資料画像中の挿絵領域の自動抽出及び画像検索システムの実装

青池亨・里見航・川島隆徳（国立国会図書館）

国立国会図書館電子情報部電子情報企画課次世代システム開発研究室では，デジタル化資料に対する検索性の向上を目的として，資料画像中の挿絵領域と文字領域を自動的に切り出す方法を検討している．

文字領域の切り出しは，文字の含まれない領域を予め取り除くことで OCR 処理の精度を向上させる目的と，利用者の可読性向上のためのコントラスト補正への応用を想定している．また，挿絵領域の切り出しについては，コンテンツベースの類似画像検索への応用を想定している．

今回は後者に焦点を当て，資料画像からの挿絵領域の切り出しとコンテンツベースの画像検索を一貫して行うシステムを実験的に作成したので報告する．本提案手法を組み込んだサービスは平成 30 年度内に NDL ラボのページ(https://lab.ndl.go.jp/)から実験的に公開予定である．

Automatic extraction of illustration from images of documents and image retrieval

Toru Aoike / Wataru Satomi / Takanori Kawashima (National Diet Library)

The National Diet Library is now developing techniques for automatically recognizing which areas of a printed page are illustrations and which are graphemes, as a means of improving the searchability of digitized material. The ability to distinguish between illustrations and graphemes is expected to improve the accuracy of OCR processing by allowing areas without graphemes to be ignored while enabling the application of contrast correction to areas with graphemes, thereby improving readability of the digital images. Moreover, the ability to extract areas with illustrations is expected to have practical applications for content-based retrieval of similar images. This paper focuses on the extraction of areas with illustrations and reports on the creation of a system that is consistently able to extract illustrations from digital images of documents as well as perform content-based retrieval of images. Services incorporating these proposed techniques will be released on a trial basis on the NDL Lab website. (https://lab.ndl.go.jp/).

１．イントロダクション

人文科学において，画像検索の取組は多く，先行する web サービスとして，jQuery の開発者である John Resig の開発した浮世絵の画像検索サービス Ukiyo-e Search (https://ukiyo-e.org)がある．文字画像の検索サービスとしては奈良文化財研究所と東京大学史料編纂所が共同で開発した木簡・くずし字解読システム MOJIZO (http://mojizo.nabunken.go.jp/)があり，研究者等から利用されている．また，国立情報学研究所の松井勇佑らは国文学研究資料館との共同研究「画像検索のための構造化問い合わせ言語による歴史的典籍画像検索システム」の中で，「日本古典籍データセット」内の類似画像検索サービスを提供している（http://yusukemat-sui.me/project/kotenseki/kotenseki_jp.html）．画像の一部から画像を検索する技術の実用化も進んでおり，Seguin らは，畳み込みニューラルネットを利用した特徴抽出により，絵画の一部を選択し，同様の部分を持つ絵画を検索する手法を提案している[1].

これらのサービスは，自動収集した美術館等の所蔵品の画像をそのまま利用することや，人手による切り出しを行うことによって，検索対象の画像を収集している．絵画等ではなく図書資料での応用を想定した場合，例えばある資料に掲載されている写真と同じ写真を掲載している別の資料を検索できることは文献探索上有意義であると考えられるが，単純に撮影した資料の画像全体で検索を行っても精度は高くなく，また大量の資料を想定した場合には，手動による写真の切り出しも現実的ではない．

他方，OCR の前処理に関わる研究としては，文書のレイアウトを自動認識する研究について，機械学習手法を使った研究[2]や認識精度を競うコンペティション[3]が行われている．国立国会図書館においてもこの文脈において研究を行っており，その成果として資料中の挿絵(写真や図表を含む）の高精度な自動抽出に成功している．

本研究では，この挿絵の自動抽出の成果と既存の画像検索の技術を組み合わせることにより，大規模な資料画像中から挿絵を自動的に切り出し，各挿絵に含まれる特徴を利用して同様の挿絵や同様の挿絵を含む資料を検索するシステムの開発を行った．

