Semantic Content Description and Object Detection: How Contemporary Deep Learning Models Describe Early Japanese Photographs

Christopher Kermorvant - TEKLIA

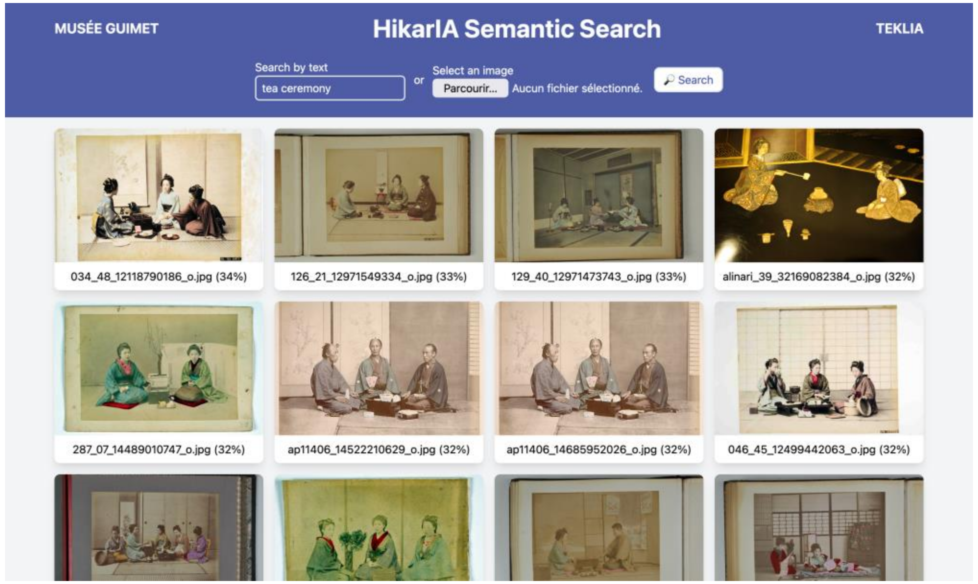
Musée national
des Arts asiatiques - Guimet

The Dubois Collection at Guimet Museum

- 284 albums, 19,000 photographs, one of the largest collection in the world
- A broad overview of the commercial photographic production in Japan during Bakumatsu and Meiji eras (1853 – 1912)
- A great variety of materials, techniques, styles, subject matters, or price ranges

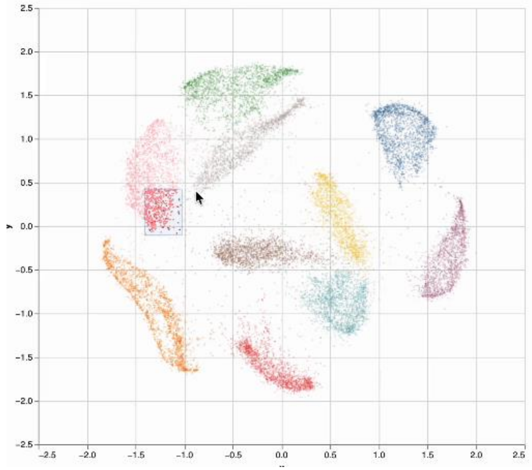
The HikarIA project

- Develop new tools based on Deep Learning to explore and analyse large collections of historical photographs
- Recognize and accurately name iconographical elements in the photographs
- Identify recurring photographs and similar compositions
- Detect potential patterns in image sequences
- Generate detailed statistical analysis



Search by text: tea ceremony

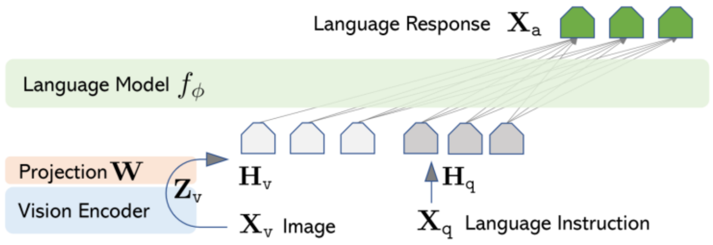
034_48_12118790186_o.jpg (34%)
126_21_12971549334_o.jpg (33%)
129_40_12971473743_o.jpg (33%)
alinari_39_32169082384_o.jpg (32%)
287_07_14489010747_o.jpg (32%)
ap11406_14522210629_o.jpg (32%)
ap11406_14685952026_o.jpg (32%)
046_45_12499442063_o.jpg (32%)



The HikarIA project

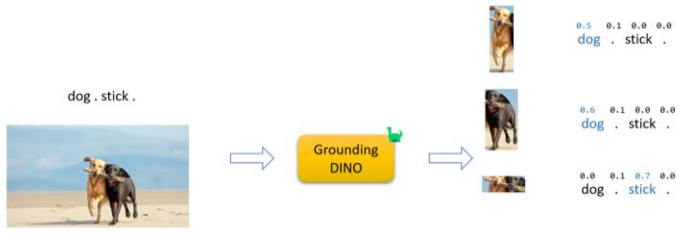
How the AI models trained on contemporary western photographs « see » the Japanese photographs from the XIXth century ?

