Multimodal Search on Iconclass using Vision-Language Pre-Trained Models

Cristian Santini*
cristian.santini@fiz-karlsruhe.de
FIZ Karlsruhe
Institute AIFB, Karlsruhe Institute of Technology
Germany

Etienne Posthumus*
etienne.posthumus@partners.fiz-karlsruhe.de
FIZ Karlsruhe
Germany

Tabea Tietz*
tabea.tietz@fiz-karlsruhe.de
FIZ Karlsruhe
Institute AIFB, Karlsruhe Institute of Technology
Germany

Mary Ann Tan*
ann.tan@fiz-karlsruhe.de
FIZ Karlsruhe
Institute AIFB, Karlsruhe Institute of Technology
Germany

Oleksandra Bruns*
oleksandra.bruns@fiz-karlsruhe.de
FIZ Karlsruhe
Institute AIFB, Karlsruhe Institute of Technology
Germany

Harald Sack*
harald.sack@fiz-karlsruhe.de
FIZ Karlsruhe
Institute AIFB, Karlsruhe Institute of Technology
Germany

ABSTRACT

Terminology sources, such as controlled vocabularies, thesauri and classification systems, play a key role in digitizing cultural heritage. However, Information Retrieval (IR) systems that allow to query and explore these lexical resources often lack an adequate representation of the semantics behind the user's search, which can be conveyed through multiple expression modalities (e.g., images, keywords or textual descriptions). This paper presents the implementation of a new search engine for one of the most widely used iconography classification system, Iconclass. The novelty of this system is the use of a pre-trained vision-language model, namely CLIP, to retrieve and explore Iconclass concepts using visual or textual queries.

CCS CONCEPTS

• **Information systems → Novelty in information retrieval**; Digital libraries and archives.

KEYWORDS

art history, classification systems, information retrieval, multimodal search, vision-language models

1 INTRODUCTION

Classification systems and other terminology sources, such as controlled vocabularies and thesauri, play a key role in the digital curation of cultural heritage objects. These lexical resources, are used by Galleries, Libraries, Archives and Museums (GLAM) to describe items in their collection; the Getty Vocabularies [1] [Harpring 2010] or the Social History and Industrial Classification (SHIC) system[2] are some examples in the GLAM domain. Iconclass[3] [Waal 1968] is the *de facto* standard for the classification of artistic representations and images. It is organized as a hierarchy which comprises 10 main categories, or notations, (abstract and non-representational art, religion and magic, etc.) and each category has several narrower terms as descendants in a tree-like structure. Notations in Iconclass are used to describe and index artworks based on their depicted objects (animals, deities, etc.) or artistic themes (biblical, mythological, etc.). Each notation is associated with a text which provides a label to the notation.

Currently, some limitations affect the accessibility of the digital platforms through which many controlled vocabularies, thesauri and classification systems are published. Specifically, the adoption of Information Retrieval (IR) techniques based on word occurrence, e.g. Term Frequency-Inverse Document Frequency (TF-IDF) or set theory, e.g. Boolean search, to retrieve entries in these lexical resources may create usability issues for non-expert users which are not familiar with the terminology of iconography in the Western arts. In order to solve this problem, word embeddings can be used to mitigate the aforementioned limitation, since they are able to encode word meanings into low-dimensional vectors, which alleviates the need for exact term matches.

The main contribution of this work is the implementation of a multimodal search engine for Iconclass. This system is designed to leverage a vision-language model, i.e. OpenAI's CLIP [Radford et al. 2021], and a visual similarity search [Posthumus and Sack 2022], to retrieve Iconclass concepts, or notations, based on either text or image inputs from the user. Section 2 of this work discusses the adapted pre-trained Vision-Language model and the implemented search system, Section 3 contains results from a small preference-based survey conducted among Iconclass users of varying art historical background and Section 4 concludes the paper.

2 MULTIMODAL IR IN ICONCLASS

Following the development of deep learning architectures trained on a single expression modality, which might consists either of texts, images, video, audio or graphs, researchers have recently investigated the potentials of training neural models on emerging patterns which can be expressed across multiple modalities, e.g. with both visual and textual features. CLIP (*Contrastive Language-Image Pre-training*) [Radford et al. 2021] follows this intuition. Trained on a

*All authors contributed equally to this research.
[1]https://www.getty.edu/research/tools/vocabularies/
[2]https://www.shcg.org.uk/About-SHIC
[3]www.iconclass.org

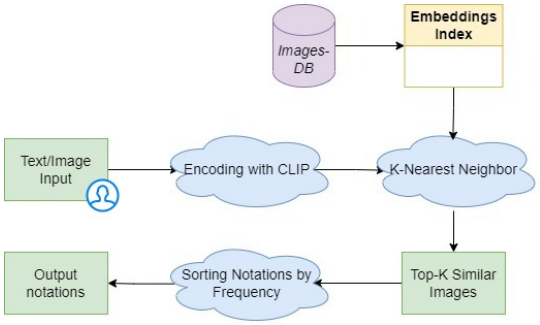

Figure 1: Diagram of the system architecture.

dataset of 400M *(image-caption)* pairs collected from the internet, CLIP, given an image, must predict which out of a set of 32,768 randomly sampled text snippets it was paired with in the dataset. The main idea behind this model is to use two different encoders, one for images and one for texts, and by formulating the learning objective as a contrastive loss, the model is able to align text and images into a shared embedding space to learn diverse visual concepts.

