1ZC-05

言語が異なる浮世絵データベース間における同一作品の同定手法の提案

木村 泰典† Biligsaikhan Batjargal‡ 木村 文則‡ 前田 亮†

立命館大学情報理工学部† 立命館大学衣笠総合研究機構‡

1 はじめに

浮世絵は江戸時代に成立した絵画のジャンルであり，人々の日常の生活や風物などを題材として描かれている．近年，美術品や芸術作品をディジタル化し，ディジタルアーカイブとして保存する動きが進んでおり，各国の美術館・博物館でさまざまな言語やメタデータ形式で浮世絵データが公開されている．

一方，浮世絵研究者からは，これらの浮世絵の画像やメタデータを網羅的に検索したいとの要望がある．また，異なるデータベース間で同じ作品のメタデータを比較することで，データの修正や補完などを行う機能が研究者から求められている．しかし，同じ作品であっても，データベースによってメタデータの内容や記述言語が異なるため，同一作品を同定することは容易ではない．このような問題を解決するために，我々は異言語かつ異種の複数浮世絵データベースから同一作品を同定する手法を提案している[1]．本研究では，先行研究で対象としていなかった，浮世絵作品名の原題と英訳のメタデータを用いて同一作品を同定する手法を提案する．

2 関連研究

レコード同定に関する研究動向については，相澤ら[2]によるサーベイ論文がある．この論文では同一言語データベース間でのレコード同定について様々な手法が紹介されているが，本研究では異言語のデータベース間でのレコード同定となるため，従来手法の適用は困難である．同言語同士で比較を行う場合は，編集距離などの文字列照合関数を用いて類似度を算出することができる．しかし，異言語同士で比較を行うには，一方の言語を他方の言語に翻訳する必要がある．本研究では，限られたメタデータから同定に有効な訳語を得ることが主な課題であると言える．

一方，画像の比較により複数データベースから浮世絵の同一作品の同定が可能なUkiyo-e.org[1]というWebサイトがある．このサイトで用いている手法と本研究との違いは，Ukiyo-e.org では画像の類似度を用いて同定を行っているのに対し，提案手法では作品のメタデータを用いて同定を行っている点である．画像を用いる手法では，言語の違いに影響されないというメリットがあるが，データベースによっては画像が存在しない場合や，浮世絵の異版など画像の類似度では同定できない場合があり，このような場合は提案手法が有効であると考えられる．

3 提案手法

本章では，作品名の原題と英訳を用いた同一作品の同定手法について説明する．提案手法全体の概要を図1に示す．

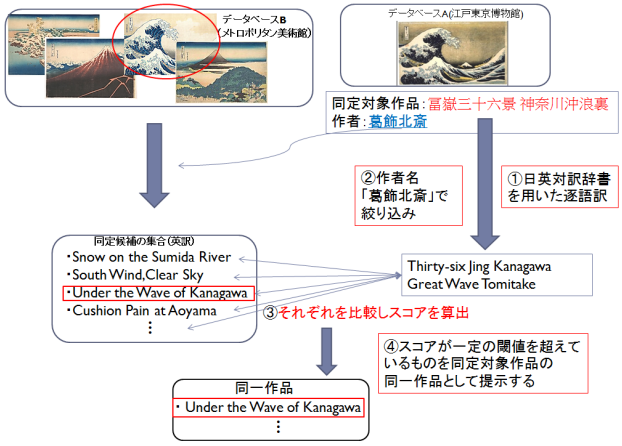

図1：提案手法の流れ

提案手法全体の流れは次の通りである．まずユーザは原題表記の浮世絵データベースから同定したい浮世絵作品を選択し，その作品名をクエリとする．次に，対訳辞書を用いてクエリの作品名(原題)を英語に逐語訳する(図1①)．そして，クエリの作品の作者名で英訳表記のデータベースから同定対象候補となる浮世絵作品を絞り込む(図1②)．その後，クエリを逐語訳したものと同定対象候補群の作品名をそれぞれ比較する（図1③）．最後に，同定対象候補群の中で類似度が閾値を超えているものをクエリの同一作品としてユーザへ提示する（図1④）．

