**How contemporary deep learning models describe early Japanese photographs**

In 2007, the Musée Guimet acquired an exceptional collection of early photographs of Japan from the collector Joseph Dubois. Comprising more than 19,000 phototypes produced between 1860 and 1920, the Dubois collection is one of the largest of its kind in the world. It is particularly representative of the 'Yokohama School' (or Yokohama Shashin) style, which was aimed at a foreign clientele and characterised by the representation of a feudal and traditional Japan in settings that were as refined as they were artificial, often hand-coloured by Japanese painters.

Raimund von Stillfried-Ratenicz & Hermann Andersen, *Views & Costumes of Japan*, photo album, 1877–1880. MNAAG, Joseph Dubois collection, AP11325.

The advent of visual deep learning models, leveraging both convolutional and transformer-based architectures, has revolutionized the analysis of photographic collections. Trained on vast arrays of internet-sourced photographs, these models have become highly efficient and widely accessible as open-source tools. This technological leap has provided unprecedented capabilities in describing, categorizing, and understanding contemporary photographic collections (Wevers, 2020; Lin Diu, 2023). However, a significant limitation arises when applying these models to historical photographs.

Often, it is observed that these state-of-the-art systems fall short in analysing historical images, primarily due to their training on modern, predominantly Western photographic datasets. As a result, these models possess a descriptive vocabulary that is too limited for an in-depth exploration of photographs that are distant from the training set, both temporally and geographically. Consequently, while these models offer powerful tools for contemporary image analysis, their application to historical and culturally diverse photographic collections necessitates careful consideration and potentially supplementary training or adaptation to accurately capture the rich and varied nuances present in such archives.

The corpus

The collection, consisting of 18,155 Japanese photographs from various sources including personal collections, was scanned and described with 2,209 different keywords during a previous study. However, this digitisation does not meet archival quality standards, both in terms of image quality and descriptive metadata. To address these issues, professional scanning and archiving of the images is currently underway. The descriptive tags cover a wide range of identifiers, including

content terms (such as 'child', 'river', 'fan'), elements of Japanese culture (such as 'geisha', 'shamisen', 'torii'), types of photographs ('group portrait', 'landscape'), and specific Japanese locations or monuments (Nikko, Yokohama, Toshogu Shrine).

## Methodology

We evaluated the automatic tagging of a sample of Japanese photographs using multi-modal LLMs. The original list of tags was reduced to common names in French using a dictionary, resulting in a complete list of 1,034 tags. This list excluded personal names, place names and words in Japanese. We then randomly selected six tagged albums, representing a total of 275 images, not all of which contained photographs as some contained covers or blank pages. These images were tagged with 190 different tags (after filtering out only the common names).

To evaluate the automatic tagging process, we tested three different multimodal LLMs: CLIP (Radford, 2021), SIGLIB (Zhai, 2023) and ChatGPT (OpenAI, 2023). The prompt for each model was identical, instructing them to tag each photo with the most appropriate tags from the list provided. Two experiments were conducted: one using the reduced tag list of 190 tags from the six albums, and another using the full list of 1,034 tags.

## Results

The results of these experiments showed that both CLIP and SIGLIB achieved very low F1 scores on the reduced tag set (12% F1) and the full tag set (9% F1). In contrast, ChatGPT performed better, achieving an F1 score of 54% on the reduced tag set and 22% on the full tag set. Specifically, ChatGPT achieved 75% precision and 43% recall on the reduced tag set and 26% precision and 21% recall on the full tag set.

To better understand the performance of ChatGPT with the full set of tags, we manually checked its predictions and validated the tags when they were correct but not in the original annotation. After this validation, we recalculated the scores of the model. We found that ChatGPT's performance increased significantly, with precision rising to 80%, recall to 43% and F1 score to 56%. This suggests that ChatGPT is able to predict correct tags 80% of the time, but predicts fewer tags than a human annotator. This experiment also highlights the subjectivity inherent in human tagging and underlines the difficulty of automatically evaluating tagging performance. It shows that predicted tags can be correct even if they were not selected by the human annotator, making it difficult to evaluate the accuracy of automatic tagging.

## References:


- Bennett, Terry. *Photography in Japan 1853-1912* [Revised second edition.] ed. North Clarendon VT: Tuttle Publishing. 2023
- Estèbe, Claude. *Les premiers ateliers de photographie japonais 1859-1872*, Études photographiques, 19, 4-27. 2006
- Wevers, Melvin and Smits, Thomas . *The visual digital turn: Using neural networks to study historical images*. Digital Scholarship in the Humanities, Volume 35, Issue 1, April 2020, Pages 194–207.
- Du, Lin and Le, Brandon and Honig, Edouardo. 2023. Probing Historical Image Contexts: Enhancing Visual Archive Retrieval through Computer Vision. J. Comput. Cult. Herit, 2023.
- GPT-4 Technical report, OpenAI, https://arxiv.org/abs/2303.08774
- Radford, Alec et al. “Learning Transferable Visual Models From Natural Language Supervision.” International Conference on Machine Learning (2021).

- X. Zhai, B. Mustafa, A. Kolesnikov and L. Beyer, "Sigmoid Loss for Language Image Pre-Training," in 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2023.