Captioning



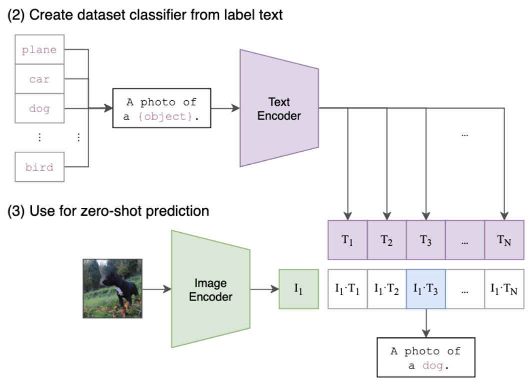
LLaVA/ChatGPT

Zero-shot object detection



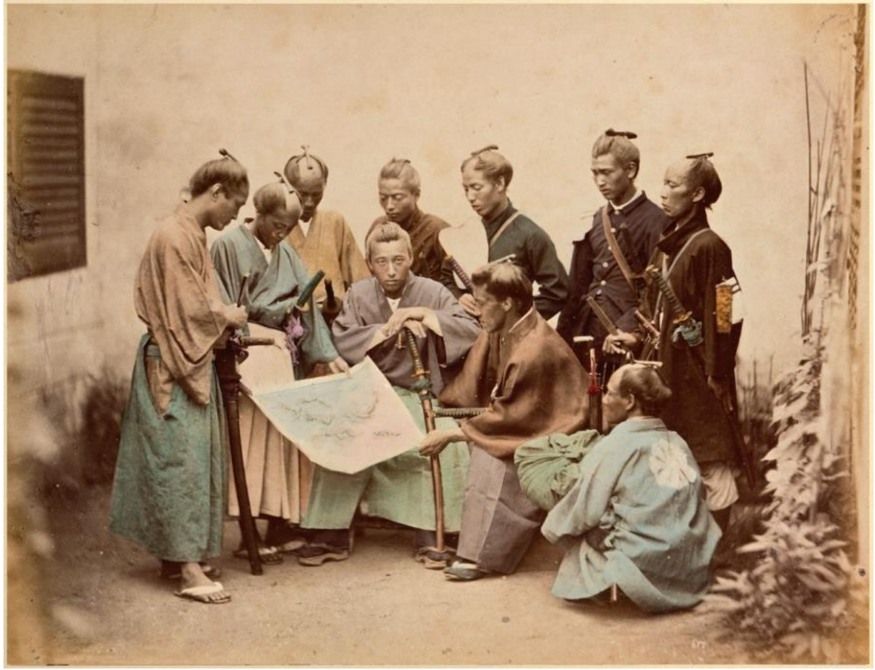
Grounding DINO

Semantic image search



CLIP

Image Captioning



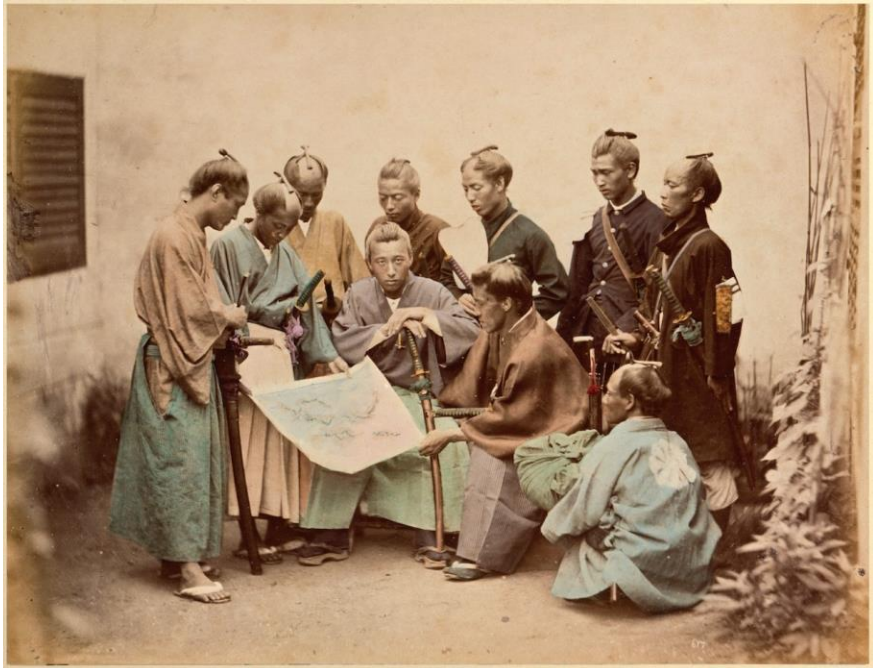
Prompt: This image is a 19th century Japanese photograph, describe it.

Answer:

The image is a black and white photograph of a group of people, possibly samurai, gathered around a map. They are dressed in traditional Japanese clothing, with some of them wearing kimonos. The group consists of at least nine individuals, with some standing and others sitting on the ground.

The map is placed in the center of the scene, and the people are positioned around it, engaged in conversation or studying the map. The photograph captures a moment of collaboration and shared interest among the group members.

Image Captioning



Prompt: Give a list of all objects present in this 19th century Japanese photograph, do not make sentences.

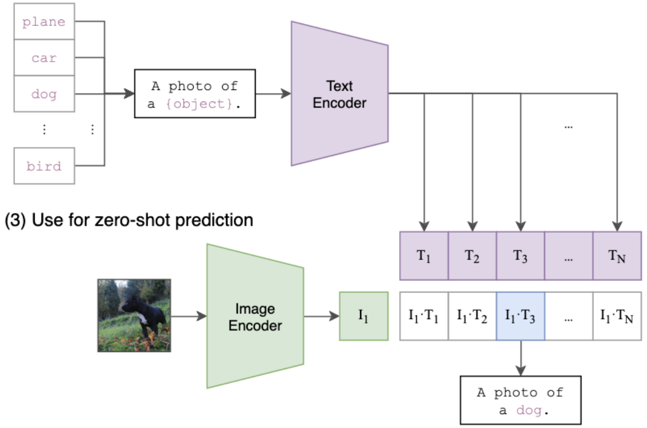
Answer:

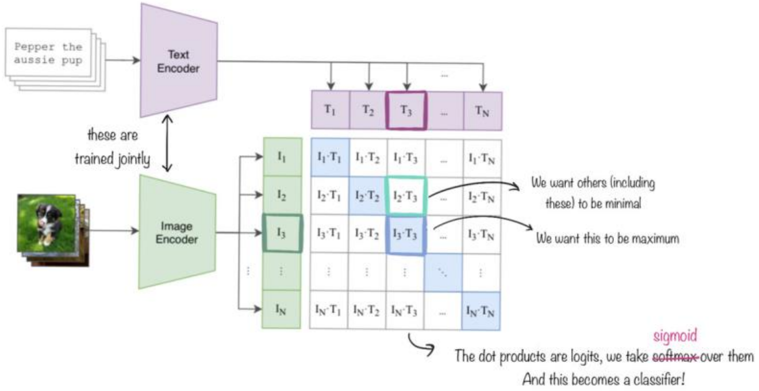
Map, umbrella, sword, hat, robe, bow, and a group of people.