２．材料・方法

２．１．　実行環境

本研究で使用した環境は以下の通りである．

CPU:Intel(R) Xeon(R) CPU E5-2620 v4 @ 2.10GHz
GPU:NVIDIA Geforce GTX1080Ti

２．２．　挿絵の抽出方法

画像を各領域の属するクラスごとに塗り分けるセマンティックセグメンテーションと呼ばれるアプローチを用いた．アルゴリズムには DeepLabV3plus の tensorflow 実装[4]を利用し

て，文書レイアウトを認識し，図表と認識された領域を切り出して検索対象とした．

学習に利用するデータセットとしては，国立国会図書館デジタルコレクション(http://dl.ndl.go.jp/)中の図書及び雑誌の画像のうち，見開きページをのど元で分割した合計 1,640 枚の画像について，

1. 挿絵領域 2. 文字領域 3. その他背景領域

の 3 種類の矩形に塗り分けたレイアウトデータセットを作成し，うち 1,400 枚を訓練データセット，240 枚をモデル検証用のデータセットとした．データセット全体で文字領域は 6,296 個，挿絵領域は 1,652 個含まれた．含まれる領域数ごとに画像枚数のヒストグラムを図 1 に示した．

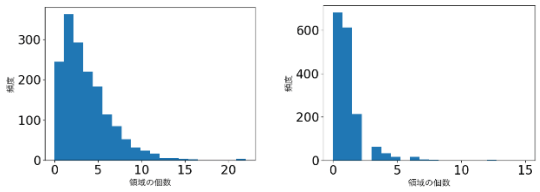

図 1 画像 1 枚当たりに含まれる文字領域の個数(左)と挿絵領域の個数(右)のヒストグラム

学習は cross entropy を損失関数として行い，モデル検証用データに対する画素単位での正解率が最も高くなった epoch で学習を打ち切った．この時の正解率は 85%であった．

２．３．　挿絵画像からの特徴抽出

画像からの特徴抽出は，切り出された挿絵領域を 224*224 のサイズに変形した画像を入力とし，ImageNet[5]で学習済みの ResNet50[6]によって抽出された 2048 次元の特徴ベクトルを利用した．特徴ベクトルの抽出に学習済みのディープニューラルネットの埋め込みベクトルを使う手法としては[7]のような先行研究がある．

２．４．　検索インデックスの構築

検索に利用する特徴ベクトルの次元が大きいため，大規模な検索対象に対して高速な類似検索を提供することを目的に，Neighborhood Graph and Tree for Indexing [8](NGT)を用いて構築した．

３．性能評価

３．１．　領域の自動認識精度

２．２．で作成したレイアウト認識のデータセットとは別に評価用の 180 枚からなるレイアウトデータセットを作成し，評価に用いた．テスト用データセット内に文字領域は 1,054 個，挿絵領域は 412 個含まれた．領域数ごとの画像枚数のヒストグラムを図 2 に示した．

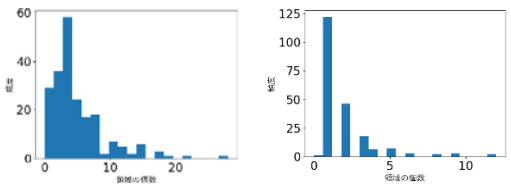

図 2 画像 1 枚当たりに含まれる文字領域の個数(左)と挿絵領域の個数(右)のヒストグラム

本テストデータセットにおける，画素単位での正解率は 83%であり，この時の正解の挿絵領域に対する挿絵と認識された領域の Intersection-Over-Union(IOU)は 0.53 であった．

自動認識された領域について，図 3 に例示する．図内の青い領域の外接矩形について，上下左右 20 ピクセルのマージンを設けて切り出した領域が本提案手法における検索対象となる．

