多言語の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（立命館大学 情報理工学研究科）
Biligsaikhan Batjargal（立命館大学 総合科学技術研究機構）
木村 文則（尾道市立大学 経済情報学部）
前田 亮（立命館大学 情報理工学部）

本論文では，世界中に散在する浮世絵データベース間における同一の浮世絵作品の同定を行う手法について述べる．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ものであるが，多くの浮世絵の複製や異版が明治時代に海外に大量に散逸し，現在は世界中の美術館や博物館のデータベースにディジタルアーカイブとして公開されている．また，浮世絵は版画であるため，同一作品が多数存在しており，浮世絵のメタデータはデータベース毎に異なる言語や形式で表記されている．メタデータの言語や形式が異なる場合，同一の作品を検索するのは非常に困難である．そこで我々は，それらの問題を解決するために，これまで浮世絵作品名の原題と英訳を用いた同一作品の同定の研究などを行ってきた．本論文では，まだ研究が進められていない浮世絵作品名の原題・英訳・蘭訳を用いた同一作品の同定手法を提案する．

Identifying the Same Ukiyo-e Prints from Multiple Databases in Different Languages

Taisuke Kimura (Graduate School of Information Science and Engineering, Ritsumeikan University)
Biligsaikhan Batjargal (Research Organization of Science and Engineering, Ritsumeikan University)
Fuminori Kimura (Faculty of Economics, Management, and Information Science , Onomichi City University)
Akira Maeda (College of Information Science and Engineering, Ritsumeikan University)

This paper discusses the method for identifying the same Ukiyo-e prints from multiple databases in different languages. Ukiyo-e is a genre of woodblock prints made in the Edo period of Japan and it depicts people's daily life, scenery, and drama. Most of these copies and variants were scattered around the world in the 19th century, and are now stored in museums and galleries in many countries. Since the Ukiyo-e is a printmaking, there are many same artworks exist and text information of digitized Ukiyo-e prints are written in different languages and formats in each different databases. Therefore, it is difficult to find the same artworks. We have been doing research on the identification of the same artworks using the original titles and English titles of Ukiyo-e for supporting humanities research. In this paper, we propose a method to identify the same artworks using the original, English, and Dutch titles.

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．現代では浮世絵の芸術性の高さが見直され，美術品としての価値が高まっており，海外でも注目されている．近年，美術品や芸術作品がディジタル化され，ディジタルアーカイブとして保存されており，その一つとして浮世絵も含まれる．浮世絵は明治時代に海外に大量に流出し，それらの浮世絵作品のディジタル画像やメタデータが世界中のデータベースに所蔵されている．

表 1：同一作品の表記の違い

作品名	画像	データベース
『雪月花 隅田』（原題）		江戸東京博物館
Snow on the Sumida River, From the series, Snow, Moon, and Flowers (英訳)		大英博物館
De Sumida rivier in sneeuw (蘭訳)		アムステルダム国立美術館

また，各美術館・博物館で公開されている浮世絵のデータは表 1 に示すように様々な言語の違いがある．例えば浮世絵研究者が，ある浮世絵作品を公開されているデータベースに対して網羅的にするということが考えられる．しかしこのような状況においては，それを行うことは困難である．なぜなら，同一作品であっても言語が違うことにより同一作品であるとみなされないため，言語ごとに検索を行わなければならないからである．このような問題を解決するために，我々は異言語間浮世絵データベースにおける同一作品の同定手法を提案している[1][2]．これらの手法では，浮世絵の作品名を用いて，音訳（ローマ字）同士など同表記同士の比較や音訳と英訳の作品の比較，原題と英訳の作品の比較が行われ，高い精度での実験結果が得られているが，日本語と英語以外の言語はまだ対象とした研究が進められていない状況である．そこで，本研究では英語に次いで浮世絵データベースでの表記が多いオランダ語を加えた多言語の浮世絵データベース間における同一作品の同定手法を提案する．

2．関連研究

本研究で扱うような，異なるデータベースに存在する同一実体を表すレコードを自動的に見つけ出す問題は，「レコード同定」「レコード照合」などと呼ばれ，古くから研究が行われている．レコード同定に関する研究動向については，相澤ら[3]によるサーベイ論文がある．この論文では同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，本研究では異言語間のデータベースでのレコード同定となるため，従来の研究とは大きく異なる．同言語同士で比較を行う場合は，たとえば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．本研究ではこの課題の解決に取り組む．