The format of the description can be adapted for archival/indexing purposes

Image Captioning: quantitative evaluation

Models:





We want others (including these) to be minimal

We want this to be maximum

sigmoid

The dot products are logits, we take ~~softmax~~ over them
And this becomes a classifier!

ChatGPT-4o

CLIP
Open Source

SigLIP
Open Source

ChatGPT
API

Learning Transferable Visual Models From Natural Language Supervision, Alec Radford et al., 2021

Sigmoid Loss for Language Image Pre-Training, Xiaohua Zhai et al., 2023

Image Captioning: quantitative evaluation

Data:

- Manual tagging by a single person of 18,155 images
- 2,209 tags including
  - content: child, river, fan
  - elements of Japanese culture: geisha, shamisen, torii
  - types of photographs: group portrait, landscape
  - Japanese locations or monuments: Nikko, Yokohama, Toshogu Shrine

Image Captioning: quantitative evaluation

Data set for the experiments:

- Random selection of 6 albums, 275 images
- Tag set reduced to 1,034 considering only common names in French
- **Goal**: being able to manually validate the predictions of the models

Image Captioning: quantitative evaluation

Model configuration for the experiments:

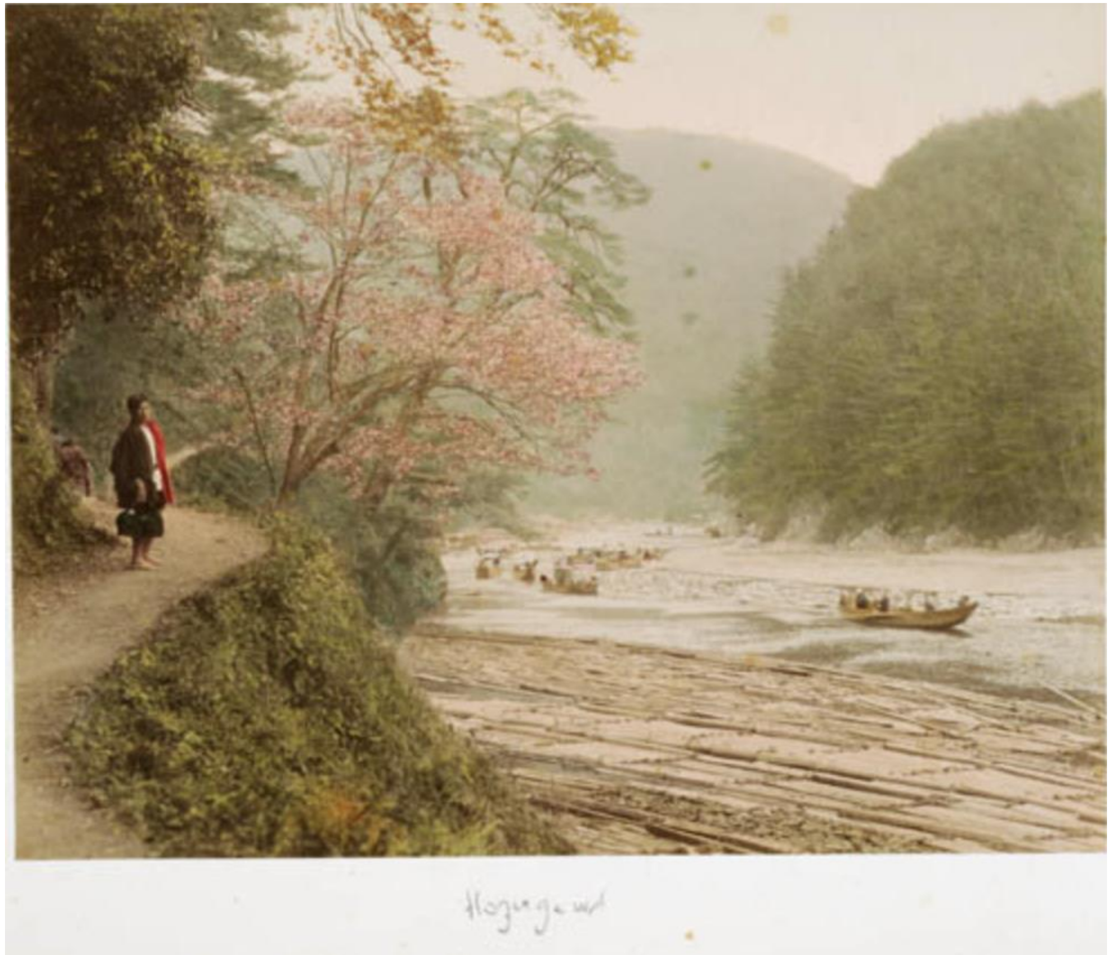
- Same prompt for the 3 models
- The prompt contains the list of possible tags
  - The reduced tag set: the 190 tags of the subset of images
  - The complete tag set: 1,034 tags

Image Captioning : quantitative evaluation

Model	Tag set	Precision	Recall	F1
CLIP	Reduced	15.7	11.5	13.3
	Full	12.9	6.6	8.7
SegLIP	Reduced	13.2	10.4	11.6
	Full	12.6	7.8	9.7
ChatGPT 4o	Reduced	74.8	42.5	54.2
	Full	25.6	20.6	22.9

- Open Source models are far behind ChatGPT
- ChatGPT has a low recall with both tag sets
- ChatGPT has a low precision with the full tag set

Image Captioning : qualitative evaluation



Manual tags	GPT4o tags
Path	
Boat	Boat
Natural landscape	Natural Landscape
Hozugawa	
Rapids	
People	People
River	River
Studio Tamamura	
Colored	
	Cherry tree