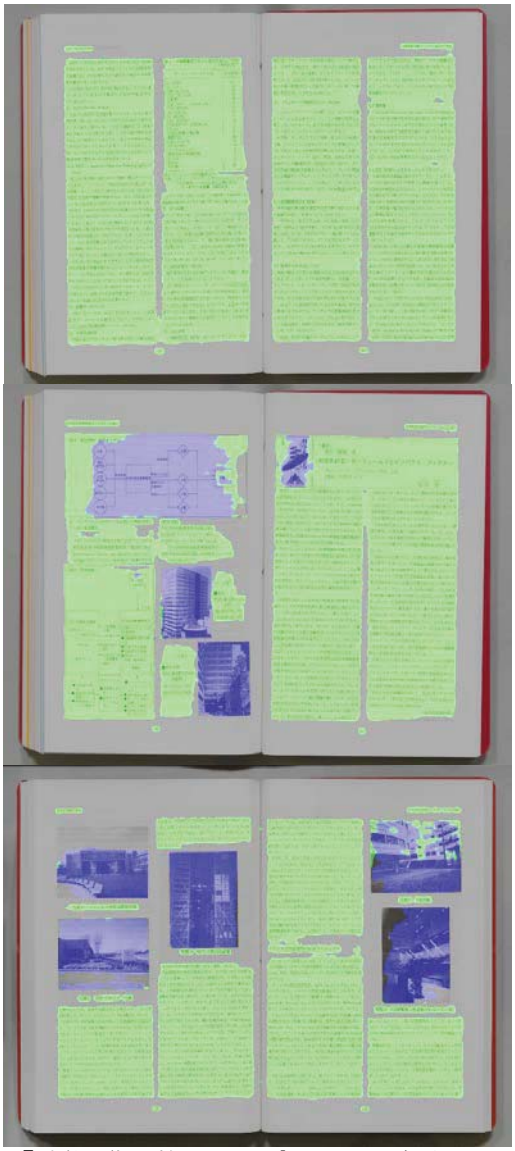

図 3『科学技術文献サービス』に対して自動認識を行った結果(緑が文字領域，青が挿絵領域)

３．２．　クエリに対する応答速度

400*600 程度のサイズの任意の入力画像からの特徴ベクトルの抽出に 200 ミリ秒前後を要した．また，抽出された特徴ベクトルをクエリとし，応答速度を確認するために用意した 18,000 枚の画像の特徴ベクトルから NGT を利用して作成した検索インデックスに対する類似検索では 2 ミリ秒を要した．

４．　画像検索システムの概要

検索機能として，

1. ２．１．において学習したモデルにより検索したい資料画像からの挿絵領域の切り出し(図 4)と，切り出された挿絵領域から特徴ベクトルを抽出する機能
2. 検索クエリとなる特徴ベクトルと類似した挿絵のメタデータを，NGT によって作成した特徴ベクトルのインデックスから検索する機能

をそれぞれ API として実装し，別途作成したフロントエンドから提供する web サービスとした．

図 4. 作成した自動認識モデルにおいてユーザがアップロードした画像(上段)と API により切り出された検索クエリ候補の例(下段)

５．インターネット公開資料への適用

５．１．材料・手法

国立国会図書館デジタルコレクションでインターネット公開を行っている資料のうち，日本十進分類表(NDC)で大分類 6(産業)に該当する著作権保護期間満了資料の中から，図書 1,481 点，計 194,916 コマを取得し，提案手法による挿絵領域の自動抽出の適用と，抽出された領域の特徴による検索インデックスを構築した．

５．２．所要時間

194,916 コマに対する挿絵抽出に 76 時間，特徴抽出に 6 時間，検索インデックスの構築に 1 分を要した．

抽出の結果，284,134 か所の挿絵と認識された領域が抽出された．これらの画像の特徴から構築した特徴ベクトルに対する類似検索の応答速度は 5 ミリ秒程度であった．

