An Image Similarity Search for the European Digital Library and Beyond

Sergiu Gordea

AIT-Austrian Institute of Technology GmbH,
sergiu.gordea@ait.ac.at

Abstract. This paper presents an Image Similarity Search service for the European Digital Library and summarizes the requirements of different stakeholders that make use of this kind of service. The current implementation is suited to support public users when exploring the content of Europeana. Further enhancements of this service are required to support design related activities by stimulating and inspiring the creativity of professional designers.

Keywords: image similarity, European digital library, creative design

1 Introduction

The European Digital Library (Europeana[1]) provides a single point of access to ca. 22 million cultural heritage objects provided by European Galleries, Libraries, Archives and Museums. This is valuable for the education, research, tourism or creative industries domains, but the heterogeneity of the content objects and poor textual descriptions raise difficulties when navigating and exploring this large repository. This arises mainly because of multi-lingual object descriptions (i.e. using one of 27 European languages), different types of content (i.e. text, image, sound, 3D) and lack of standardized classifications among the 2000+ content providers. Additionally, some of the collections provide poor descriptions of their objects, especially in the case of image content (see objects in photography collections[2]). Within this context, content based retrieval services are providing complementary solutions to overcome the limits of text based search. By using an image similarity search service, the user has the possibility to select or provide a picture and find objects with similar visual content (available within the image index). For example one could search for buildings that are similar to St. Mary monastery in Goranxi, Albania (built in the 16th century). In this case churches or old buildings from different countries having their description in different languages are retrieved (see Figure 1).

Within this paper we present a service supporting exploration of the Europeana repository using a content-based image retrieval approach. The service

[1] see http://europeana.eu

[2] http://www.europeana.eu/portal/search.html?query=title:Collectie+Heidemaatschappij

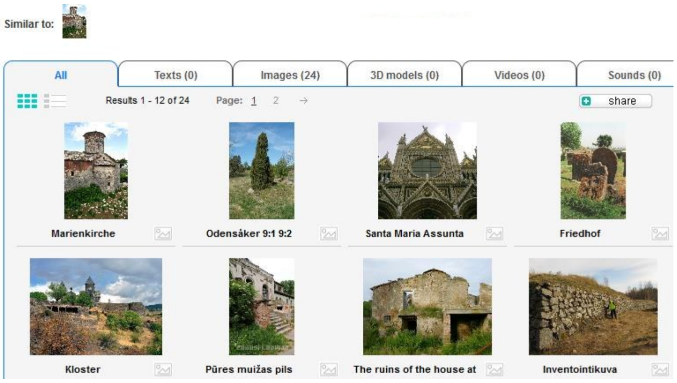

Fig. 1. Searching buildings similar to Goranxi monastery

supports well the search strategy of the public user, while on-going work concentrates on extending it to satisfy the expectations of creative design professionals. A new dataset with objects suited to inspire design activities was created and made available for evaluation and demonstration purposes.

2 Image similarity search

There are three main categories of stakeholders interested to use content-based search functionality when exploring the Europeana repository: public users exploring cultural heritage assets, professionals in different application domains (e.g. design, history education, art, etc.) and librarians. Typically, they have different levels of knowledge with respect to the Europeana content, different search intentions and different expectations when using an image retrieval service [8, 4]:

Public users typically use image search functionality to explore the content of the repositories, eventually searching for objects similar to the ones they know. They expect to retrieve results that present a certain degree of similarity to the query object. This less constraining degree of similarity can be achieved by using global image descriptors like color and edge histograms. Real-time system response is very important in this context [3].

Professional users search for image content able to inspire their work. The employees of creative industries sector are representative stakeholders. They are either interested in the semantic of the image content (i.e. designers, journalists) or in particular geometries and proportions (i.e. architecture designers) [8]. Scalable solutions were proposed recently supporting this kind of search strategies in [6, 2]. A trade off between good time performance in favor of search quality is accepted within this group of users.

Librarians have the highest level of knowledge regarding the content available in the GLAM collections, consequently they have the highest expectations from image retrieval services. They are interested to perform precise categorizations of the images (e.g. baroque buildings) [7], find multiple copies of the same object (e.g. compare various copies of Gioconda) or recently to find duplicate digitized content in large repositories [5]. The search service demonstrated within this

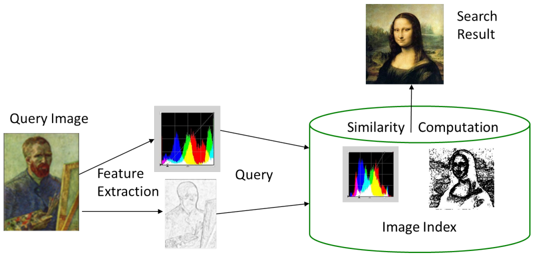

Fig. 2. Image similarity search process

paper was developed in the context of ASSETS[3] project to improve the user experience when exploring Europeana. The implementation is based on open source technologies. The MPEG-7 descriptors are computed by using the LIRE[4] feature extractors and the approximate similarity search algorithm [3] uses the Melampo[5] library. Figure 2 sketches the principle used within the similarity search process. The core of the retrieval process is represented by the computation of the similarities between the query and result images. This is performed by employing the cosine similarity computed on color and shape features extracted from the original images, namely the MPEG-7 Sclable Color, Color Layout and Edge Histogram descriptors.

