歴史的文書の文書画像解析

A.　メンバー

【事業推進担当者】八村広三郎、赤間亮
【客員研究員】當山日出夫
【PD】李亮
【RA】Chulapong Panichkriangkrai
【学内研究協力者】ウォーカー，ロス、山本泰則、竹久修平

B.　研究目的

江戸時代を中心として多く出版された古典籍の画像および浮世絵の画像について、画像解析に基づく分析システムとその教育・研究面への応用の研究を行う。

古典籍関連では、挿絵と文書部分の分離、画像ベースでの文字の切り出しなどによる、文書画像の構造記述とそのXML化、文字や単語のスポッティング、インデックスの作成、翻刻・解読支援などの諸機能について研究を行う。

浮世絵については、文字列部分の抽出を基礎とする、画像のXMLによる構造記述、落款の字形に基づく作者同定や編年、色彩の利用などの情報を利用した木版画の類型判定と画像検索などについての技術的研究を行う。

C.　本年度の成果

1）古典籍からの文字切り出しシステム

江戸時代に大量に出版・流通した、木版印刷による古典籍については、本学ARCをはじめ、国立国会図書館など、いくつかのところでデジタル化が始まっている。これらはページイメージのデジタル化であり、PCやタブレット型の端末などで、アクセスし読むことができる。これらは日本文化･出版文化についての貴重なアーカイブであり、今後もこのような活動は広まっていくことは確実である。

一方で、このような形での古典籍の公開では、当時の文字、また言語表現等に不慣れな者にとっては、外国語と同等あるいは場合によってはそれ以上に難解な書物ともみなされ、必ずしもこのままの形で一般人また学生などに普及するとも思えない。

一つの方法はこれらの古典籍のそれぞれの文字を認識して、現代の文字に置き換える方法、すなわち古典籍のOCR（文字認識）の開発である。これには、一定の需要があると思われる。本研究でもそれを最終目標として目指すが、当面は、認識レベルまでを想定せず、まず古典籍のページを構成する各要素、すなわち、行、本文文字、振り仮名、挿絵などを自動的に分離抽出することを目標としている。

このため、当面は、文字行の組み方が明確な文献について、行の位置決め、行中の個別の文字の切り出し抽出の自動処理を目標と定め研究を行っている。

現在の書籍や文書とは違う、古典籍特有の扱いにくい性質がある。たとえば、ページの汚れ、破損、デジタル化の際の文書の傾き、続け字などがある。これらのためには、さまざまな画像処理機

能が必要である。また、同じカテゴリー内に属する古典籍といっても、それぞれが特有の特徴をもっており、ひとつの処理方法ですべてのものに対応できることはあり得ない。

現在、100 ページ程度の古典籍を対象として想定し、最初の数ページは、コンピュータによる処理をオペレータが修正・ガイドする形で対応し、その時に得た。対象の書物の性質・特徴に基づいて、残りのページを自動的に処理するという半自動での解析をめざしている。

この研究が完成すると、一冊の本に現れるすべての文字図形についてのインデックスが作成でき、同じ文字がどの程度の字形の揺れで表現されているかを知ることができる。また、文字同士の類似性を判定し、ある文字が、本の中のどこで現れているかを特定する、キャラクタスポッティングが行える。また、ある文字がどの文脈で現れるかを一覧表示する、文字図形ベースでのKWIC(Key Words In Context) が作成できる。

現時点では、挿絵のない「椿説弓張月」を対象とし，本文文字の切り出し処理の完成をめざしているところである。

本研究のこれまでの成果は 2012 年 3 月に国際会議で発表する予定である。

2）浮世絵画像の構造解析

本学ARCには大量の浮世絵画像のデジタルアーカイブが構築されている。本研究では、このアーカイブ内の浮世絵デジタル画像を対象として、浮世絵の中に描かれている、あるいは、書かれているものを分離抽出することを目標としている。すなわち、描かれている対象物や情景を切り出すようなこと重要と思われるが、これ以外にも、重要で、比較的容易と思われる、絵の中に書かれている文字、文字列の抽出を行っている。

