一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
LOIS2019-66(2020-03)

光学的文字認識を活用した植物標本画像のラベル自動マスキング方法の検討

日野 遥　　檜垣 泰彦†

千葉大学工学部 〒263-8522 千葉県千葉市稲毛区弥生町 1-33
† 千葉大学アカデミック・リンク・センター 〒263-8522 千葉県千葉市稲毛区弥生町 1-33
E-mail:　† higaki.yasuhiko@faculty.chiba-u.jp

あらまし　故萩庭氏らが収集した萩庭植物標本は，収録数・標本の採集地の両面において国内最大規模であり，デジタル・スカラシップ開発の一環で運用を開始した千葉大学学術リソースコレクション（c-arc）での公開が予定されている。しかし絶滅危惧種の収録や，活字と手書き文字が混在した詳細な採集地の記載などがみられ，IIIF・Right Statements を採用するシステムで公開するには資源保護の観点で懸念がある。そこで今回は，標本画像中に記載された採集地の部分を自動的に検出・マスキングする方法に関して，光学的文字認識（Optical Character Recognition）を中心に検討した。検討の結果，OCR エンジンは Cloud Vision API を利用することとし，画像を含めたパラメータについても絞り込むことができた。マスキング範囲の決定に関しては若干の調整は必要であるが，マスキング処理を自動的に行うめどが立った。

キーワード　植物標本画像，光学的文字認識(OCR)，自動マスキング，デジタル・スカラシップ開発

Investigating of automatic label masking methods for plant specimen images using Optical Character Recognition

Haruka Hino†　and　Yasuhiko Higaki†

Faculty of Engineering, Chiba University　1-33 Yayoi, Inage-ku, Chiba, 263-8522 Japan
† Academic Link Center, Chiba University　1-33 Yayoi, Inage-ku, Chiba, 263-8522 Japan
E-mail:　† higaki.yasuhiko@faculty.chiba-u.jp

Abstract　The Haginiwa plant specimen images, are one of the largest collection of flowering plants in Japan, are expected to be published at Chiba University Academic Resource Collections (c-arc) that adopts IIIF and Right Statements. However, there is concern about resource conservation that endangered species are recorded in the collection, and detailed collection sites are described. In this study, we focused on optical character recognition (Optical Character Recognition), which is a method for automatically detecting and masking the collection area described in the sample image. We mainly use “Cloud Vision API” for OCR, and refined parameters including images itself. As a result of study, the masking area determination algorithm needs some adjustment, but we obtained the prospect of performing masking automatically.

Keywords　Plant Specimen Images, Optical Character Recognition(OCR), Automatic Masking for Images, Digital Scholarship Development

1. はじめに

1.1. 研究背景

萩庭植物標本画像データベースは，千葉大学名誉教授であった故萩庭氏らが 1895～2003 年にかけて国内外で採集したさく葉標本を基に，あのはな山岳会を中心に結成された萩庭植物標本データベース作成協力会が整備したものである。本データベースには 51,819 点の標本画像が含まれており，顕花植物の標本では国内最大規模とされる[1]。

萩庭植物標本データベースに収録された標本画像

図 1 千葉大学学術リソースコレクション(c-arc)の画面例

等は，千葉大学学術リソースコレクション（c-arc[1]、画面例は図 1）の公開が予定されている。これはデジタル･スカラシップ開発の一環で整備されたシステムで，国際的な画像相互運用規格 IIIF と Right Statements を採用している[2]。このことから研究資源としての利活用の機会が広げられ，不明種の同定や科名・学名の不一致の修正など研究の進展が期待されている。

1.2. 研究目的

しかし，現状のままでの公開には資源保護の観点で懸念がある。具体的には，本データベースには多数（8,488 件が 2019 年環境省発行のレッドデータブック掲載種[3]に該当）の絶滅危惧種が含まれているほか，標本画像には詳細な採集地が記載されているものが散見される。

そこで本研究では標本画像中に記載された採集地を検出し，自動的にマスキングを行う方法の検討を目的とした。

2. 下準備

2.1. 用語の定義

本研究では図 2 のように標本画像中にある標本情報（科名・学名・和名・採集地・採集日など）が記載されている領域を「標本ラベル」と呼称する。図 2 では詳細な採集地に相当する箇所はモザイク加工を施している。