Image Captioning : quantitative evaluation

New reference : *Full validated* , manual validation of the predictions

For each prediction, if the tag is correct, the tag is added to the ground-truth

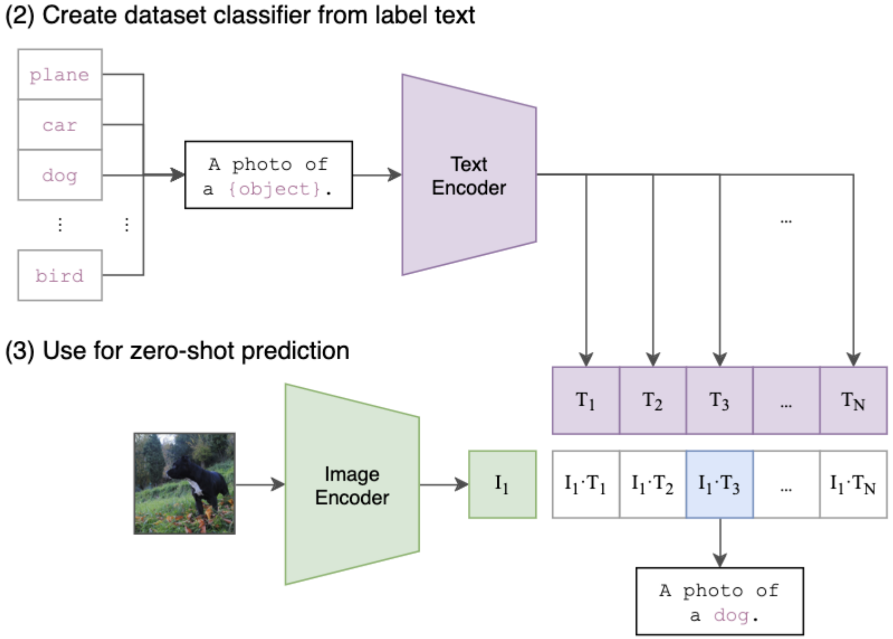
Model	**Tag set**	**Precision**	**Recall**	**F1**
ChatGPT 4o	Reduced	74.8	42.5	54.2
	Full	25.6	20.6	22.9
	Full validated	**80.2**	**43.5**	**56.4**

ChatGPT4o predicts fewer tags but is accurate in its predictions

Limitation of the captioning approach: only allows for (closed) keyword search

Image semantic search

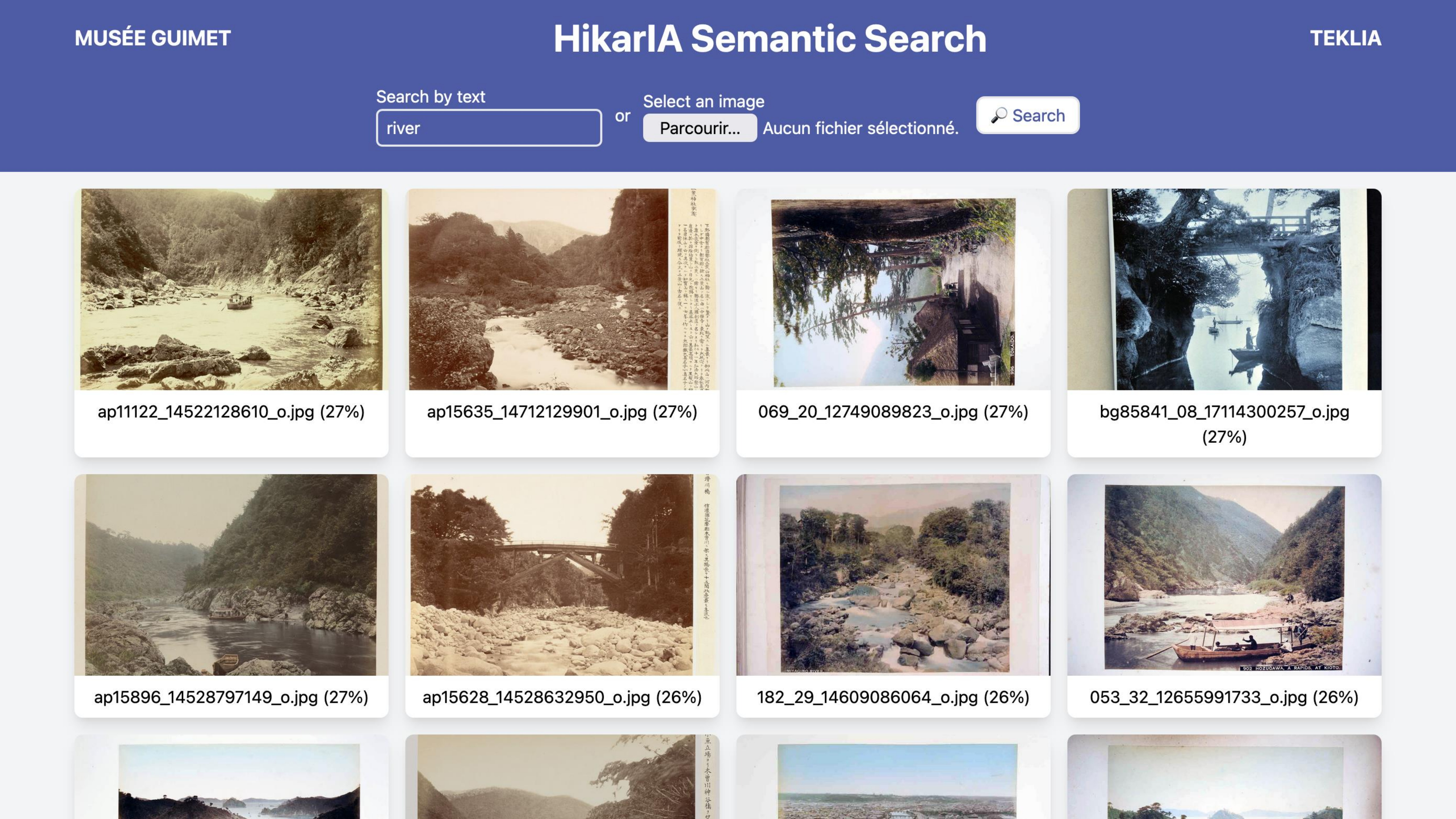
CLIP Model


(2) Create dataset classifier from label text
plane
car
dog
bird
A photo of a {object}.
Text Encoder
(3) Use for zero-shot prediction
Image Encoder
I_1
T_1 T_2 T_3 ... T_N
$I_1 \cdot T_1$ $I_1 \cdot T_2$ $I_1 \cdot T_3$ ... $I_1 \cdot T_N$
A photo of a dog.