５．３．検索結果の例示と考察

図 5a におけるクエリ画像は『千九百年巴里万国博覧会臨時博覧会事務局報告．下』(http://dl.ndl.go.jp/info:ndljp/pid/801823)457 コマ目に掲載されているメダルである．

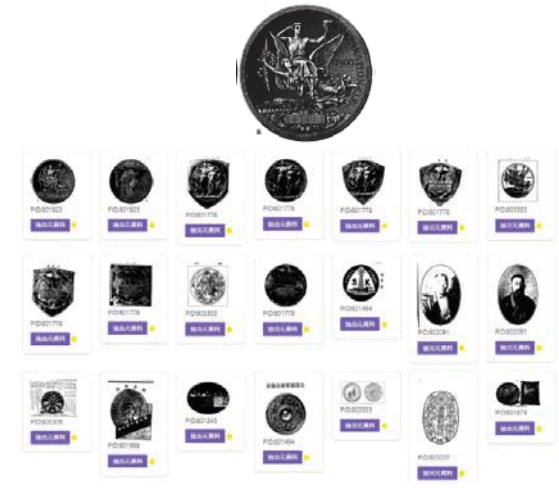

図 5a クエリ画像(上段)と検索結果例(下段)

下段の検索結果は左上から右下に向かって，クエリ画像に特徴ベクトルが近い順にソートされている．

上位 3 番目から 8 番目に見られるメダル画像は，図 5b で示した通り同一の資料ページから提案手法を用いて自動的に切り出された画像である．

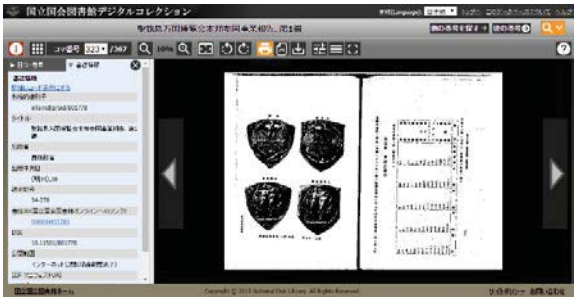

図 5b 図 5a の検索結果に表示された画像の抽出元となった資料(『聖路易万国博覧会本邦参同事業報告．第 1 編』323 コマ目
http://dl.ndl.go.jp/info:ndljp/pid/801778/)

図 5a において，メダル画像の次には，円形に縁どられた肖像画や，円形の徽章等，丸みを帯

びた形状の中に模様があるものが上位に見られた．モノクロの資料においては輪郭線が重要な特徴として機能しているものと推察される．

また，図 6 のように，形状によっても認識が可能である．

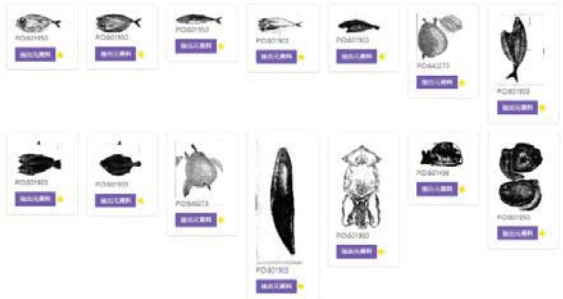

図 6 勧業博覧会の品目として掲載された魚の干物の写真に対する画像検索結果

形状が類似するものについて上位に並んだが，形状の類似していない画像についても同一の資料に掲載があるものは上位にヒットする傾向があった．同一資料の場合，抽出された背景領域のパターンや挿絵部分の濃淡が酷似しているため，類似度が高く見積もられたことが原因とみられる．