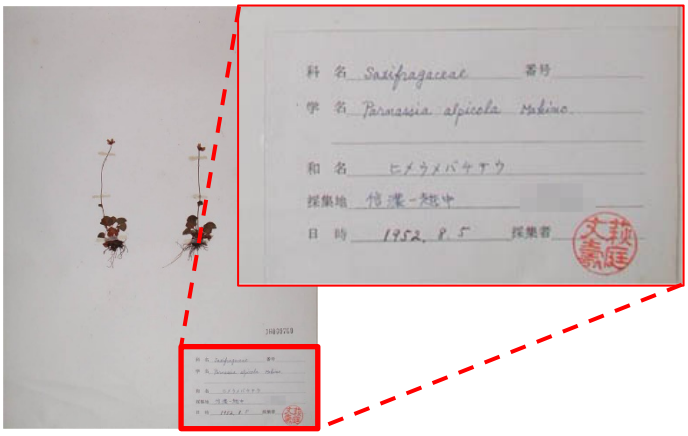

図 2 標本画像と標本ラベルの拡大図

また，標本ラベルの見出し部分が図 2 のように「日本語であるものを「日本語ラベル」，英語であるものを「英語ラベル」と呼称する。英語ラベルの一例を図 3 に示す。

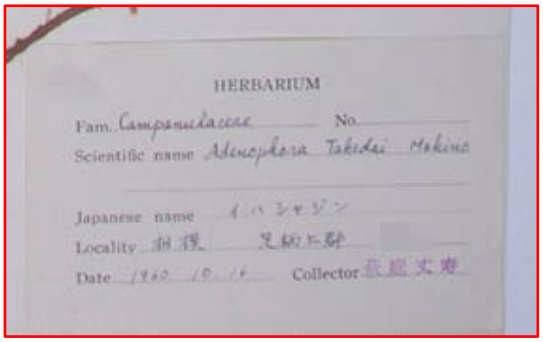

図 3 英語ラベルの例

2.2. サンプリングの実施

本研究は検討段階にあるためサンプリング調査で実施した。目視ベースで標本ラベルの形式毎に標本画像を分類し，採集地の表記がないもの・大きく標本で隠れたもの等自動処理が困難と思われるものを除外の上，4 グループの異なる標本ラベルの集合（日本語ラベル･英語ラベルでそれぞれ 2 グループ）を作成した。そしてグループごとに ID をベースにした単純ランダムサンプリングを行い，以降の調査を実施している。

3. OpenCV による矩形認識の検討

オープンソースの画像処理ライブラリ OpenCV[2]に含まれる，画像中の領域を認識できる矩形認識を用いて標本ラベルの検出・切り出しが可能であるか調べた。通常の画像から矩形認識を行う場合，画像を二値化する必要があるが，今回は大津の二値化[4]と適応的閾値処理 2 種・Canny 法によるエッジ検出[5]を試した。実装には OpenCV の Python ラッパーである opencv-python[3]ならびに拡張パッケージの opencv-python-contrib を利用した。

結果としては，一部の画像で Canny 法によるエッジ検出および大津の二値化を利用した場合を中心に，標本ラベルの周辺や文字の輪郭の検出とみられる矩形が確認できたものの，標本本体の認識が多く期待された形で標本ラベルが検出できたとは言えないものであった。

また，画像調整（統計的正規化・ヒストグラム平坦化及び適用的ヒストグラム平坦化）と閾値処理をそれぞれ組み合わせた場合でも，矩形認識の傾向に若干の変化は見られたものの，大津の二値化を利用した場合を中心にノイズ様の認識結果が増加するなど，期待した形での標本ラベルの認識には至らなかった。

[1] https://iiif.ll.chiba-u.jp/

[2] https://opencv.org/
(リポジトリは https://github.com/opencv/opencv)

[3] https://github.com/skvark/opencv-python

4. 光学的文字認識（OCR）の利用方法の検討

画像から文字情報（文字コード）及び画像中の位置を得る技術として光学的文字認識(Optical Character Recognition、OCR)がある。今回はこれを利用し，画像中に記載された採集地の位置の検出を試みた。

