**Finding the Same Artworks from Multiple Databases in Different Languages**

Taisuke Kimura

College of Information Science and Engineering, Ritsumeikan University, Japan

Biligsaikhan Batjargal

Kinugasa Research Organization, Ritsumeikan University, Japan

Fuminori Kimura

Kinugasa Research Organization, Ritsumeikan University, Japan

Akira Maeda

College of Information Science and Engineering, Ritsumeikan University, Japan

**1. Introduction**

This paper discusses a method for identifying the same artworks across multiple databases using textual metadata written in different languages. As more and more libraries, museums, galleries and archives are making their collections available online, it is becoming essential to develop methods for accessing these vast and valuable collections of cultural heritage easily and thoroughly.

Although some databases are available online through standard access methods or APIs, there are many databases that still use non-standard metadata schemas and interfaces. Even if standard access methods are available, there still exist metadata inconsistencies in different languages. These problems hinder the development of a search system for accessing multiple databases in different languages simultaneously.

We aim to identify the same artworks in different languages since the most of the existing approaches such as *record linkage* [1], *duplicate detection* [2] or supervised machine learning techniques [3] consider mainly the same language.

Recently, we have developed a method for identifying the same Ukiyo-e prints from databases in English and Japanese [4]. This method is particularly useful since the Japanese traditional woodblock printing – *Ukiyo-e* is engraving and many copies or variants of one particular work were made from the same woodblock, and most of these copies were scattered around the world in the 19th century, and now stored in museums and galleries in many countries. Most of the metadata of these databases are available only in English or in the native language of that country. Titles are mostly written either as the transliteration of the original Japanese title, or a translation in that language. Table 1 shows some examples of databases in Europe that contain Ukiyo-e in their collections, and the languages in which the title is written.

One of the effective approaches for identifying the same artworks from multiple image databases is to utilize image similarity calculations. Ukiyo-e.org[[1]](#footnote-1)[5] is the most successful example of identifying the same Ukiyo-e prints. It purely uses image similarities rather than textual data. Our textual-metadata-based approach and image-similarity-based approach both have advantages and disadvantages. One of the advantages of our approach is that we do not have to harvest all the data from the databases beforehand. Furthermore, artefacts other than printings might not be suitable for using 2-dimensional image similarities. Perhaps, it might be better to combine both methods for obtaining the most accurate results.

Table 1. The same Ukiyo-e print in different databases

|  |  |  |
| --- | --- | --- |
| **Original Ukiyo-e print** | **Title** | **Database** |
| 説明: Katsushika Hokusai: Under the Wave off Kanagawa (Kanagawa-oki nami-ura), also known as the Great Wave, from the series Thirty-six Views of Mount Fuji (Fugaku sanjûrokkei) - Museum of Fine Arts  **神奈川沖浪裏**  (original title in Japanese) | 神奈川沖浪裏  (original title in Japanese) | The Edo-Tokyo Museum |
| Kanagawa oki nami ura  (transliteration)  The great wave off shore of Kanagawa (in English) | The Library of Congress |
| The great wave off Kanagawa  (in English) | National Gallery of Victoria |
| * La grande vague * Sous la grande vague au large de Kanagawa * Sous la vague au large de Kanagawa   (in French) | French Photo Agency |
| De grote golf bij Kanagawa  (in Dutch) | Netherlands State Museum |
| De grote golf bij KanegawaFugaku Sanjrokkei  (in Dutch) | Centre Céramique |
| * Große Woge * Der Fuji hinter der großen Woge * Die Welle   (in Germany) | Bildarchiv Foto Marburg |

**2. Proposed approach**

We extend our previous method that identifies the same artworks between transliterated Japanese title and its English translation so that it can identify the same artworks between the original Japanese title and its translations in other European languages. Our method utilizes proper nouns in a title as key elements for matching, and it also utilizes other words’ translations to further improve the matching accuracy. If a given word in the title is not a proper noun, we perform a literal translation using bilingual dictionaries, i.e., all the words excluding the proper nouns are translated. The similarity degree is based on the weighting of matching words between titles in different languages. The similarity degree increases as many matching words as the titles contain. Not only the weighting of matching words but also partial string matching score is considered in the similarity calculation.