Learning Transferable Visual Models From Natural Language Supervision, Alec Radford et al., 2021

Search application

- Encode the low resolution images with CLIP
- Store them in a vector database
- Encode the textual query with CLIP
- Search the closest vector in the database
- Display the best results



Search by text: river

ap11122_14522128610_o.jpg (27%)

ap15635_14712129901_o.jpg (27%)

069_20_12749089823_o.jpg (27%)

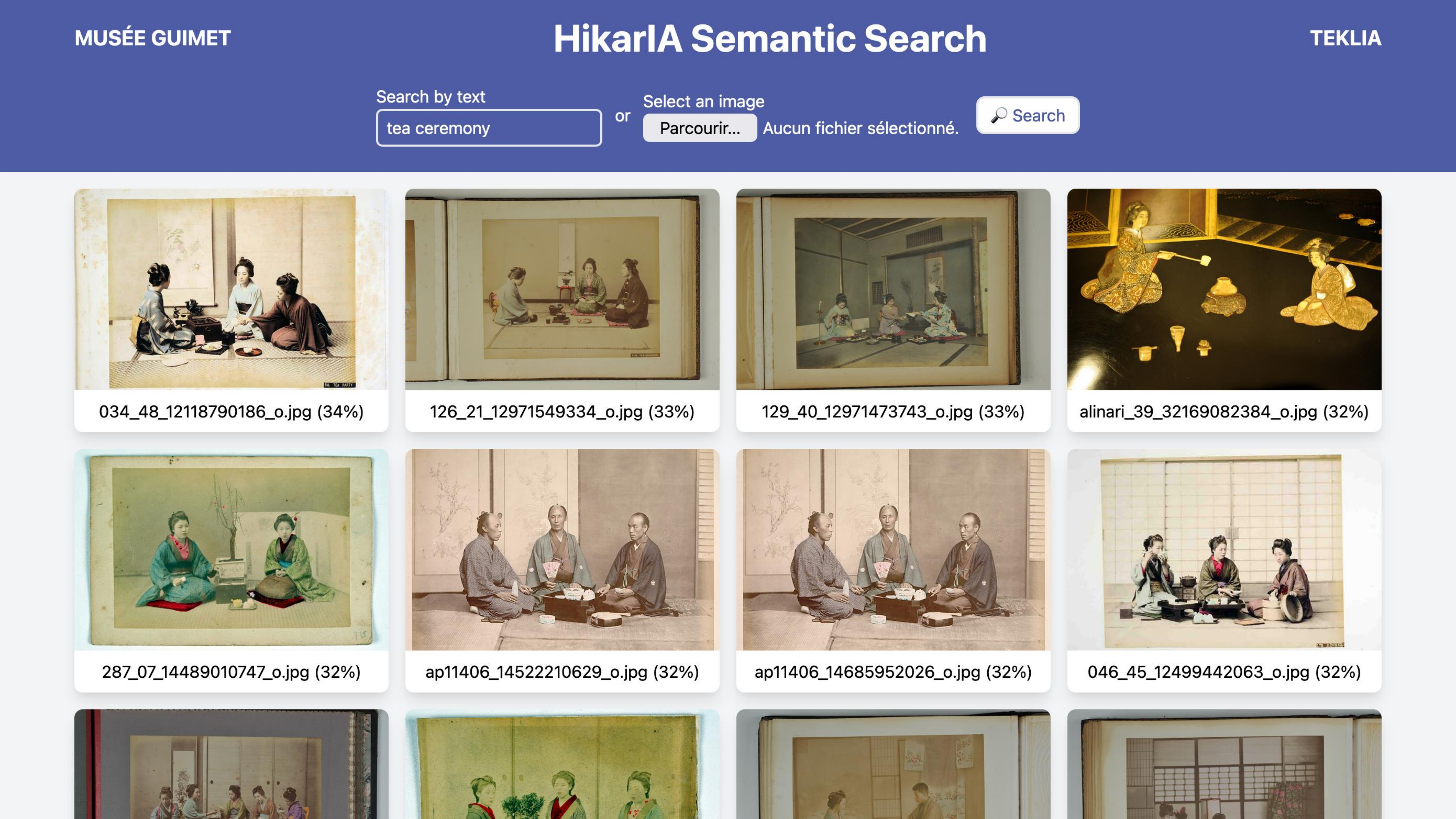
bg85841_08_17114300257_o.jpg (27%)

ap15896_14528797149_o.jpg (27%)

ap15628_14528632950_o.jpg (26%)

182_29_14609086064_o.jpg (26%)

053_32_12655991733_o.jpg (26%)



HikarIA Semantic Search

Search by text: tea ceremony

034_48_12118790186_o.jpg (34%)

126_21_12971549334_o.jpg (33%)

129_40_12971473743_o.jpg (33%)