図 7a に示した通り，輪郭がはっきりしていれば建物の外観写真で検索を行うこともできる．

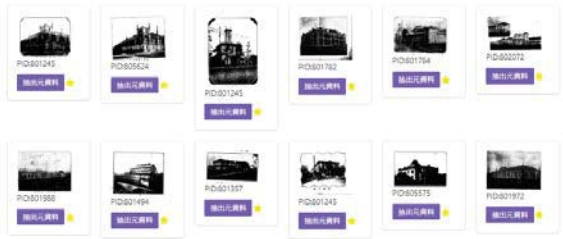

図 7a 生糸検査所の写真で検索した結果(※検索インデックス内に検索クエリと同一の画像が存在するため，左上の検索順位第 1 位は検索クエリ画像と一致している)

図 7b『神奈川県産業要覧』42 コマ目
(http://dl.ndl.go.jp/info:ndljp/pid/801245)

図 7c『横浜開港五十年史・下巻』313 コマ目
(http://dl.ndl.go.jp/info:ndljp/pid/805624)

検索結果(図 7a)の左上にある『横浜開港五十年史』に掲載されていた生糸検査所の写真(図 7b)から，検索結果の左上から 2 番目にある『横浜開港五十年史・下巻』に掲載されていた同一の生糸検査所の写真(図 7c)を発見することができた．検索クエリと一致しないが，類似しているとされた画像については，空の部分が白く抜かれていて屋根の形状が類似した建物が上位を占める傾向にあった．

また，図 8，図 9，図 10 に示す通り，蔵書印等，印影に対する検索も行うことも可能である．

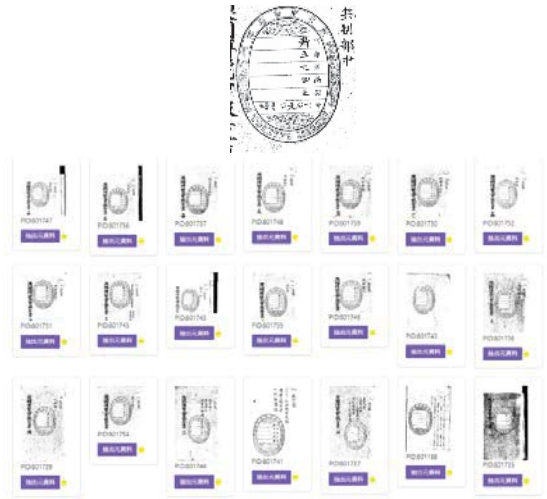

図 8 東京書籍館の蔵書印(上段)に対する検索結果(下段)

図 8 において，検索結果の画像が実際の蔵書印よりも過大な大きさの領域として抽出されている．東京書籍館の蔵書印は資料の表紙に押される傾向があり，強いノイズの乗った画像では，蔵書印と何も書かれていない背景領域の境界を正確に捉える事が出来ず，表紙全体が挿絵として抽出されたことが原因とみられる．

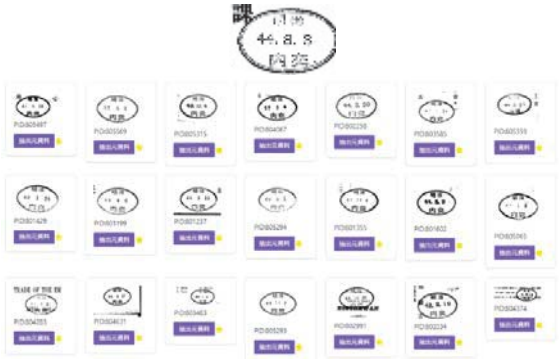

図 9 内務省交付本の印影(上段)に対する検索結果(下段)