Figure 1 illustrates how the same artworks are identified from databases in different languages. First, the metadata elements are translated (1) as explained above. Then, the artworks are filtered (2) by an artist name by utilizing our previous method that uses artist’s various notations or aliases [6]. The proposed method calculates the similarity degree (3) and the artworks that have higher scores are treated as the same (4) as the given artwork.

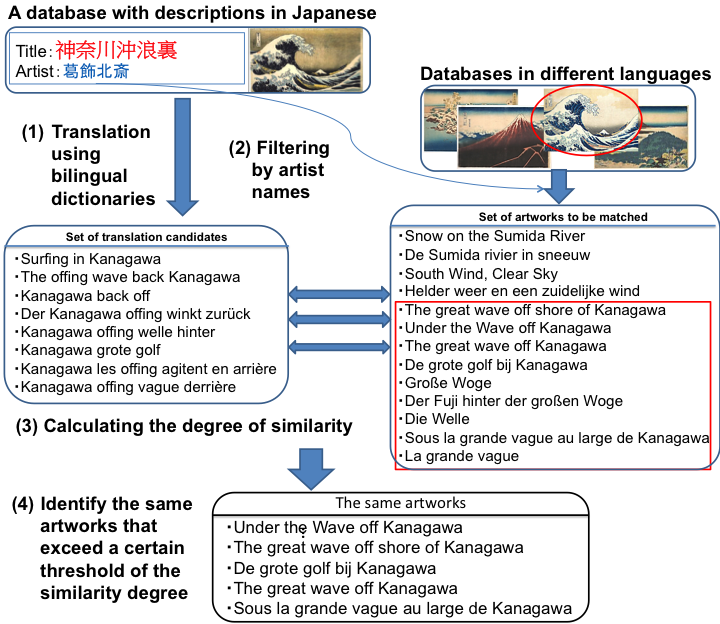
Although this method has been tested only on Ukiyo-e databases, it does not depend on a particular type of art. Besides, since this method requires only bilingual dictionaries as the language resource, it can be applied to other languages easily.

Figure 1. An illustration of the proposed method

**3. Conclusion**

In this paper, we proposed a method for identifying the same artworks across multiple databases using textual metadata written in different languages. The proposed method would be useful for both humanities researchers and database administrators. It provides researchers easy and efficient ways of finding the same artwork of their interest in other databases regardless of the language. For database administrators, aggregating multiple metadata of the same artwork from various databases makes it possible to correct and/or enrich the existing metadata records for improving the quality of their database. Besides, the proposed method could be used to link the same or similar artworks across different databases, and it will contribute to enriching the Linked Open Data in the field of humanities.

**References**

[1] I. P. Fellegi and A. B. Sunter, “A Theory for Record Linkage,” *Journal of the American Statistical Association*, vol. 64, no. 328, p. 1183, Dec. 1969.

[2] M. Bilenko and R. Mooney, “On evaluation and training-set construction for duplicate detection,” in *Proceedings of the KDD-03 Workshop on Data Cleaning, Record Linkage, and Object Consolidation*, pp. 7–12, 2003.

[3] A. Elmagarmid, P. Ipeirotis, and V. Verykios, “Duplicate Record Detection: A Survey,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 19, no. 1, pp. 1–16, Jan. 2007.

[4] B. Batjargal, T. Kuyama, F. Kimura, and A. Maeda, “Identifying the Same Records across multiple Ukiyo-e Image Databases Using Textual Data in Different Languages,” in *Proceedings of Digital Libraries 2014: ACM/IEEE Joint Conference on Digital Libraries* (*JCDL 2014*) *and International Conference on Theory and Practice of Digital Libraries* (*TPDL 2014*), Sep. 2014.

[5] J. Resig, “Aggregating and analyzing digitized Japanese woodblock prints,” in *3rd Annual Conference of the Japanese Association for Digital Humanities*, Sep. 2013.

[6] B. Batjargal, T. Kuyama, F. Kimura, and A. Maeda, “Linked data driven multilingual access to diverse Japanese Ukiyo-e databases by generating links dynamically,” *Literary and Linguistic Computing*, vol. 28, no. 4, pp. 522-530, Dec. 2013.

1. http://ukiyo-e.org/ [↑](#footnote-ref-1)