一方，多数の浮世絵データベース中から同一作品を見つけることができる Web サイトとして Ukiyo-e.org[1]がある．このサイトで用いている手法と本論文の提案手法との違いは，Ukiyo-e.org では画像の類似度を用いた同定を行っているのに対し，本提案手法では浮世絵作品のメタデータによる同定を行っている点である．データベースによっては，一部のレコードに画像が無くメタデータのみ存在する場合があり，提案手法により Ukiyo-e.org では同定できなかった作品を抽出できる可能性がある．それに加えて，浮世絵には原画を修正し，出版された異版が存在する．異版は画像の特徴差が大きい場合があるため，類似画像検索の手法では区別することができない作品がある．上記のように，本提案手法はメタデータ処理ならではのメリットを持つ．

我々は，世界中のデータベースにある浮世絵情報を検索する複数言語に対応したシステムとして FeSSU（Federated Searching System for Ukiyo-e prints）[4]を構築している．本システムでは，ユーザが作者名や作品名などのクエリを入力すると，システムは各データベースに対して SRU（Search Retrieve via URL）またはスクレイピングを用いて検索を行い，クエリに関する浮世絵作品をユーザに提示する．SRU とは，横断検索用プロトコルのことで，検索要求情報を含んだ URL をサーバへ送り，その検索結果を XML

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

形式で返すものである．スクレイピングとはウェブサイトから必要な情報を自動で収集する処理のことを指す．また，このシステムは検索の際メタデータを多言語に翻訳し検索を行っており，世界中の多くの浮世絵データベースに対応した多言語横断アクセスを可能にしている．浮世絵横断検索システムの概要を図 1 に示す．本論文で提案する同一作品の同定手法は，将来的に本システムに組み込み，複数データベースから同一作品を提示する機能として実装する予定である．

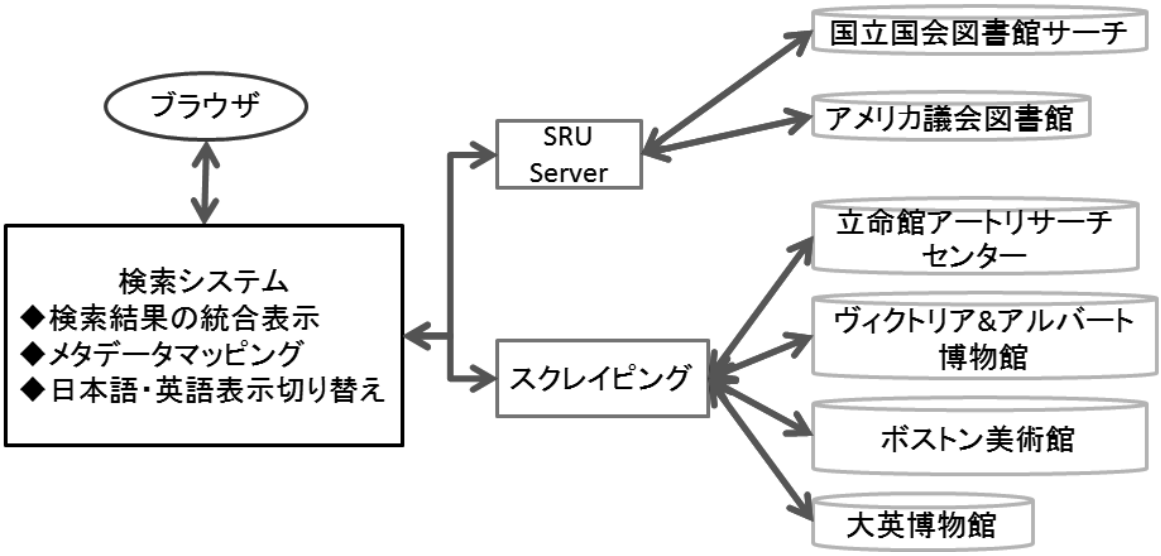

図 1：FeSSU の処理概要図

3．提案手法

本章では，今回対象とした日本語，英語，オランダ語の多言語データベース間における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 2 に示す．