図 9 の内務省交付本の印影の抽出については，文字のあるページに文字と重ならないように押印される傾向があり，比較的正確に目的の領域の抽出に成功している．

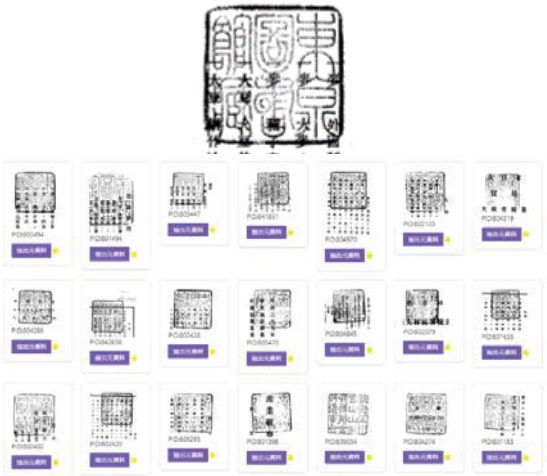

図 10 「東京図書館蔵」の印影(上段)による検索結果(下段)

図 10 の検索結果には類似の蔵書印の誤検出が含まれた．図 10 下段に示した上位 20 件の内訳は，「東京図書館蔵」の蔵書印が 10 件，誤検出として「帝国図書館蔵」の蔵書印が 9 件，その他の印影が 1 件であった(図 11).

図 11 誤検出された 2 種類の印影

６．検索結果の性能評価

５と同様の検索対象に対して，蔵書印画像を利用して，本画像検索システムの検索適合率を評価した．

６．１．材料・手法

図 10 の上段で示した「東京図書館蔵」の蔵書印画像を検索クエリとして，上位 500 件の画像を手作業で分類した．内訳は表 1 の通り．

表 1 検索結果上位 500 件に出現した画像

順位	蔵書印文	個数
1	東京図書館蔵	155
2	帝国図書館蔵	147
3	東京図書館蔵書之印	31
4	国立国会図書館蔵書	14
5	内務省書庫	12
6	衆議院図書印	4
－	その他の印文	23
－	蔵書印でないもの	114

表 1 の 2 から 6 について印影の例を図 12 に左から順に示した．

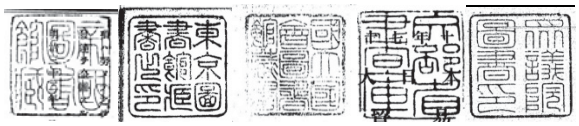

図 12 検索対象以外の印影の例

蔵書印でないものの割合は検索順位が下位になるほど増える傾向にあり，内訳としては商標等印影でない挿絵や版権者を表す印影が見られた．

表 1 で分類した結果に基づいて各画像をラベル付し，「東京図書館蔵」の蔵書印画像 155 枚を正解ラベルとして，以下の実験を行った．

実験 A. 各画像を検索クエリとしたときの上位 10～100 件の検索結果に対する平均適合率と平均再現率を計算した

実験 B. 155 枚の中からランダムに 2 枚の画像を選択し，両画像の特徴ベクトルを平均したベクトルを検索クエリとして 1,000 回検索を行い，実験 1 と同様に平均適合率と平均再現率を計算した

６．２．１．　実験 A の結果

155 枚の正解画像による検索結果の上位件数に対する評価を表 2 に示した．なお，155 枚はいずれも検索インデックス内に存在するため，本結果において上位に検索クエリ画像自体が出現する可能性が高いことに留意されたい．

表 2 クエリ画像 155 枚の平均適合率と平均再現率

上位(件数)	平均適合率	平均再現率
10	0.644	0.042
20	0.593	0.077
30	0.548	0.106
40	0.524	0.135

50	0.500	0.161
60	0.482	0.187
70	0.466	0.211
80	0.451	0.233
90	0.440	0.255
100	0.428	0.276

６．２．２．　実験 B の結果

ランダムに抽出した 2 枚の特徴ベクトルの平均で 1,000 回検索を行い，検索結果の上位件数に対する評価を表 3 に示した．

表 3 ランダムに抽出した 2 枚の特徴ベクトルの平均で検索した際の平均適合率と平均再現率

上位(件数)	平均適合率	平均再現率
10	0.786	0.051
20	0.707	0.091
30	0.658	0.127
40	0.631	0.163
50	0.608	0.196
60	0.575	0.223
70	0.569	0.257
80	0.547	0.282
90	0.536	0.311
100	0.526	0.339