4.1. OCR エンジンの検討

OCR を提供するプラットフォームは 2020 年現在多数あるが，今回は Google 提供の Cloud Vision API[4]とオープンソースの OCR エンジン Tesseract OCR[5]を候補として選択した。

しかし，実際の標本画像を用いて OCR を行った際の両者の認識精度には大きな隔たりがみられた。にその一例を示した。Cloud Vision API はパラメータの変更により認識結果が変化・認識文字数が少ない場合でも一定の認識結果が得られる傾向にあった。一方で，Tesseract OCR では大半の画像で文字列が取得できないという結果となった。表 1 に例を示す。これは後述の OCR 時のパラメータ（Language Hint)を変更しても改善は見られなかった。

表 1 Tesseract OCR と Cloud Vision API による認識結果の比較

Tesseract OCR	Cloud Vision API
‘HERBARIUM eit Gmpensiond Pea ne dln ade Jopanive ame 408 He | sHo2s479 ents ata, Reem sel Dine /162 18.4 − Coleco Bt ME ‘4 ‘	2 HERBARIUM Pam Campanulace - No Scientific name Adenophera Takedai Makine TH023979 Japanese samt Locality at all. Date 1960. 10 anvera 2ort to at Collector AS

なお，Tesseract OCR については LSTM(Long Short-Term Memory)を用いて学習を行った独自モデルも利用可能である[6]が，今回は UB Mannheim の Windows 版[7]に付属の日本語モデルを利用している。独自にモデル作成した場合はまた異なった結果が得られた可能性はあるものの，独自モデルの構築には相応の時間を要するということもあり，本研究の目的達成の観点では Cloud Vision API の利用が適すると判断した。

4.2. OCR のパラメータ検討

OCR 利用時のパラメータ（バージョン・Language Hint）の変更による認識結果の変化を見た.

2020 年 1 月現在 Cloud Vision API では α 版 1 つを含め 4 つのバージョンが選択可能[7]であり，今回はそのうち v1・v1p3beta を用いて結果の変化を見たが，すべてのパターンで認識結果にバージョン間での差異を認められなかった。このため，以降の検討はチュートリアル[9]に従い，v1p3beta を使用した。

もう一つのパラメータ Language Hint は，OCR 実行時に指定できる画像に記載された言語のヒントである。今回は日本語・英語・ラテン語系（後者 2 つは手書き文字対応)をそれぞれ指定し，認識結果の変化を見た。次のグラフ(図 4・図 5)は日本語ラベル・英語ラベルそれぞれにおいて，Language Hint が日本語・英語それぞれの場合の認識文字数を，10 文字毎の階級に分けた上でグラフ化したものである。

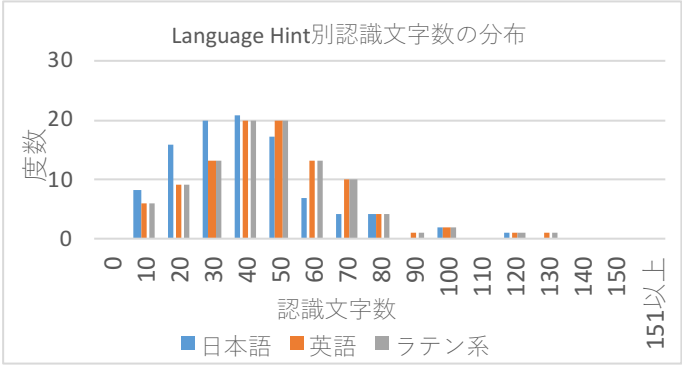

図 4 日本語ラベルの場合の Language Hint 別認識文字数の分布

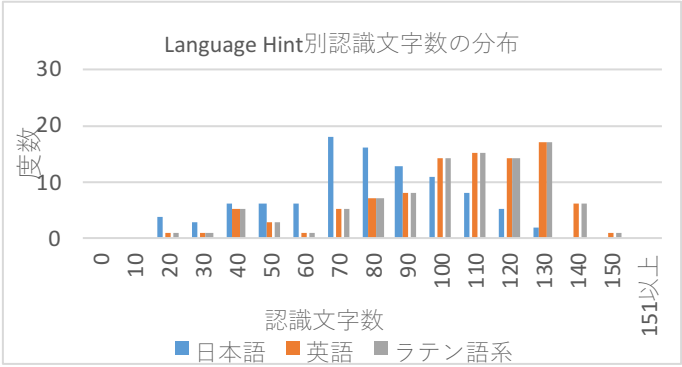

図 5 英語ラベルの場合の Language Hint 別認識文字数の分布

図 4・図 5 に示したように，Language Hint が英語の場合とラテン語系の場合の認識結果(認識文字数の分布)については，すべてのサンプルグループで一致した。

