010）の結果は単語ごとの抽出精度に関するものであるが、再現率に関してはほぼ変化がないと考えられるため、その結果から有効性を確認できる。

6. 考察

6.1 背景と文字が同色の画像

図5は文字色が背景色と同色のため、文字領域を抽出できなかった例である。MSERを利用しても文字色と背景色が同じ場合は閾値によって2値化した場合に同じタイミングで変化するため抽出ができないと考えられる。

図5：背景と文字が同色の場合

表1：実験結果

	Recall（%）	Precision（%）	F-measure（%）
Proposed method	75.8	73.4	74.6
Yin et al., 2014	69.5	77.1	73.1
Pan et al., 2011	68.0	67.0	67.0
Lee et al., 2011	66.0	75.0	70.0
Epshtein et al. 2010	73.0	60.0	66.0

6.2 文字削減の際に誤って文字を削除した画像

図6(c)は、ヒストグラムの類似度を比較したときに誤って削除した例である。左から抽出した全文字候補、OCRスコアで文字削減を行った後、ヒストグラムの比較後の画像である。OCRスコアで文字削減を行った際に、本来文字領域である候補のスコアが低かったため誤削除が生じている。図6(b)では、抽出した全文字候補のスコアが閾値より低かったため、スコアが最も高い文字が1個抽出されている。しかし、1個の場合、今回のように文字以外の候補を選択する可能性が高くなるため、適切な個数を抽出するための改善が必要である。

(a) (b) (c)

図6：文字候補の誤削除の例

注：(a) 文字候補、(b) OCRスコア適用後、(c) ヒストグラム比較後

6.3 適合率に関して

本研究では、OCRスコアを利用したため、ある程度文字候補が多くても非文字列を抽出した数を減らすことができた。しかし、画像によっては図7のように特に窓やレンガなどの規則性のある個所を文字列と誤判別して抽出してしまうため、文字部分は正しく抽出できているが、適合率が著しく悪くなる画像が存在した。したがって、抽出した文字列が実際に文字列かどうかを判定する処理を加える必要がある。

図7：適合率が悪くなる画像の例

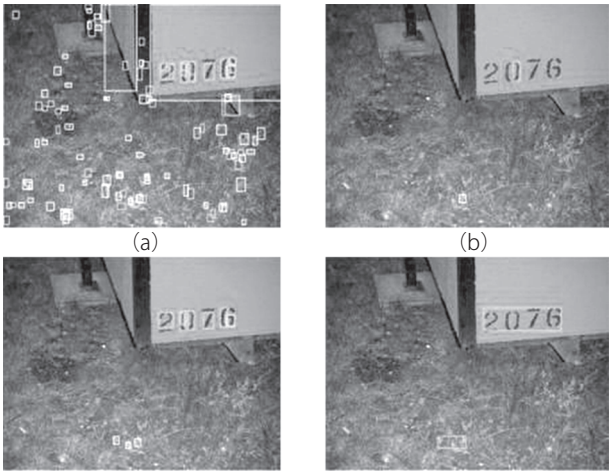
6.4 単一文字の抽出

本研究では、OCRスコアを利用したため、文字列のみを対象としていた従来手法では対象外となっていた単一文字の抽出を行うことができた。今回は適合率の関係から数字に限定して抽出を行ったため、本来英字で抽出ができていた文字を棄却した画像が数枚あった。したがって、今後は数字のみではなく英字を抽出できるように工夫する必要がある。

図8：単一文字の抽出例

6.5 同一文字の要素が離れている画像

MSERやCSERでは、文字が連結している場合は1つの領域として抽出が可能であるため、英数字に対しての文字抽出は有効であるが、平仮名の「い」や「う」のように離れている場合には有効ではないと考えられる。本研究では英数字を対象としているが、図9のように英字でも要素が離れている画像が存在した。しかし、本手法ではOCRスコアを利用して抽出した文字候補から水平方向に類似したヒストグラムを持つ候補を文字候補として採用することができるため、図9(d)に示すように、文字の構成要素が離れていても抽出が可能である。



(a) (b)

(c) (d)

図9：非連結文字抽出の成功例

注：(a) 全文字候補、(b) OCRスコア適用後、(c) 類似画像検出後、(d) 文字抽出後

7. まとめ

現在、カメラ付き携帯端末の普及に伴い利用者は様々な画像を撮影することが可能である。文書画像内の文字を認識することは可能であるが、情景画像内の文字を認識することは、背景の複雑さや外乱の影響を受けるため、一般的に困難である。本研究では、同一文字のピクセルは類似した特性を持つことから連結成分ベース処理を利用した。また、MSERをベースとした手法であるCSERを利用した後、文字候補の削減を目的にOCRスコアを利用した手法を提案した。その結果、文字列に対しての抽出精度はF値で74.6 %の結果が得られた。またOCRスコアを利用することで、今回は数字のみに限定したが、単一文字に対しての抽出も可能になった。今後は図6

のような文字削減の失敗画像を改善することや抽出した文字列が文字列かどうかの分類を行うことで適合率の向上が見込めると考えられる。

引用文献

Arthur, D. (2007). k-menas++: The advantages of careful seeding. *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithm*, 1027-1035.

Chen, H., Tsai, S., Schroth, G., Chen, D., Grzeszczuk, R., and Girod, B. (2011). Robust text detection in natural images with edge-enhanced maximally stable extremal regions. *Proceedings of the IEEE International Conference on Image Processing*, 2609-2612.

Chen, X. and Yuille, A. (2004). Detecting and reading text in natural scenes. *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vison and Pattern Recognition*, Vol. 2, 366-373.

Epshtein, B., Ofek, E., and Wexler, Y. (2010). Detecting text in natural scenes with stroke width transform. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR2010)*, 2963-2970.

Kim, K., Jung, K., and Kim, J. (2003). Texture-base approach for text detection in images using support vector machines and continuously adaptive mean shift algorithm. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, Vol. 25, No.12, 1631-1639.

Lee, J., Lee, P., Lee, S., Yuille, A., and Koch, C. (2011). AdaBoost for text detection in natural scene. *International Conference on Document Analysis and Recognition (ICDAR2011)*, 429-434.

Lucas, S. M., Paneretos, A., Sosa, L., Tang, A., Wong, S., and Young, R. (2003). ICDAR2003 robustcompetitions. *International Conference on Document Analysis and Recognition (ICDAR2003)*, 682-687.

Matas, J., Chum, O., Urban, M., and Pajdla, T. (2004). Robust wide baseline stereo from maximally stable extremal regions. *Image and Vison Computing*, Vol. 22, No. 10, 761-767.

Neumann, L. and Matas, J. (2012). Real-time scene text localization and recognition. *2012 IEEE Conference on Computer Vison and Pattern Recognition*, 3538-3545.

OpenCV (2015). http://opencv.org/（2015/1/15 アクセス）.

Pan, Y., Ahu, Y., Sun, J., and Naoi, S. (2011). Improving scene text detection by scale-adaptive segmentation and weighted CRF verification. *International Conference on Document Analysis and Recognition (ICDAR2011)*, 759-763.

Tesseract-OCR–Google Code (2015). https://code.google.com/p/tesseract-ocr/（2015/1/15 アクセス）.

Yin, X. C., Yin, X., Huang, K., and Hao, H. W. (2014). Robust text detection in natural scene images. *IEEE Transaction on Pattern Analysis and Machine Intelligence*. Vol. 36, No. 5. 970-983.

（受稿：2016年5月16日　受理：2016年5月30日）