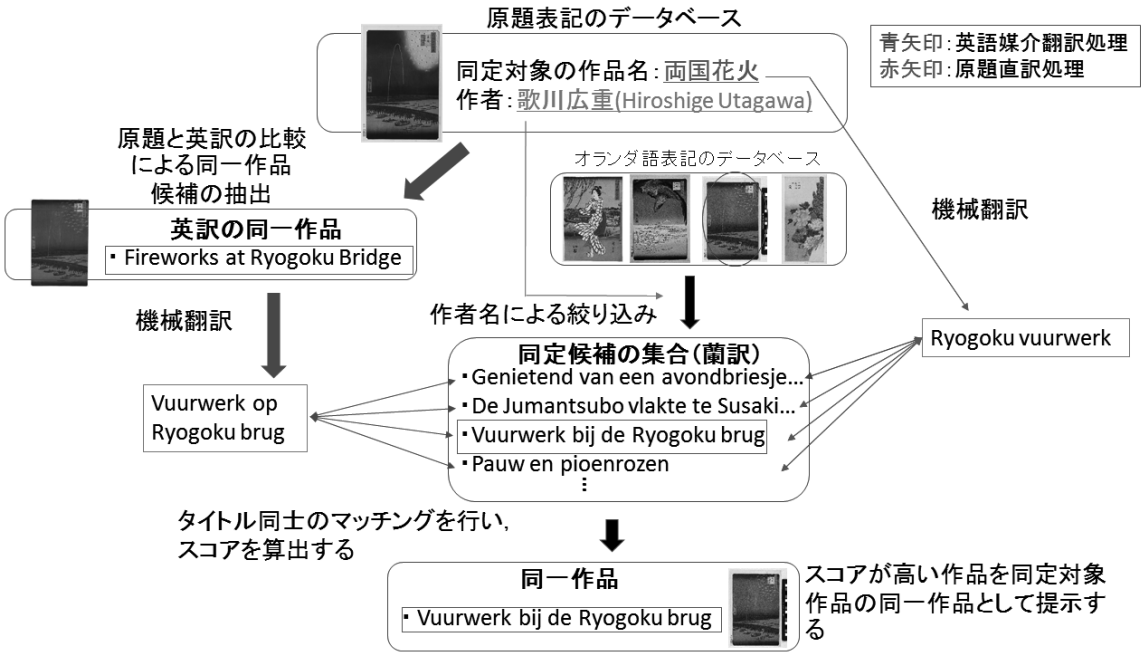

図 2：提案手法の流れ

提案手法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対象作品として浮世絵作品を一つ選択する．そして，選択した作品の作品名を用いて英語媒介翻訳処理と原題直訳処理（後述）を行い，二つの比較処理のスコアを統合し，その結果スコアが閾値を超えている作品をユーザに同定対象作品の同一作品として提示する．

3.1 浮世絵のタイトルの特徴

浮世絵のタイトルは大きく二つに分類することができる．すなわち，原題が直訳されたものと，浮世絵に描写されているものをタイトルに反映しているものである．図 3 に直訳的なタイトルの例を，図 4 に描写的なタイトルの例をそれぞれ示す．