また，全体的な傾向として Language Hint に日本語を指定した場合よりも，英語・ラテン語系を指定した場合のほうが認識文字数は多くなる傾向にあった。日本語ラベルにおける，Language Hint が日本語の場合と英語の場合の認識文字数の関係を散布図としてあらわしたものが図 6 である。

[4] https://cloud.google.com/vision?hl=ja

[5] https://github.com/tesseract-ocr/tesseract

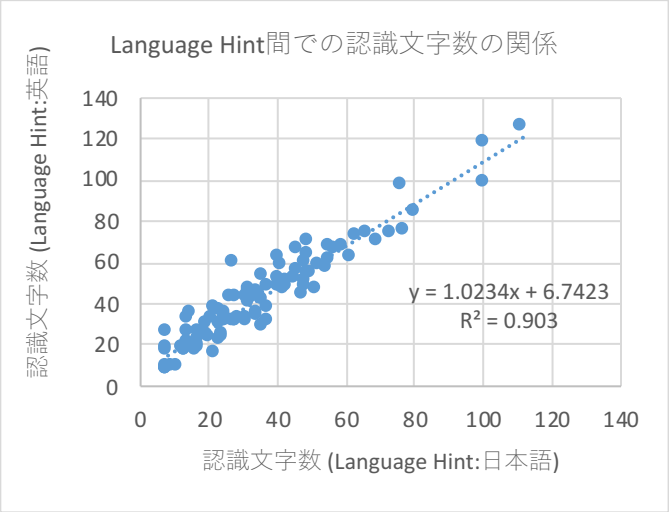

図 6 日本語ラベルにおける Language Hint 間（英語・日本語）の認識文字数の関係

ここで得られた相関係数 r （今回のサンプルでは 0.91～0.95），危険率 0.95 の無相関の検定にかけたところ，いずれのサンプルにおいても相関係数 r から得られる検定統計量 t の確率分布は棄却域にあった。検定統計量 t は n,r を用いて次の式により算出される。

$$t = \frac{|r|\sqrt{(n-2)}}{\sqrt{(1-r^2)}}$$

したがって，Lanugage Hint の指定が日本語の場合と英語の場合のそれぞれの認識文字数の間には強い正の相関があり，回帰式と分布からは英語(・ラテン語系)を指定する方が多くの文字を認識できる傾向があるといえる。

認識内容に関しては，活字・手書きともに日本語で記載された部分（例:採集地・採集者）よりも英語等で記載された部分（例:科名・学名）の方が認識される傾向が見られた。これらの結果と科名にラテン語が含まれうる点を踏まえ，Language Hint はラテン語系手書き文字を意味する mul-Latn-t-i0-handwrit を指定することとした。

4.3. 入力画像の検討

続いて画像もパラメータの一種と考え，画像調整（ヒストグラム平坦化・適用的ヒストグラム平坦化、実行時のパラメータは矩形認識の際に試したものと同じ）の有無及びカラー画像（元の標本画像と同等）とグレースケール（OpenCV での読み込み時にグレースケールを指定）を利用した場合の認識結果の変化をみた。

カラー画像とグレースケール画像に関しては，認識文字数に関して，カラー画像を用いた方が多い場合とその逆がおよそ 1:1 の割合で観測され，どちらが優位であるか判断できない結果となった。一方，ヒストグラム平坦化・適用的ヒストグラム平坦化を施した場合に関しては，に示すように適用的ヒストグラム平坦化を施した場合の認識文字数が多くなる傾向がみられた。また，カラー画像に対しても認識文字数がやや増加する傾向が確認された。