６．３．　考察

実験 A から適合率と再現率のトレードオフについて，蔵書印に限れば，検索結果の 1 画面に表示すべき件数は 20 件から 30 件程度が好ましいと考えられる．

また，実験 B から，利用者に検索クエリ画像を複数枚選択してもらうことで，検索したい画像の部分に共通する特徴をとらえることができると考えられるため，検索パフォーマンスを大きく向上させることが可能と推察される．

７．全体考察

挿絵領域の切り出しについて，今回の実験では，学習データセットの作成コストの観点から，学習データセット・評価用データセットとも正解ラベルの付与を矩形に限定して行った．実際の雑誌では挿絵の領域は必ずしも矩形の形状をとるとは限らず，正確な学習データの作成と評価方法について課題が残った．輪郭検出等で実際の形状に沿った形のデータセットを作成することができれば，より正確な領域の認識と評価が可能になると推察される．

また５に挙げたような著作権保護期間満了資料にはノイズのある資料が多く含まれるため，領域の誤認識を防ぐためにノイズに対してロバストな領域認識方法やグレースケール画像から性質の良い特徴を抽出する方法についても検討が必要と考えられる．

また，図書・雑誌全般の挿絵を対象とすることで，検索対象となる画像のバリエーションが豊富であることから，特徴が似ている画像であったとしても，利用者の視点からは妥当とは見えない場合があり得ると考えられる．６で蔵書印画像を利用した検索結果の定量評価と複数画像を利用した際の性能改善の評価を試みたが，検索にヒットした挿絵の掲載されている資料のメタデータを利用する等，定量的な評価方法を引き続き検討する必要がある．

８．実験サービスの公開

５に挙げた，NDC 大分類 6(産業)に分類される著作権保護期間満了資料を対象とした検索システムについて，本研究を含む次世代システム開発研究室における研究成果を「次世代デジタルライブラリー」として平成 30 年度内を目途に公開予定である．並行して機械学習用データセットの公開準備を進めている．

参考文献

1) Seguin, B., di Leonardo, I. and Kaplan, F., Tracking Transmission of Details in Paintings, the Digital Humanities 2017, (2017)

2）Esposito, F., Malerba, D., Semeraro, G. and Ferilli, S.: Machine learning methods for automatically processing historical documents: From paper acquisition to XML transformation, *Document Image Analysis for Libraries, 2004. Proceedings. First International Workshop* pp. 328-335. (2004)

3) Gao, L., Yi, X., Jiang, Z., Hao, L., and Tang, Z.: ICDAR2017 Competition on Page Object Detection. *Document Analysis and Recognition, 14th IAPR International Conference*, vol. 1, pp. 1417-1422 (2017).

4）Chen, L. C., Zhu, Y., Papandreou, G., Schroff, F., and Adam, H. :Encoder-decoder with atrous separable convolution for semantic image segmentation. arXiv preprint arXiv:1802.02611, available from 〈https://arxiv.org/abs/1802.02611〉 (2018).

5) Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., and Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In Computer Vision and Pattern Recognition,(CVPR 2009). pp. 248-255 (2009).

6）He, K., Zheng, X., and Ren,S.: Deep residual learning for image recognition, *Proceedings of the IEEE conference on computer vision and pattern recognition* (IEEE, 2016), pp. 770-778 (2016).

7) Matsuo, S. and Yanai, K.: CNN-based style vector for style image retrieval, *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval* (ACM, 2016), pp.309-312 (2016).

8）Sugawara. K., Kobayashi. H., and Iwasaki. M.: On Approximately Searching for Similar Word Embeddings, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (ACL, 2016）, pp.2265-2275 (2016).