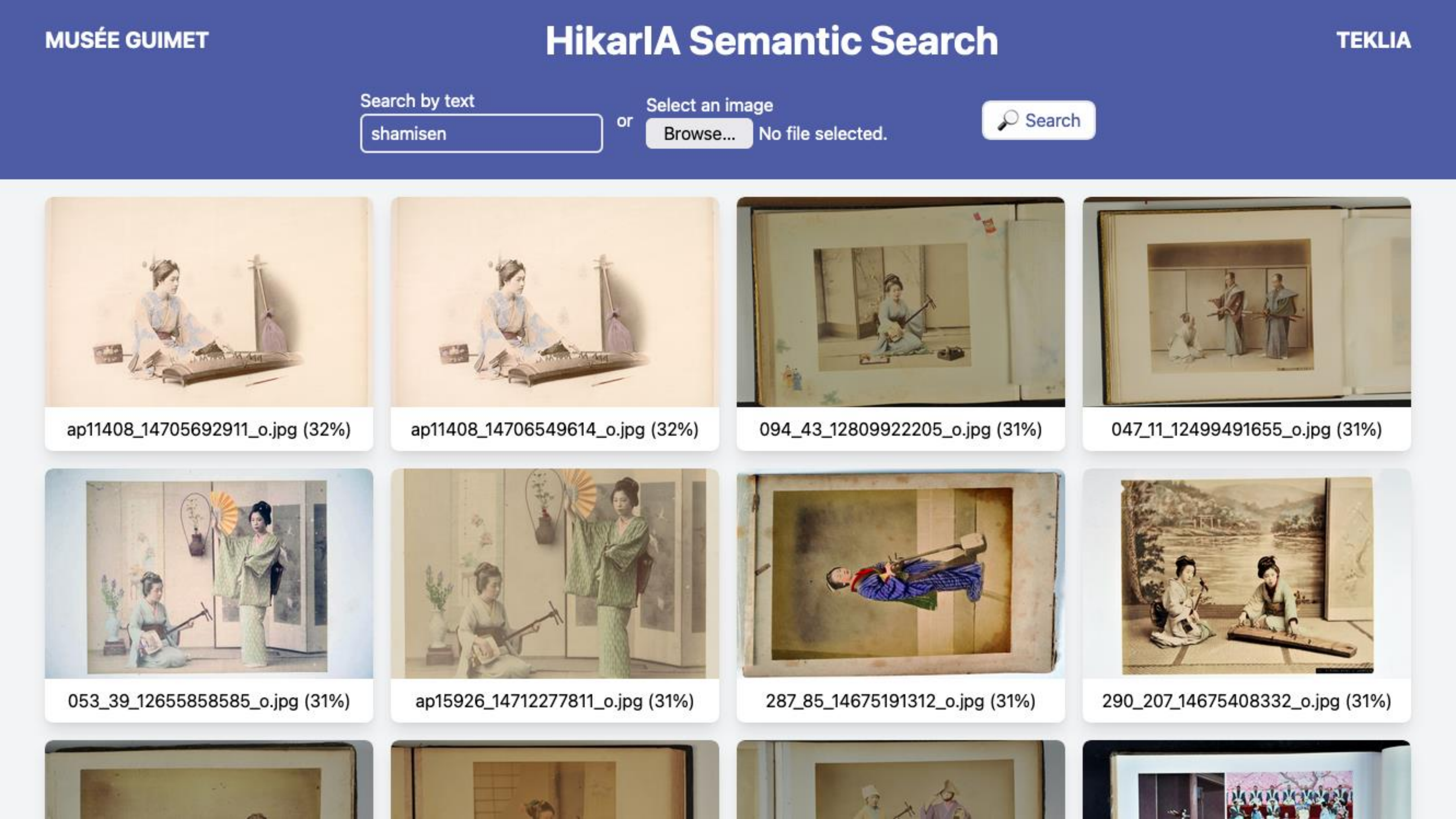
alinari_39_32169082384_o.jpg (32%)

287_07_14489010747_o.jpg (32%)

ap11406_14522210629_o.jpg (32%)

ap11406_14685952026_o.jpg (32%)

046_45_12499442063_o.jpg (32%)



Search by text: shamisen

Select an image: No file selected.

ap11408_14705692911_o.jpg (32%)

ap11408_14706549614_o.jpg (32%)

094_43_12809922205_o.jpg (31%)

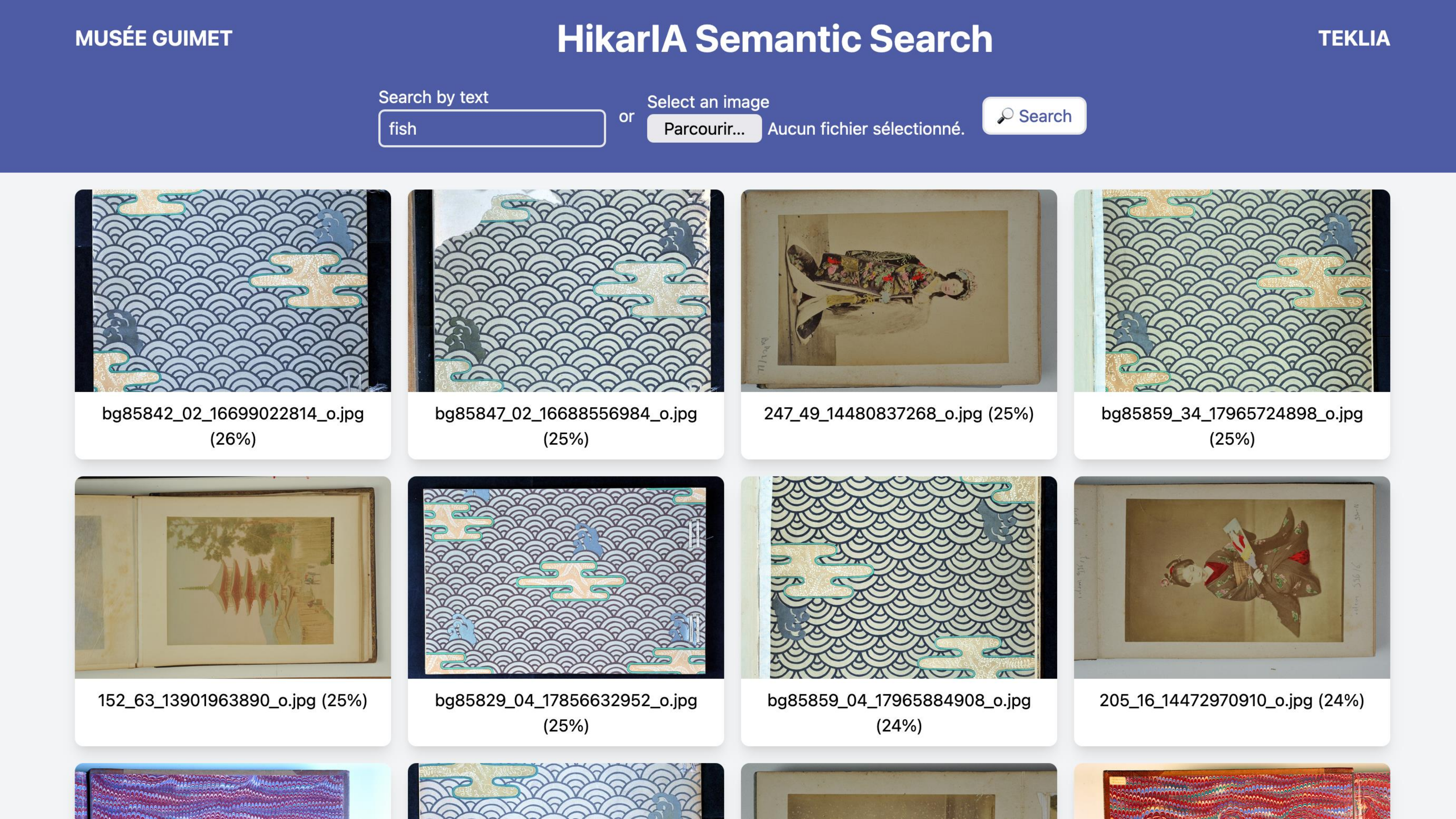
047_11_12499491655_o.jpg (31%)