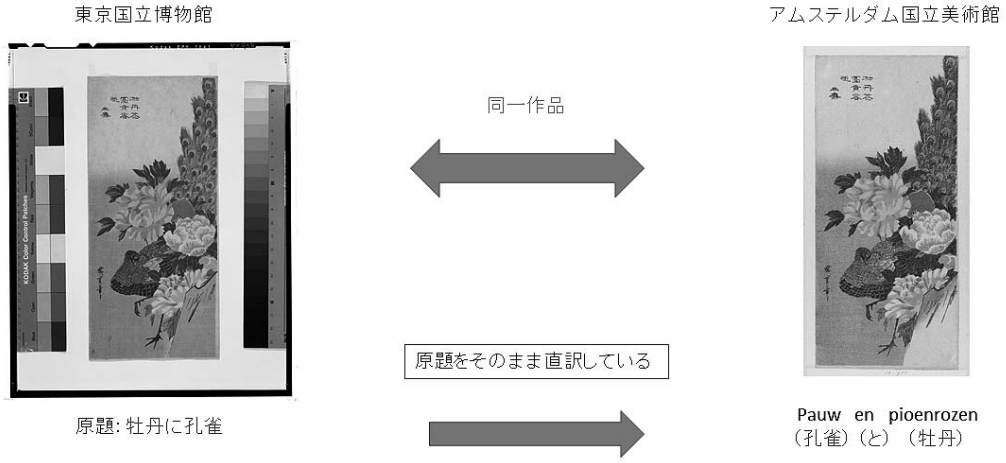

図 3：直訳的なタイトルの例

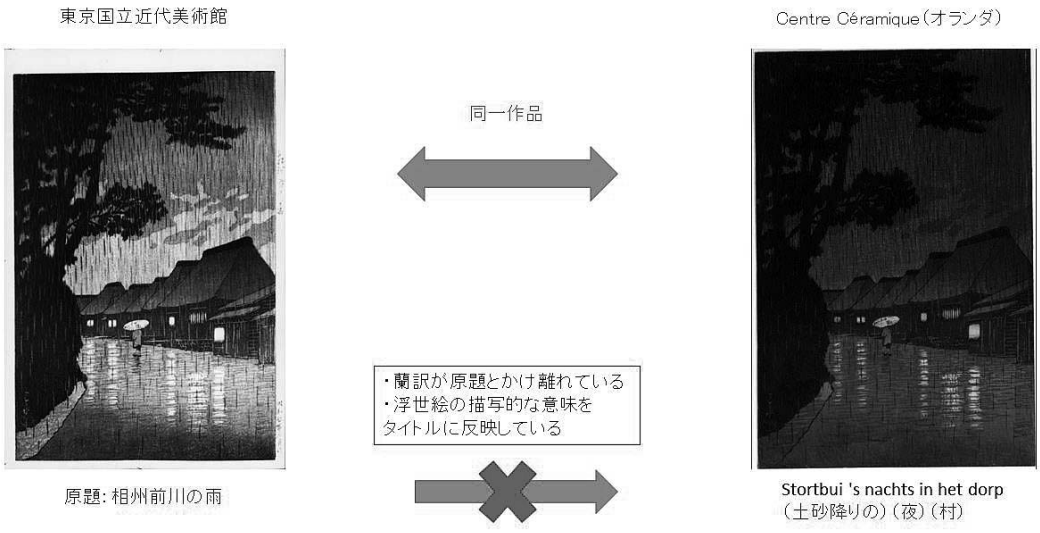

図 4：描写的なタイトルの例

図 3 は直訳的なタイトルの一例である. 図中左の浮世絵が東京国立博物館のデータベースにある歌川広重の『牡丹に孔雀』という作品である. そして, 図中右の浮世絵がオランダのアムステルダム国立美術館のデータベースにある同一作品である. タイトルに注目すると, 蘭訳の『Pauw en pioenrozen』の意味は「孔雀と牡丹」であり, 原題を忠実に翻訳していることがわかる.

図 4 は描写的なタイトルの一例である. 図中左の浮世絵が東京国立近代美術館に所蔵されている川瀬巴水の『相州前川の雨』という作品である. そして, 図中右の浮世絵がオランダの Centre Céramique のデータベースにある同一作品である. 図中右の浮世絵のタイトルは『Stortbui's nachts in het dorp』となっているが, これは日本語では「土砂降りの夜の村」のような意味である. しかし, 原題に含まれている「相州」や「前川」のような地名が蘭訳には含まれていない. また, 浮世絵を見ると激しい雨が降っている風景を描写していることが分かる. つまり, この蘭訳は原題を直訳したのではなく, 浮世絵画像の描写的な意味を反映していると言える.

タイトル同士を比較する際, 直訳的なタイトル同士と描写的なタイトル同士で場合分けをした比較方法が必要となる. 次項で詳細を説明する.

3.2 描写的なタイトルを基準とした比較

3.2.1 原題の同一作品（英訳）の抽出

原題の同一作品（英訳）の抽出について説明する. 抽出手法全体の概要を図 5 に示す.