3.1 原題の逐語訳の手順

ここでは，浮世絵作品の原題を逐語訳する方法について説明する．逐語訳の概要を図2に示す．

例)神奈川沖浪裏
↓ ①最長一致法により文字列を区切る
神奈川 | 沖浪 | 裏
↓ ②それぞれの辞書に対応した訳語に変換する
Kanagawa | offing wave | back
↓ ③全て連結する
Kanagawa offing wave back

図2：逐語訳の流れ

逐語訳全体の流れは次の通りである．まず原題表記の浮世絵作品名を辞書の見出し語との最長一致法により単語に分割する（図2①）．最長一致法については次節で説明する．次に，分割した各単語に対して対訳辞書を用いてそれぞれ逐語訳していく(図2②)．対訳辞書については3.1.2節で詳しく説明する．最後に分割していた訳語を連結することにより逐語訳を得る（図2③）．

3.1.1 最長一致法による単語分割

本提案手法では，原題表記の浮世絵作品名を対訳辞書で適切に翻訳するために，最長一致法を用いる．最長一致法

Identification of the same artwork across diverse Ukiyo-e databases in different languages using metadata
Taisuke Kimura†, Biligsaikhan Batjargal‡, Fuminori Kimura‡, Akira Maeda†
†College of Information Science and Engineering, Ritsumeikan University
‡Kinugasa Research Organization, Ritsumeikan University

[1] http://ukiyo-e.org/

とは，形態素解析においてよく使われる手法で，文字列を先頭から解析し，辞書に登録されている最長の単語を選択しながらマッチングを進める方法である．最長一致法の使用例を図 3 に示す．

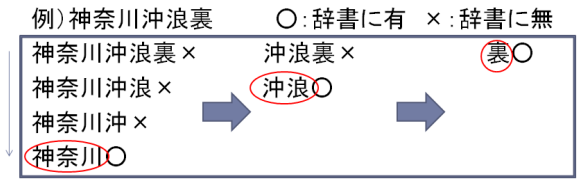

図 3：最長一致法を用いた単語分割の例

ここでは「神奈川」「沖浪」「裏」という単語が辞書に登録されていると仮定し，「神奈川沖浪裏」を解析する．図 3 に示すように，最初のステップでは「神奈川」が辞書に最長一致し，次のステップでは「沖浪」，最後のステップでは「裏」が最長一致する．結果として「神奈川 | 沖浪 | 裏」に分割される．

3.1.2 翻訳に使用する辞書

浮世絵作品名の原題は，現在では使用されない単語や，同じ単語でも読み方が異なる場合があるため，そのまま翻訳できない場合が多い．そこで，日英対訳辞書の「英辞郎第五版」，浮世絵関連語辞書（「日本演劇辞典」，「浮世絵大辞典」など浮世絵関連の辞書を電子化したもの），地名辞書（旧国名とその略称のペアを Web サイトの情報を参考に作成したもの）の 3 種類の辞書を用いて翻訳を行う．

日英対訳辞書は，主に固有名詞以外の名詞の英訳に使用する．浮世絵関連語辞書には，浮世絵作品名に頻出する語句の読みが含まれており，固有名詞を正しくローマ字で音訳化するために使用する．地名辞書は，旧国名とその略称を対応付けたもので，データベースによって異なる地名の表記を対応付けるために使用する．

3.2 原題の逐語訳と英訳の比較方法

ここでは，浮世絵作品名の原題と英訳の完全一致によるマッチングについての説明と，マッチングの結果を用いたスコア算出方法を説明する．

3.2.1 完全一致によるマッチング

原題を逐語訳したものと同定対象候補群の比較について説明する．ここでは，逐語訳の各単語と英訳の各単語の全ての組合せに対して文字列が完全に一致しているかどうかを判定する．比較対象は名詞のみとし，それ以外の品詞は比較対象としない．また，アルファベットの大文字・小文字の区別はしない．マッチングの一例を図 4 に示す．