While navigating through the digital library, users are allowed to use a query image available in Europeana, in Internet or on his/her computer for accessing similar objects in the repository. An usability evaluation was performed in Assets indicating clear interest of public and professional users for image search functionality [1]. The evaluation methodology used a task-oriented and supervised approach. The users were separated in two groups (i.e. public and professional) and interviewed to provide their feedback after interacting with the Assets portal. The conclusions on the image similarity indicated a strong interest and acceptance of the service, but also discovered that the semantics of the similarity is perceived differently by the two groups. In particular, the professional users were not satisfied by the quality of results when using their own query objects that are not available in Europeana. Currently, development efforts are invested to integrate this service in the Europeana search API. A demonstration of the baseline algorithm and the description of the dataset used are available online[6].

2.1 Content reuse in Creative Design

Even if most cultural heritage objects have a certain potential to inspire creative design activities, there are several categories of content that present particular interest. Fashion and web designers present use attractive color combinations and particular textures in their daily work. These can be found in nature (e.g. coloring of birds or butterflies), in older ornamental objects or clothing (e.g.

[3] http://ec.europa.eu/information_society/apps/projects/factsheet/index.cfm?project_ref=250527

[4] http://www.semanticmetadata.net/

[5] https://github.com/claudiogennaro/Melampo

[6] see http://62.218.164.177:8080/imagedemo/

court dressing in 19th century). The film and advertisement sectors often use costumes and images of historical characters available in portraits of famous historical characters. Game designers are inspired by mythology and animals, while architects intensively study the symmetry and shapes of older buildings.

3 Experimental Evaluation

By using the Europeana search API we aggregated a dataset of 5125 cultural heritage objects, classified in 14 categories, that present high value for creative design professionals (see Table 1). We set up an experiment having the goal to verify the following two hypotheses:

Hypothesis 1. The current search algorithm presents a level of retrieval accuracy that is able to satisfy the expectations of the public users exploring the Europeana repository.

Hypothesis 2. The image retrieval service may effectively support creative designers in their work by satisfying their expectations in particular search contexts.

In order to verify the given hypotheses, we manually classified the images in the dataset by using two level of categorizations. In order to formalize the different retireval expectations, we consider that the public users are satisfied by retrieving objects from the same top level category (i.e. Top-Category), while the professionals expect results from the more precise categorization (i.e. Sub-Category). We use the *Precision at N* ($P@N$)[7] metric to measure the accuracy of the retrieval algorithm and we vary N from 5 to 25.

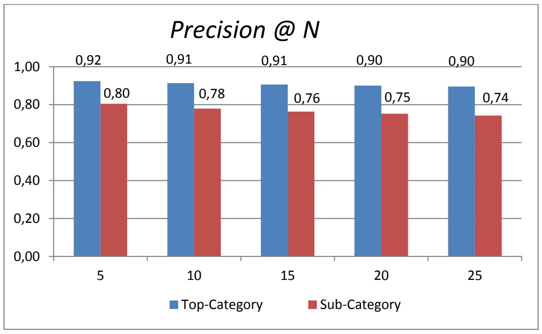

Fig. 3. Average *Precision at N*

Figure 3 presents the average $P@N$ both, at Top-Categoy and Sub-Category levels (i.e. a search result is considered a hit if it belongs to the same Top-Category or same Sub-Category as the query image). One can identify a tendency of decreasing the retrieval accuracy with the increase of results list. Additionally, the difference between the precision within the *Top-Category* and in the *Sub-Category* increases with N. On average, we can conclude that the search algorithm performs well on the given dataset.

[7] http://www.springerreference.com/docs/html/chapterdbid/63595.html

Table 1 presents the accuracy of the search algorithm tailored by categories and computed within *Top5*, *Top15* and *Top25* results, respectively. In column headings *P5tc* stands for *P@5* using *Top-Category* matching, *P5sc* uses *Sub-Category* matching, and so on.