抽出手法全体の流れは次の通りである. はじめに, 原題に対して日英対訳辞書を用いて英語に逐語訳する（図 5①）. 次に, 原題の作者名でデータベース B（英訳）から同定候補を絞り込む（図 5②）. 日本語表記以外のデータベースでは, 作者名はローマ字表記となっているので原題をローマ字に変換することで作者名を一致させることができる. そして, 絞り込んだ同定候補群の英訳と原題を逐語訳したものを, それぞれ固有名詞がマッチした際の重みを大きくする比較を行い, タイトル間の類似度のスコアを算出する(図 5③). そして, スコアが一定の閾値を超えている作品の英訳のタイトルを抽出する（図 5④）.

3.2.2 英訳と蘭訳の比較

英訳と蘭訳の比較の流れについて説明する. 3.2.1 項の手法で抽出した英訳を, 機械翻訳を用いてオランダ語に翻訳する. その後, 原題の作者名で絞り込んだ同定対象候補（蘭訳）とオランダ語に翻訳したものをそれぞれ比較してスコアを算出する.

図 6 に英訳をオランダ語に翻訳したものと蘭訳とのマッチングの一例を示す.

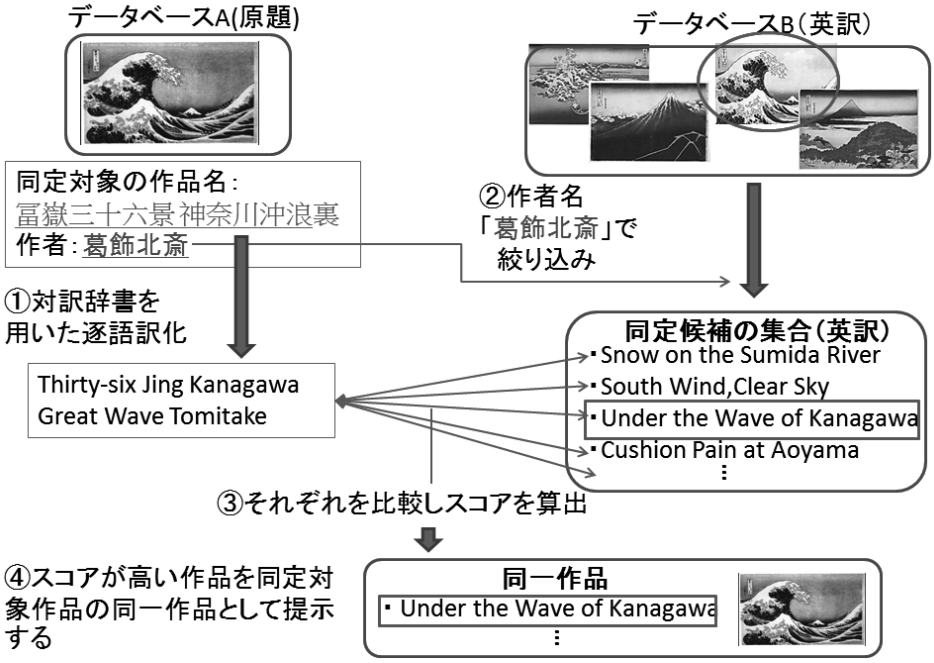

図 5：英訳抽出手法の流れ

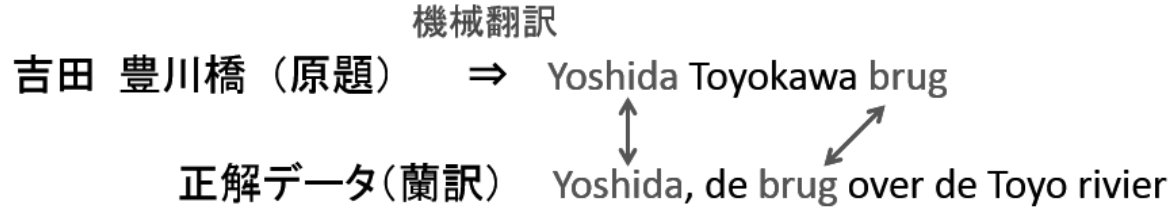

図 6：マッチングの一例

原題を英語に翻訳したものの“Yoshida”と“brug”が完全に一致しているので，これを一致単語とみなす．

続いて，スコアの算出方法を以下に示す．

$$S = \frac{N}{L}$$

ここで S をスコア，N を名詞（noun）の一致数，L を翻訳後のタイトルの名詞の数とする．図 7 の例をこの計算式に当てはめた場合，N に 2，L に 3 が代入されるので S は 0.6666 となる．