053_39_12655858585_o.jpg (31%)

ap15926_14712277811_o.jpg (31%)

287_85_14675191312_o.jpg (31%)

290_207_14675408332_o.jpg (31%)



Search by text: fish

bg85842_02_16699022814_o.jpg (26%)

bg85847_02_16688556984_o.jpg (25%)

247_49_14480837268_o.jpg (25%)

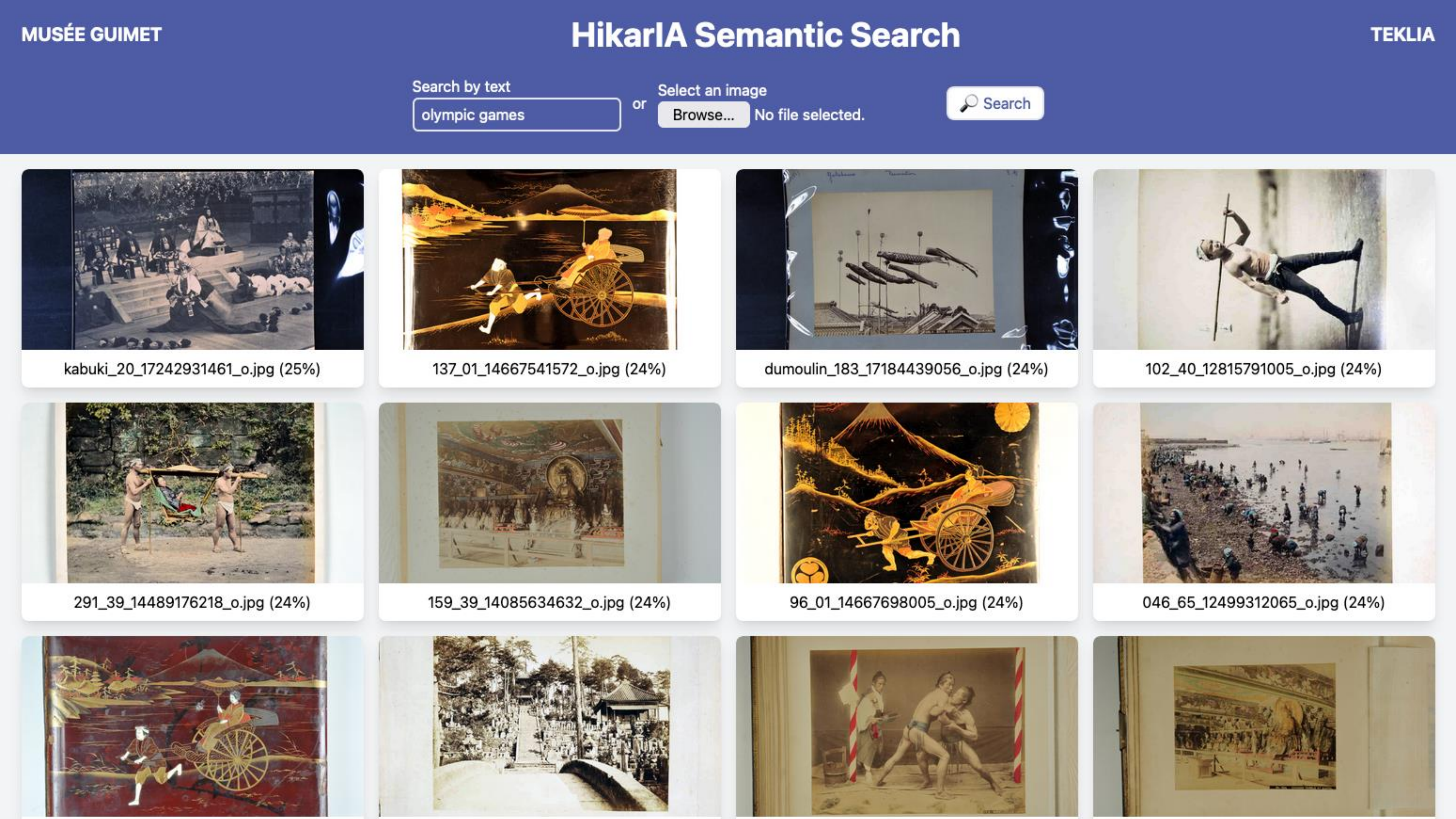
bg85859_34_17965724898_o.jpg (25%)

152_63_13901963890_o.jpg (25%)

bg85829_04_17856632952_o.jpg (25%)

bg85859_04_17965884908_o.jpg (24%)

205_16_14472970910_o.jpg (24%)



Search by text: olympic games

Select an image: No file selected.

kabuki_20_17242931461_o.jpg (25%)

137_01_14667541572_o.jpg (24%)

dumoulin_183_17184439056_o.jpg (24%)