このため，入力に用いる画像については「カラー画像」「グレースケール画像」「適用的ヒストグラム平坦化を施した画像」の 3 種類を用いることとした。

5. マスキング範囲の決定

OCR の結果を踏まえ，マスキング範囲の決定を行うこととした．採集地の直接の検出は難しい傾向にあるため，OCR で取得できた文字から標本ラベルごとの差異を考慮したうえで採集地の位置を推定するアルゴリズムを作成した。マスキングを施した標本画像の例（全体と標本ラベルの拡大）を図 7 に示す。

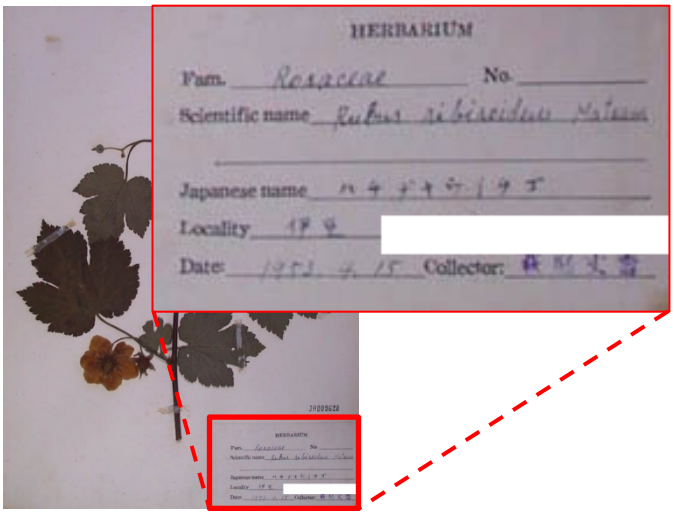

図 7 マスキングを施した標本画像の例（英語ラベル・右上は標本ラベル拡大）

今回，入力画像にはカラーの標本画像のほか，グレースケール画像・適用的ヒストグラム平坦化を施した画像の 3 種類を用いており，OCR 時の Language Hint はラテン語系を指定している。1 枚の標本画像に対して平均 3.6 枚のマスキング画像が得られたため，それらを目視確認の上，「1.採集地部分のマスキングができたもの（図 7 はこれに該当）」「2.はみ出しがあり，詳細な採集地を推測可能であるもの（例:図 8）」「3.大幅にずれたもの・マスキング範囲を確認できないもの（例:図 9）」の 3 段階で分類した。

その結果，日本語ラベル・英語ラベルともに，入力画像の種類によるマスキング結果には差異が見られなかった一方で，全体的なマスキングの成功度合いには差異が見られた。英語ラベルでは 7 割から 8 割のマスキングに成功した画像（3 段階のうち 1 に相当）が得られた一方，日本語ラベルの場合 4 割弱にとどまった。

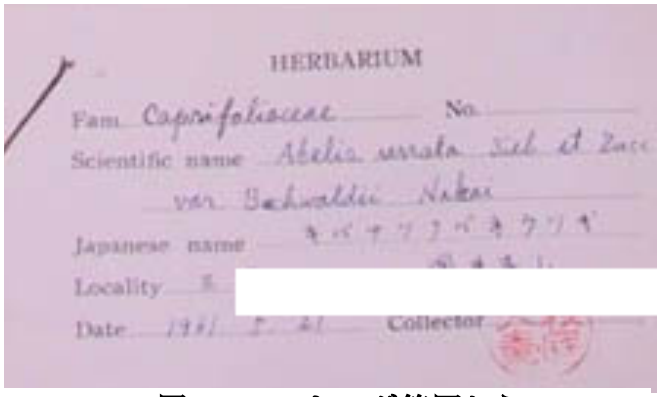

図 8 マスキング範囲から はみ出しが見られる例

図 9 マスキング範囲が大幅にずれた例

これに関し，マスキング範囲の決定に至った認識結果(認識文字列)を分析すると，標本ラベルの見出し部分（例:科名を意味する「Faculty」「Fam.」など)に相当する文字列から採集地の位置を推定しているケースが多くみられた。残りは手書き文字で記載された英語・ラテン語の科名からの推定が主であった。

また, OCR 実行時のパラメータである Language Hint が英語・ラテン語系の場合，日本語で記載された部分（見出しも含む）はほぼ認識ができない傾向にある。日本語ラベルに対し，OCR 時の Language Hint を日本語として実行すると，英語部分に比べると認識の度合いは悪化するものの，活字で記載された見出し等は認識不可能ではない傾向にあった。