3.3 二つの処理の統合

原題直訳処理と英語媒介翻訳処理でそれぞれ抽出された同一作品候補を統合して，最終的な同一作品を決定する．統合の方法として，片方の処理にのみ同一作品候補があればそれを同一作品とし，両方の処理に同一作品候補があればスコアが高い方の作品を同一作品とする．

4．実験

第 3 章で述べた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

4.1.1 使用するデータ

実験の準備として，江戸東京博物館のデータベース[1]にある歌川広重の浮世絵 32 件の作品名の原題と，メトロポリタン美術館のデータベース[2]にある歌川広重の浮世絵 133 件の作品名の英訳と，アムステルダム国立美術館のデータベース[3]にある歌川広重の浮世絵 207 件の作品名の蘭訳を用意した．英訳 133 件，蘭訳 207 件の中には原題 32 件の同一作品が含まれている．

4.1.2 翻訳に使用するもの

機械翻訳には Microsoft 社の Microsoft Translator API を使用した．また，原題と英訳の比較による同一作品候補の抽出では日英対訳辞書の「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った．

4.1.3 重みと閾値の設定

原題と英訳の比較による同一作品候補の抽出では固有名詞の重みを 2,固有名詞以外の名詞を 1 として計算している．

今回の実験では，全ての処理においてスコアが 0 より大きければそれらは同一作品候補の条件に当てはまるものとしている．

4.2 実験結果

実験の結果を表 2，図 7 に示す．表 2 の左から 4 列目の「処理の統合」は「原題直訳処理」と「英語媒介翻訳処理」の統合結果である．右端の列は描写的なタイトルに絞った結果を示している．原題 32 件の中には描写的タイトルが 10 件含まれており，その 10 件の中から英語媒介翻訳処理にて何件抽出できたかを示している．図 7 は表 2 の正解率を棒グラフで表したものである．

4.3 実験結果の考察

表 2 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 32 件中 21 件であり，改善の余地は大きいと思われる．原題直訳処理で正しく同定できた例として，原題の『両国花火之図』を直訳した結果が『yogoku vuurwerk, figuur』となり，正解データの『Vuurwerk bij de Ryogoku

表 2：実験結果の正解データに対する正解数と正解率

	原題直訳処理	英語媒介翻訳処理	処理の統合	描写的なタイトルの抽出
正解データがスコア 1 位の件数（割合）	19/32(0.5937)	12/32(0.375)	21/32(0.6562)	7/10(0.7)
正解データがスコア 5 位以内の件数（割合）	20/32(0.625)	14/32(0.4375)	22/32(0.6875)	7/10(0.7)