102_40_12815791005_o.jpg (24%)

291_39_14489176218_o.jpg (24%)

159_39_14085634632_o.jpg (24%)

96_01_14667698005_o.jpg (24%)

046_65_12499312065_o.jpg (24%)

Conclusion

Models can analyse old photographs using contemporary, common vocabulary in English/French

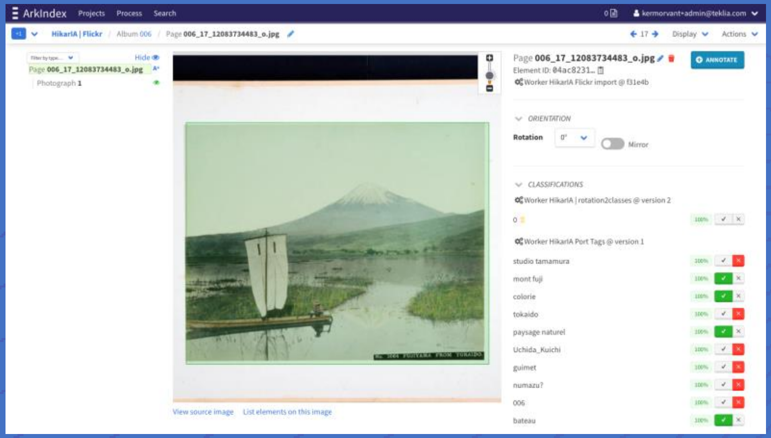
- Multi-lingual interactions should be evaluated
- The interaction with the thesaurus can be improved (hierarchical approach)
- Automatic evaluation is complex due to subjectivity

Moving from lab evaluation to field evaluation

TEKLIA’s open-source software suite for document processing

Arkindex

Document processing



Page 006_17_12083734483_o.jpg

Photograph 1

Page 006_17_12083734483_o.jpg

Element ID: 04ac8231...

Worker HikarIA Flickr import @ f31e4b

ORIENTATION

Rotation 0°

Mirror

CLASSIFICATIONS

Worker HikarIA | rotation2classes @ version 2

0	100%

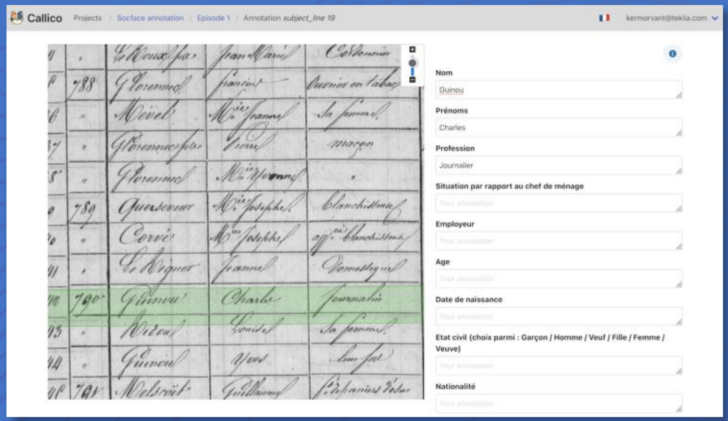
Worker HikarIA Port Tags @ version 1

studio tamamura	100%
mont fuji	100%
colorie	100%
tokaido	100%
paysage naturel	100%
Uchida_Kuichi	100%
guimet	100%
numazu?	100%
006	100%
bateau	100%

Open-source and Enterprise licences

Callico

Collaborative annotation campaigns



Nom: Guinou

Prénoms: Charles

Profession: Journalier

Situation par rapport au chef de ménage: Your annotation

Employeur: Your annotation

Age: Your annotation

Date de naissance: Your annotation

Etat civil (choix parmi : Garçon / Homme / Veuf / Fille / Femme / Veuve): Your annotation

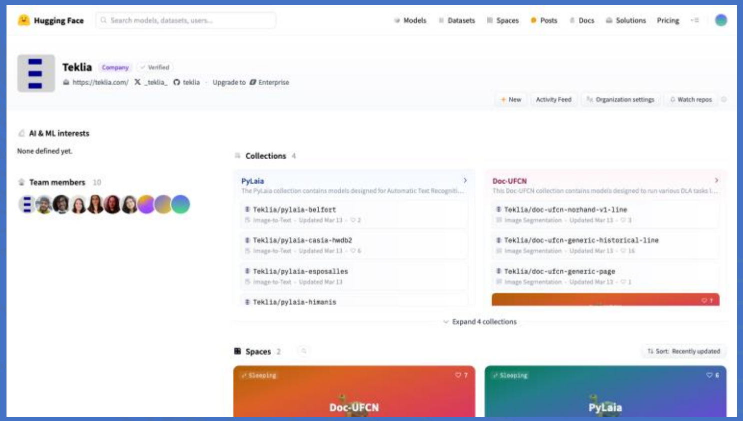
Nationalité: Your annotation

Open-source

https://gitlab.teklia.com

HuggingFace

Models and datasets



Teklia Company Verified

https://teklia.com/ _teklia_ teklia · Upgrade to Enterprise

AI & ML interests

None defined yet.

Team members 10

Collections 4

PyLaia

The PyLaia collection contains models designed for Automatic Text Recogniti...

- Teklia/pylaia-belfort — Image-to-Text · Updated Mar 13 · ♡ 2
- Teklia/pylaia-casia-hwdb2 — Image-to-Text · Updated Mar 13 · ♡ 6
- Teklia/pylaia-esposalles — Image-to-Text · Updated Mar 13
- Teklia/pylaia-himanis

Doc-UFCN

This Doc-UFCN collection contains models designed to run various DLA tasks i...

- Teklia/doc-ufcn-norhand-v1-line — Image Segmentation · Updated Mar 13 · ♡ 3
- Teklia/doc-ufcn-generic-historical-line — Image Segmentation · Updated Mar 13 · ♡ 16
- Teklia/doc-ufcn-generic-page — Image Segmentation · Updated Mar 13 · ♡ 1

Spaces 2

Open-source

https://huggingface.co/Teklia