Table 1. Image retrieval accuracy using Top- and Sub-Category matching

Top-Category	Sub-Category	#	P5tc	P15tc	P25tc	P5sc	P15sc	P25sc
birds	ducks	121	0.77	0.65	0.60	*0.54*	*0.37*	*0.31*
birds	eagles & hawks	145	0.93	0.91	0.90	0.72	0.68	0.65
birds	parrots	105	0.92	0.86	0.85	0.62	0.39	0.34
birds	woodpeckers	210	0.88	0.83	0.82	0.51	0.46	0.43
drawings	landscapes	699	0.94	0.93	0.92	**0.96**	**0.95**	**0.94**
insects	butterflies & moths	371	*0.69*	*0.67*	*0.65*	0.69	0.67	0.65
objects	bottles	144	0.94	0.92	0.90	0.74	0.63	0.56
objects	decor miniatures	69	0.97	0.94	0.93	0.65	0.49	0.41
objects	electrical engineering	231	0.88	0.82	0.80	0.56	0.39	0.34
objects	musical trumpets	1092	0.98	0.97	0.97	**0.92**	**0.91**	**0.90**
objects	optical engineering	195	0.81	0.79	0.77	*0.38*	*0.29*	*0.25*
objects	porcelain	131	0.96	0.93	0.91	0.83	0.73	0.63
paintings	landscapes	425	0.88	0.85	0.83	0.69	0.64	0.61
paintings	portraits	1187	0.97	0.97	0.97	**0.90**	**0.90**	**0.89**

Within this experiment one can identify that the matching on Top-Categories provides good results. The lowest precision (i.e. marked with italic font) was identified for butterflies & mots category, which can be explained by the fact that the real nature background is more relevant for the algorithm that the small insect in foreground. By using the Sub-Category matching, the current algorithm (heavily color based) provides best results for drawings and paintings (i.e. marked with bold fonts). The musical trumpets collection contains very similar objects, therefore the high retrieval accuracy. The worst results were obtained for the optical engineering images. The low precision encountered on the ducks category, is obtained because of the similarities with the other categories of birds but also because the nature of images. Many of these contain more than three birds placed with different positions and orientations.

Discussion. With the current experimental evaluation we validate both research hypotheses. The current search algorithm has a retrieval accuracy able to satisfy the expectations of public users navigating through the Europeana repository, and it may satisfy the expectations of professional designers searching for certain types of content like: paintings, drawings, glass or porcelain objects. For other categories of objects, shape-based descriptors and algorithms may perform better. The retrieval accuracy obtained on the current dataset is higher that the regular one, because of the nature of aggregated images. The content was contributed by museums that used a standardized method to generate the physical artifacts and to digitize them. Consequently, the similarity between objects within the same category (or collection) is higher than in other datasets. In any case, the objects were selected by using the Europeana Search API within collections with high quality content that can be reused for supporting creative design work.

4 Conclusions

Within this paper we presented the image similarity search service developed for Europeana using available open source technologies. On-going research activities are related to the reuse of cultural heritage content for inspiring creative design. Several categories of cultural heritage objects were identified to present particular interest in this context. The current version of the search algorithm provides a good overall retrieval accuracy on the given dataset, but there are certain types of images for which search performance doesn't meet the expectations of professional users. Future work will focus on identifying algorithms that are able to maximize the retrieval performance on each of the selected classes of images. We plan to evaluate additional image descriptors (e.g. local and shape-based descriptors) and search algorithms able to better satisfy the expectations of design communities.

Acknowledgments. This work was partially funded by the ASSETS4Europeana and EuropeanaCreative projects, co-funded by the Commission of the European Communities under the ICT Policy Support Programme (ICT PSP). This publication reflects only the author's/authors' views and the European Union is not liable for any use that might be made of information contained therein.

References

1. C. M. I. E. M. N. P. C. Agnes Saulnier, Preben Hansen. Final report on evaluation of assets services. project deliverable, ASSETS4Europeana, May 2012.
2. G. Amato, P. Bolettieri, F. Falchi, C. Gennaro, and F. Rabitti. Combining local and global visual feature similarity using a text search engine. In *CBMI*, pages 49–54, 2011.
3. G. Amato and P. Savino. Approximate similarity search in metric spaces using inverted files. In *Proceedings of the 3rd international conference on Scalable information systems*, InfoScale '08, pages 28:1–28:10, ICST, Brussels, Belgium, Belgium, 2008. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering).
4. T. Colombino, D. Martin, A. Grasso, and L. Marchesotti. A reformulation of the semantic gap problem in content-based image retrieval scenarios. In *Int. Conf. on the Design of Cooperative Systems. France-19-21 May*, 2010.
5. R. Huber-Mörk and A. Schindler. Quality assurance for document image collections in digital preservation. In *ACIVS*, pages 108–119, 2012.
6. E. Spyrou, Y. Kalantidis, and P. Mylonas. Exploiting a region-based visual vocabulary towards efficient concept retrieval. In *in Proceedings of Recognising and tracking events on the Web and in real life, in conjunction with SETN 2010 (EVENTS 2010)*, May 2010.
7. G. Tolias, Y. Kalantidis, and Y. Avrithis. Symcity: Feature selection by symmetry for large scale image retrieval. In *in Proceedings of ACM Multimedia (Full paper) (MM 2012)*, Nara, Japan, October 2012. ACM.
8. S. J. Westerman and S. Kaur. Supporting creative product/commercial design with computer-based image retrieval. In *Proceedings of the 14th European conference on Cognitive ergonomics: invent! explore!*, ECCE '07, pages 75–81, New York, NY, USA, 2007. ACM.