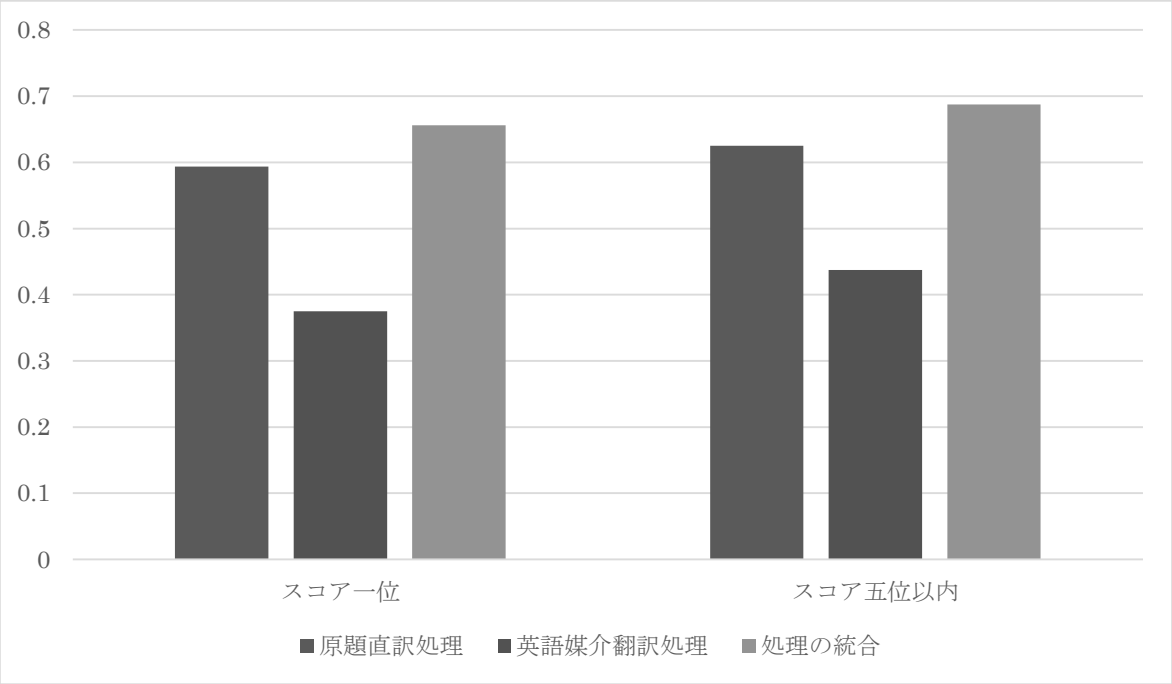

図 7：実験結果の正解データに対する正解率

brug』に対して名詞が二つマッチした．同定できなかった例としては『水口 名物干瓢』が“mizuguchi specialiteit transportband”と翻訳され，正解データの『Minakuchi, het beroemde streekproduct kanpyo』と全く一致しなかった．

英語媒介翻訳処理では，一例として『深川洲崎十万坪』の同一作品を抽出することができた．これは，抽出した英訳の『Jumantsubo Plain at Susaki，Fukagawa』が“Jumantsubo Plain op Fukagawa Susaki”と翻訳され，正解データの『De Jumantsubo vlakte te Susaki bij Fukagawa』と名詞が三つマッチしたためうまく同定できたと考えられる．しかし，今回の描写的タイトル抽出の結果からは描写的な単語のマッチングは確認できなかった．例を挙げると，抽出された英訳の『View of the Kanagawa station at sunset』を機械翻訳すると『Weergave van het Kanagawa station bij zonsondergang』となったが，これを正解データの『Kanagawa, bergopwaarts』と比較したところ“Kanagawa”以外の部分はマッチしなかった．この原因として，“sunset”は日没，“bergopwaarts”は下り坂を意味するように，お互い描写的な単語ではあるものの，それぞれの意味が翻訳前から異なっていたことが考えられる．

５．おわりに

本論文では，日本語とオランダ語を対象とした多言語の浮世絵データベース間における同一作品の同定手法を提案した．これまでの研究では日本語と英語表記のレコードのみ同一作品の同定が可能だったが，本手法を用いることにより，オランダ語表記のデータベースも含めた浮世絵レコード同定が可能となったことに加えて，描写的なタイトルにも対応できるようになった．提案手法の精度確認の実験では，描写的なタイトルに対して有意な結果が得られたが，改善の余地はまだあると思われる．今後の課題として，同定精度の向上や他の言語に対応した同一作品の同定手法の構築が挙げられる．

謝辞

本研究の一部は，文部科学省科学研究費補助金基盤研究（C）「多言語ディジタルアーカイブの統合検索に関する研究」（研究代表者：前田亮，課題番号：24500300）の支援を受けている．

参考文献

1) 木村 泰典，Biligsaikhan Batjargal，木村文則， 前田 亮：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，第 77 回情報処理学会全国大会講演論文集，第 4 分冊，pp.639-640 (2015).
2) 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 2013 論文集，pp.225-232 (2013).
3) 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).

4) Biligsaikhan Batjargal, Fuminori Kimura, and Akira Maeda : Metadata-related Challenges for Realizing Federated Searching System for Japanese Humanities Databases. Proc. *11th International Conference on Dublin Core and Metadata Applications (DC-2011)*, pp. 80-85 (2011).

1 http://digitalmuseum.rekibun.or.jp/index.html

2 http://www.metmuseum.org/collection/the-collection-online

3 https://www.rijksmuseum.nl/en