これらを踏まえ，日本語ラベルの 1 グループに対して OCR 時の Language Hint を日本語に変更した場合のマスキング結果を見る追加実験を行った。実施にあたっては，Language Hint 以外の条件に変更は加えていない。その結果は，得られるマスキング画像の枚数は大きく増加するものではなく，また入力画像の種類による差異が見られなかった点も従前の結果と変化がなかったが，すでに Language Hint に英語を指定して得られた結果と合わせると，トータルではマスキングに成功したものが増加することとなった。

これらの結果から，OCR で認識できた情報から，画像中の採集地の位置を間接的に推定するアプローチは有効であると考えられる。しかし，マスキング結果の分類の 2 に示すように，実際の位置と少しずれた位置に文字を検出するケースも見られるため，マスキング範囲決定のアルゴリズムなどには若干の調整は必要といえる。

6. フィルタ処理が OCR 結果に与える影響

Cloud Vision API は内部に CNN (Conventional Neural Network) を用いているという文献[10]があり，これら機械学習による処理はノイズによる影響を受けることが知られており，耐ノイズ性(Adversarial example・Perturbation)に関する研究が進められている。代表的な研究として 2016 年の Seyed-Mohsen Moosavi-Dezfooli らによる研究[11]があるほか，Cloud Vision API で用いる画像に対してノイズフィルタを適用すると，OCR 内部のモデルを変更せず認識精度を向上しうるという研究結果[12]もある。

このため，本研究においても OCR 前の画像に対するノイズフィルタの適用が有効であるかを確かめることとした。前項のマスキング範囲決定方法の検討に用いた 3 種の入力画像に対し，それぞれローパスフィルタ・ガウシアンフィルタを適用し，OCR を実行した。

しかし結果は，ほとんどのケースで認識文字数・内容ともに著しく悪化した。

フィルタサイズやフィルタ適用前の入力画像の選択などのパラメータの調整次第では有効である可能性が否定されたわけではないものの，今回の場合，ここまでに得られた結果で自動的にマスキングを行うめどは立てられる状況にあり，追加の検証については見送った。

7. マスキング方法に関する結論

今回の調査を通じ、残りの萩庭植物標本画像については，次の図 10 に示す形でマスキング処理を施すものとする。

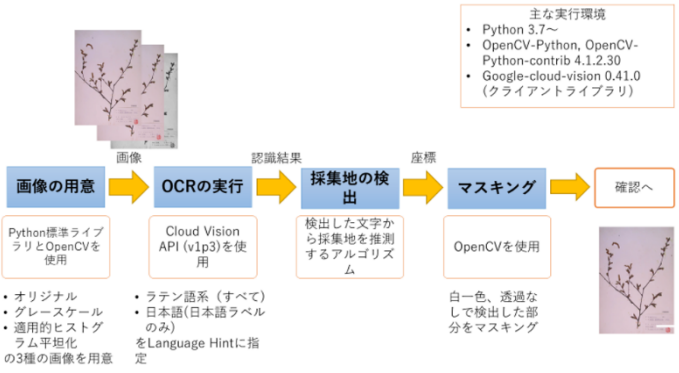

図 10 本研究をもとにした植物標本画像のラベル自動マスキングの方法の流れとパラメータ

なお，今回の手法では 1 枚の標本画像に対して複数のマスキング済みの画像が生成されることになるが，今回の実験で行ったように，目視で判断する方向で現在は検討を進めている。

8. 今後の展望・課題

本研究における手法に関して，他の史料への応用等を検討するうえではいくつかの課題が挙げられる。

まず，OCR 結果に影響を与える因子の分析は課題の一つである。本稿の範囲では一部データが不足しているなどの理由から詳述を見送ったものの，画像調整に用いた統計的正規化は画像の明度・コントラストを個別に変更できるため，これらが OCR 結果に与える影響の評価にも利用が可能と考えられる。

次に，画像の分類作業の自動化である。今回，標本ラベルごとに仕訳ける作業ならびにマスキング結果の分類・取捨選択については目視による手作業で実施した。この部分の自動化による作業負荷軽減を検討することが，他の史料への適用の際を含め，今後の大きな課題の一つである。