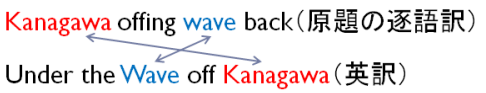

図 4：完全一致によるマッチング

図 4 のように，原題の逐語訳（神奈川沖浪裏）と同定対象候補の作品名の 1 つ（英訳）をマッチングした結果，それぞれの作品名に“Kanagawa”と“wave”が含まれているのが分かる．よって，完全一致数は 2 となる．

3.2.2 スコアの算出方法

原題と英訳の比較スコアの算出式は以下の通りである．

$$S = \frac{(w_1 N_1 + w_2 N_2)}{L}$$

ここで S をスコア，N_1 を固有名詞の一致数，N_2 を固有名詞以外の名詞の一致数，L を原題の逐語訳の単語数，w_1 を固有名詞の重み，w_2 を固有名詞以外の名詞の重みとする．

固有名詞を一般名詞と区別する理由として，固有名詞は多くの場合，作品を特定するための重要な情報であるためである．また，浮世絵作品名の英訳表記の中に多く使われており，原題を英訳化する際に一意に翻訳しやすいため，一般名詞と比較して曖昧性が少ないという特徴がある．よって，一般名詞よりも固有名詞にスコア比重を置く．

4 実験

提案手法による浮世絵作品の同一レコードの同定の精度を確認するために実験を行った．

4.1 実験方法

実験の準備として，江戸東京博物館のデータベース[1]にある葛飾北斎の浮世絵作品名の原題 13 件（全て富嶽三十六景のシリーズ作品）と，メトロポリタン美術館のデータベース[2]にある葛飾北斎の浮世絵作品名の英訳を 437 件用意した．なお，英訳 437 件の中には原題 13 件の同一作品（正解データ）が含まれている．そして，原題作品名を提案手法により逐語訳し，437 件の同定対象候補すべてと比較する．その際，スコア算出式の重みはw_1=2，w_2=1 とした．

4.2 実験結果

実験の結果，13 件の原題作品のうちランク 5 位以内に同一作品を含むものは 10 件であり，この場合の正解率は約 0.77 であった．また，同一作品がランク 1 位であったものは 7 件であり，この場合の正解率は約 0.54 であった．

4.3 実験結果の考察

実験結果より，ランク 1 位に正しく同一作品の同定ができたものは 13 件中 7 件であり，改善の余地は大きいと思われる．正しく同定できた例として，「武州千住」の逐語訳が“musashi province senju”となり，正解データの「Senju in Musashi Province」に対して固有名詞 2 つ，名詞 1 つがマッチした．同定できなかった例としては，「甲州犬目峠」が“Kai province inunometouge”と逐語訳され，正解データの“Fuji from Inume Pass”と一致しなかった．

5 まとめと今後の課題

本論文では，異言語の浮世絵データベースから作品名の原題と英訳を用いて同一作品を同定する手法を提案した．

提案手法の精度の改善案として，完全一致のみの比較だけではなく，“inunometouge”と“Inume”のような先頭一致する文字列を比較し，一定の文字数が一致すれば一致単語と認めるなどの手法を取り入れることが考えられる．

参考文献

[1] 久山岳夫, Biligsaikhan Batjargal, 木村文則, 前田亮：複数の異種浮世絵データベース間における同一浮世絵の同定手法の提案, 人文科学とコンピュータシンポジウム論文集, pp.225-232 (2013).

[2] 相澤彰子, 大山敬三, 高須淳宏, 安達淳：レコード同定問題に関する研究の課題と現状,電子情報通信学会論文誌, DI, Vol.J88-DI, No.3, pp.576-589 (2005).

[1] http://digitalmuseum.rekibun.or.jp/index.html
[2] http://www.metmuseum.org/collection/the-collection-online