特に、浮世絵中の絵師の落款などの情報、また、役者絵の場合などは、訳者名や演目名などは浮世絵の分析に大きな手掛かりになる。

もちろん、浮世絵研究の専門家にとっては、このような情報は即座に判断できるが、入門研究者が、世界中に散在している、浮世絵画像データベースの数万枚、数十万枚の画像を対象として検索や比較を行おうとする際には、コンピュータの画像処理による、文字列の特定や認識技術が有用となる。最終的には、浮世絵中の複数の文字列の存在する部分だけを分離抽出し、さらにその中から文字だけを抽出する。これにより何が可能になるかというと、たとえば、同じ絵師の描いた浮世絵の「落款」文字の字形を、大量の浮世絵の中で相互比較することができる。同じ版で作られた、浮世絵は複数存在するが、その刷りの順番を文字の形のわずかな変化から読み取ることもできる。

落款などの文字列は背景の何もないところ、あるいは、一様な色のところに書かれることはまれで、浮世絵の主たる対象である風景や役者の絵の上に刷られている。これが問題を困難にしている。

したがって、今年度は、人間の目視により抽出した大量の落款部分の画像データから背景の部分を取り除き文字だけを抽出する処理を行っている。将来的には、落款部分を自動的に特定し、切り出すことも行う。

現在までのところの成果は 2012 年 3 月にオーストラリアで開かれる国際会議で発表する。

3）利用者の検索意図を反映させた類似画像検索システム

現在、膨大な量の画像データを簡単に誰でも利用可能になった。それに伴い、目的の画像を得る

ための効率的な手法の開発が求められ、多くの画像検索技術が提案されてきた。

画像に付与されたメタデータによる検索だけではなく、画像そのものが持つ情報を用いて、画像の内容に基づく「類似画像検索」を行うことが課題となっている。

その場合、画像から抽出する 1 種類の特徴量のみを用いることでは、必ずしも意図した検索結果となるとは限らない。一般に複数の特徴量を組み合わせて用いることで、適切な検索を行うことが可能となる。

しかし、その一方、複数の特徴量を用いた場合には、どの特徴量をどの程度重視して検索を行うかが課題となる。このためには、検索の目的に合わせて、各特徴量に重み付けを行った上で検索を行うことが一般的に行われている。最も簡単には、検索を行う際に、利用者にその重み付を行わせることで行えるが、これは、画像の特徴量の意味などを理解して行う必要があり、あまり現実的ではない。

そこで、本研究ではシステムが利用者の検索の意図を自動的に推定し、その結果に従って各特徴量に重み付を行うことを試みている。

すなわち、検索を行う際、利用者に複数の画像を選択させることで、その利用者がどの特徴量に注目しているかをシステムで推定し、各特徴量に与える重みを決定する。これにより利用者の検索意図を反映させた検索を行うことを目的とする。

すでに本研究室では、検索利用者に複数枚の候補画像を提示させ、それぞれの画像における各特徴量の散らばり具合を求めた上で、これがある程度まとまっているものが利用者の求める画像の特徴であると判断するシステムを開発してきた。これでは、特徴量の分散をその尺度とする単純なものであったが、良好な結果を得ることができた。

本研究では昨年度、

・TF-IDFを用いた重み付手法

・テキスチャ特徴の追加

の 2 つの改良を行い、検索評価実験を行って、よい結果を得ているが、今年度はさらに、画像の構造的な特徴を表現するために最近よく利用されるようになった、BOF（Bag Of Features）特徴も組み込んだ．その結果，さらに検索の精度が向上した。

今後はさらに他の画像特徴量についても利用し、例示画像から特徴量の重視の度合いを類推する手法をより精緻化すること、また、このシステムを、ARCで公開されている浮世絵画像データベースなどの検索インタフェースとして利用し、一般の利用者にも利用してもらえるようにしたいと考えている。

本研究の成果は、画像電子学会論文誌に投稿済みであり、現在査読中である。

4）メタデータと画像特徴による画像の類似性基づく画像検索システム

本学アートリサーセンターでは多くの浮世絵画像がデータベース化され一般にも公開されている。ここでは、浮世絵に対して、絵師、画題、出版年、版元などのメタデータ項目が付与されており、これらの属性値によって、特定の絵師によってある年に出版された浮世絵などを検索することができる。

一方、画像データのデータベースにおいては、前述のとおり、画像の類似性に基づく「類似画像検索」機能の重要性が指摘され、自分が手元に持っている画像と類似する画像をデータベースから検索することができる機能が望まれ、多くの研究