上記に関連して，OCR 実行前に認識結果を予測できる指標の導入も検討課題として挙げられる。本研究では OCR 結果の定量的評価指標として認識文字数を主に利用した。しかし，これは OCR 実行後でなければ得られない情報であり，本研究の範囲では事前に知り，入力画像の調整やパラメータの選択に生かすことができない。

このため，OCR 実行前にある程度認識結果を予測できる指標が求められる。画像を利用するほかの分野，例えば超解像や機械学習の分野では SSIM(Structural Similarity)や PSNR(Peak Signal-to-Noise Ratio)を用いて画質を評価することが多く，比較画像を必要としない BRUSQUE[13]なども登場している。これらの指標と OCR の認識結果（認識文字数）を結びつけることができた場合，OCR 時のパラメータをさらに絞り込むことができ，分類等の負荷軽減にもつながると考えられる。

9. まとめ

本研究では標本画像に記載された採集地を検出し，マスキング範囲の決定・実行までの一連の操作を自動的に行う方法について検討を行った。複数手法を試したところ，採集地の直接的検出は難しいものの，光学的文字認識を活用することで，間接的に位置を推定するアプローチが有効と言える結果が得られた。

まだ調整を要する部分などはあるものの，標本画像のラベル自動マスキング方法のめどは得られ，OCR の活用やデジタル・スカラシップ開発の上でも，本研究では一定の成果および示唆が得られたと考えられる。

なお、萩庭植物標本画像の c-arc での公開に向けては、本稿で述べた事項・手法は変更を生じる可能性がある。

謝辞: 本研究に際し，資料提供など多くの場面で協力いただいた千葉大学図書館の皆様へ，この場を借りて心より感謝の気持ちとお礼を申し上げます。

文　献

[1] 萩庭植物標本画像データベース作成協力会, 萩庭植物標本画像データベース作成プロジェクト 総括報告書. 千葉市稲毛区 : 萩庭植物標本画像データベース作成協力会, 2008. p. 17. 第 1 巻.

[2] 檜垣泰彦，ほか. “千葉大学学術リソースコレクション(c-arc) ～大学図書館における情報システム開発事例～”. 一般社団法人 電子情報通信学会，信学技報, vol. 118, no. 420, LOIS2018-51, pp. 51-56, 2019 年 1 月.

[3] 環境省. 環境省レッドリスト 2019 の公表について | 報道発表資料. 環境省. （オンライン） 2019 年 1 月 24 日. （引用日: 2020 年 1 月 13 日.） https://www.env.go.jp/press/106383.html.

[4] OtsuNobuyuki. “A threshold selection method from gray-level histograms”, “IEEE Transactions on Systems, Man, and Cybernetics”, vol 9, pp.62-66. January 1979.

[5] CannyJ. “A Computational Approach To Edge Detection”. IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. PAMI-8, no. 6, November 1986.

[6] Tesseract OCR, Training Tesseract - tessdoc – Tesseract Documentation （オンライン）, 2020 年 2 月 7 日（引用日: 2020 年 2 月 13 日.）

[7] Stefan Weil, et, al. Home – UB-Mannheim/tesseract wiki – GitHub （オンライン）, 2019 年 10 月 31 日, （引用日:2020 年 2 月 13 日）

[8] Google Inc., Python Client for Google Cloud Vision （オンライン）, 2017, （引用日:2020 年 2 月 13 日）

[9] Google Inc., 光学的文字認識 (OCR) – Cloud Vision API ドキュメント （オンライン）, （引用日:2020 年 1 月 31 日）

[10] Jake Walker, Yasuhisa Fujii and Ashok C. Popat, （Google Inc.） “A Web-Based OCR Service for Documents”, 13th IAPR International Workshop on Document Analysis Systems, pp.22-23, 2018.

[11] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi and Pascal Frossard, “Universal Adversarial Perturbations”, 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 21-26 July 2017.

[12] Hossein Hosseini, Baicen Xiao and Radha Poovendran, “Google’s Cloud Vision API Is Not Robust To Noise”, 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA), pp.101-105, 18-21 December, 2017.

[13] Anish Mittal,Anush Moorthy Krishna and Alan Bovik Conrad., “No-Reference Image Quality Assessment in the Spatial Domain”, IEEE Transactions on Image Processing, Vol.21, 12, December 2012.