The new multimodal search system in Iconclass takes advantage of the aforementioned vision-language model to find appropriate Iconclass notations given a text or image as input. The system consists of a database of ≈500K images of artworks and photographs, hereinafter referred to as *Images-DB*, and a Faiss index [Johnson et al. 2017] of 2GB which stores the embeddings of the images from the aforementioned database, retrieved from the pre-trained CLIP model of [Radford et al. 2021]. *Images-DB* collects 531,172 images described with Iconclass notations, which were kindly provided by the Arkyves initiative[4]. The database contains 2,526,145 Iconclass notations, of which 90,347 were unique.

For visual search, the text or image initially given by the user as input is encoded into a multidimensional vector by using either the text or image encoder of CLIP. The obtained embedding is then used as seed for a similarity search on the Faiss index, which returns the top-K most similar annotated images in *Images-DB*, by using a k-nearest neighbor algorithm. The output of the similarity search is then the set of similar images and the list of notations used to describe them, sorted by the number of assigned images in reversed order. A diagram of the underlying algorithm of the system is presented in Figure 1.

It is important to note that a user does not directly query the complete Iconclass hierarchy, but a set of annotated images. This represents a novelty since it indirectly exploits the information contained in the annotation of ≈500K samples in *Images-DB*, which was carried by expert annotators. This multimodal search aims to exploit semantic similarity between text and images to retrieve, from a candidate set of images, potentially relevant notations provided by human experts, which is a relevant feature for users which are not well-versed in iconography. The new multimodal search engine was made available on a public demo[5].

[4]www.arkyves.org
[5]https://github.com/ISE-FIZKarlsruhe/iconclass/tree/main/multimodal

	Multimodal Search	TF-IDF
#Preferences	105	104
#Preciseness	64	72
#Exhaustiveness	30	17

Table 1: Results from preference-based survey aimed to compare visual-similarity search and TF-IDF search for Iconclass.

3 PREFERENCE-BASED EVALUATION

Currently, Iconclass provides a TF-IDF based text search. Through a preference-based survey, the proposed multimodal search engine (*System A*) was compared to the current approach (*System B*). For objectivity, the users were unaware of the system designation. 10 participants were gathered with a public call for volunteers. Each respondent had to fill a spreadsheet containing overall 10 artwork images and 25 query strings. Given a query string, the top-10 results of system A and B were placed side-by-side. The users were then asked to select their preferred results, and to specify the reason for this preference: *Preciseness*, the correctness of the returned notations (with respect to the image), and *Exhaustiveness*, the recall of valid notations in the result list. Results from the survey are reported in Table 1.

Overall, respondents did not express a marked preference for one system over another. However, from the survey emerged that, when multimodal search was preferred, it was mainly due to exhaustiveness of the results. This result may derive from the fact that CLIP-based search does not aim to retrieve a single *pinpoint* notation but a range of Iconclass codes related to a visual concept. For example, the query string *Street* returns, when using multimodal IR, Iconclass notations which describe not only this iconographic element (*25I141: street*), but also some related elements which are likely to occur in pictures of streets, such as humans (*31D14: adult man*) or animals (*34B11: dog*). The same does not happen for TF-IDF.

4 DISCUSSION AND CONCLUSION

This paper presents the new multimodal search engine of Iconclass. This system leverages pre-trained vision-language embeddings and a database of human-annotated images to return Iconclass notations based on either textual or visual inputs. The multimodal search engine offers users the advantage to query Iconclass with both images or free-text descriptions, which is a relevant feature to image curators which are not accustomed to the underlying vocabulary in Iconclass. However, the quality of CLIP-based results is still to be adequately estimated by using objective ground-truths. As a consequence, multimodal search can be exploited in order to complement search results from TF-IDF, rather than be considered as an equivalent alternative. As future work, the possibility to combine results coming from different search systems for Iconclass, e.g. both from visual similarity and word similarity, can and will be taken in consideration. Moreover, new features and services for Iconclass users will be introduced, such as the publication of a SPARQL endpoint to enable external web services to exploit the new similarity-based search.

REFERENCES

Patricia Harpring. 2010. Development of the Getty Vocabularies: AAT, TGN, ULAN, and CONA. *Art Documentation: Journal of the Art Libraries Society of North America* 29, 1 (2010), 67–72. https://www.jstor.org/stable/27949541 Publisher: [The University of Chicago Press, Art Libraries Society of North America].

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2017. Billion-scale similarity search with GPUs. https://doi.org/10.48550/arXiv.1702.08734 arXiv:1702.08734 [cs].

Etienne Posthumus and Harald Sack. 2022. The Art Historian's Bicycle Becomes an E-Bike. (2022).

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. https://doi.org/10.48550/arXiv.2103.00020 arXiv:2103.00020 [cs].

H. van de Waal. 1968. *Decimal index of the art of the Low Countries; D.I.A.L* (abridged ed. of the iconclass system ed.). Rijksbureau voor Kunsthistorische Documentatie, The Hague. OCLC: 27696.