開発がおこなわれている。

木版画の伝統絵画である浮世絵の特性、たとえば、同じ版から多くの製品が作りだされ、しかもこれらが世界中に分散して存在する。一説によると全世界で 100 万枚の浮世絵が存在しているともいわれている。将来的にこれらの浮世絵画像の多くがデータベース化された暁には、すべてをメタデータだけを頼りに検索することもできないし、また、メタデータもあるとしても、たとえば、手元にある絵と「同じ」、または同じ版木が用いられているが、刷られた時期が違うため版木の摩耗などにより「ほぼ同じ」絵となったものの存在を世界中から探すことは興味ある研究課題になると考えられる。

以上のような観点から、ここでは、当面はアートリサーチセンターでデータベース化されている浮世絵画像データを対象として、メタデータによる検索、および、それと画像間の類似性に基づく画像類似検索の機能を持ったシステムを作成し、浮世絵研究、美術史研究に役立てようという目標を立てている。

画像の類似性についても、一般的にたとえば、画像中の色彩の度数分布（ヒストグラム）や色の空間分布、画像のテキスチャ特徴の類似などを組み込んだものが多いが、たとえば浮世絵など特定のジャンルの絵画を対象とした場合、そのジャンルでの「絵画の類似性」には、特有の観点がありうると考えられる。

このため、画像からの画像特徴の抽出機能についても、モジュラー性を確保して、必要に応じて追加などができるように、プラットフォームのデザインを考えている。

昨年度は、一般の利用者が手軽に使えるように、GUIなどのユーザインタフェースを工夫し、また、ユーザからの希望に応じて、改善やバージョンアップ、カスタマイズなどが行いやすいシステム環境を構築することを目指した。

今年度はこれを引き継ぎ、さらにテキスチャ特徴量について、DCT（離散コサイン変換）係数を用いる方法を検討し、実験を行った。評価実験の結果、検索性能の向上を得ることができた。

来年度はこのシステムをプラットフォームとして用い、GUIをさらに改良すること、画像特徴の種類を増やすこと、前述したような複数の例示画像を提示して利用者の検索意図を類推するような機能を付加することなどを考えている。

D. 論文・学会発表以外の活動の記

特記事項はありません。

E. 業績一覧

〈著書〉

八村広三郎, 田中弘美編『デジタル・アーカイブの新展開』ナカニシヤ出版, p.343, 2012 年 3 月 30 日,

Kozaburo Hachimura, and Tiromi T. Tanaka eds., "*New Developments in Digital Archives*", Nakanishita Shuppan, 343p., 30 March 2012

〈著書（分担執筆）〉

八村広三郎, 田中弘美「デジタル・ミュージアムの実現に向けて」八村広三郎, 田中弘美編『デジタル・アーカイブの新展開』ナカニシヤ出版, pp.16-37, 2012 年 3 月 30 日, Kozaburo Hachimura, and Tiromi T. Tanaka, 'Towards the Realization of the Digital Museum', Kozaburo Hachimura, and Tiromi T. Tanaka eds., "*New Developments in Digital Archives*", Nakanishita Shuppan, pp.184-205, 30 March 2012

八村広三郎「デジタル・アーカイブ技術の現状と課題」八村広三郎, 田中弘美編『デジタル・アーカイブの新展開』ナカニシヤ出版, pp.1-15, 2012 年 3 月 30 日, Kozaburo Hachimura, 'The Current Status and Issues of Digital Archiving Technology', Kozaburo Hachimura, and Tiromi T. Tanaka eds., "*New Developments in Digital Archives*", Nakanishita Shuppan, pp.169-183, 30 March 2012

〈口頭発表〉

【審査付き】Liang Li, Chulapong Panichkriangkrai, Chihiro Tsunoda, and Kozaburo Hachimura, 'A binarization approach for Ukiyo-e Rakkan extraction', *The 10th IAPR International Workshop on Document Analysis Systems (DAS2012)*, Griffith University Gold Coast Campus(Gold Coast, Australia), 27-29 March 2012 (Poster)

【審査付き】Chulapong Panichkriangkrai, Liang Li, and Kozaburo Hachimura, 'Character segmentation for Japanese woodblock printed historical books', *The 10th IAPR International Workshop on Document Analysis Systems (DAS2012)*, Griffith University Gold Coast Campus(Gold Coast, Australia), 27-29 March 2012 (Poster)