Searching Silk Fabrics by Images Leveraging on Knowledge Graph and Domain Expert Rules

Thomas Schleider
Raphael Troncy
Thibault Ehrhart
thomas.schleider@eurecom.fr
raphael.troncy@eurecom.fr
thibault.ehrhart@eurecom.fr
EURECOM
France

Mareike Dorozynski
Franz Rottensteiner
dorozynski@ipi.uni-hannover.de
rottensteiner@ipi.uni-hannover.de
Leibniz University Hannover
Germany

Jorge Sebastián Lozano
jorge.sebastian@uv.es
University of Valencia
Spain

Georgia Lo Cicero
georgia.locicero@unipa.it
University of Palermo
Italy

ABSTRACT

The production of European silk textile is an endangered intangible cultural heritage. Digital tools can nowadays be developed to help preserving it, or even to make it more accessible for the public and the fashion industry. In this paper, we propose an image-based retrieval tool that leverages on a knowledge graph describing the silk textile production as well as rules formulated by experts of this domain. Out of several possible similarity scenarios, two have proven to work best and have been integrated into an exploratory search engine.

CCS CONCEPTS

• **Computing methodologies → Visual content-based indexing and retrieval**.

KEYWORDS

knowledge graph, image retrieval, deep learning, cultural heritage

ACM Reference Format:
Thomas Schleider, Raphael Troncy, Thibault Ehrhart, Mareike Dorozynski, Franz Rottensteiner, Jorge Sebastián Lozano, and Georgia Lo Cicero. 2021. Searching Silk Fabrics by Images Leveraging on Knowledge Graph and Domain Expert Rules. In *Proceedings of the 3rd Workshop on Structuring and Understanding of Multimedia heritAge Contents (SUMAC '21), October 20, 2021, Virtual Event, China.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3475720.3484445

SUMAC '21, October 20, 2021, Virtual Event, China

ACM ISBN 978-1-4503-8668-5/21/10…$15.00
https://doi.org/10.1145/3475720.3484445

1 INTRODUCTION

The knowledge about the production of European silk across centuries is an intangible cultural heritage that can unfortunately be considered endangered. Many silk items that were originally produced in Europe do, however, still exist - mostly in museums and other collections around the world. Many of these museums give public access to the images and metadata of these objects through their websites and make it therefore possible to study them and their material or get inspired from their design. This data has originally not been made accessible from a single location or web address: even if digitized, the information about these objects and their images are scattered on the Web.

In this paper, we describe how we make use of domain-expert rules and a knowledge graph about European silk textiles to develop an image-based retrieval system to search and find related silk fabrics. We briefly describe the development of this knowledge graph and its underlying ontology, which has also been aided by experts in the domain of silk textiles, as it offered semantically annotated data about the depicted objects that could be used for the image retrieval component. The domain experts have helped to establish not only one of the definitions of image similarity based on formulated rules. They also designed an important part of an evaluation framework, which strongly shaped the assessment process towards good results for other actual experts, but more importantly the public and other users.

The aim of all developed tools is to make the history of the silk heritage accessible to everyone: From domain experts to enthusiasts and historians, it shall be possible to overcome physical distances and learn about, see images of and study textile artefacts held in many collections about European silk textiles. Using digital images and metadata makes it possible to make quick comparisons between different search results and to study even those fragile fabrics that are impossible to manipulate physically.

The remainder of this paper is structured as follows. In Section 2, we describe some related work. In Section 3, we detail our approach. We evaluate our method in Section 4 and we illustrate the integration of this component into an exploratory search engine

in Section 5. Finally, we conclude and outline some future work in Section 6.

2 RELATED WORK

Knowledge Graph. The fact that most Cultural Heritage (CH) organisations or portals manage their own metadata using proprietary and ad-hoc formats makes integration often challenging and time consuming. We hypothesize that using Semantic Web technology, and in particular building a Knowledge Graph (KG) based on an ontology that follows the CIDOC Conceptual Reference Model (CRM) helps not only with making such integration tasks easier, but provides also tools for further semantic annotation. The CIDOC-CRM model was specifically developed for information in the field of CH and is the outcome of more than 20 years of development by ICOM's International Committee for Documentation (CIDOC). CIDOC-CRM [9] is an official ISO standard since 2006, which has been renewed in 2014 and it can be found at ISO 21127:2014.

The experiments described in this paper used training data that are extracted from a Knowledge Graph that relies on classes and properties defined by CIDOC-CRM and its direct extensions CRMsci (Scientific Observation Model) and CRMdig (Model for provenance metadata). Many other similar CH projects and models rely on or extend CIDOC-CRM, such as SCULPTEUR [1], which aims at improving the integration, browsing and retrieval of museum metadata or the more recent example of WarSampo [20], a Knowledge Graph about Finland in the Second World War. ArchOnto [19] is another example, which tries to specifically address challenges of integrating CH data from and for national archives.

Image Retrieval. In general, information retrieval aims to provide useful information to a user [28]. The particular case of *image retrieval* focuses on searching a database on the basis of images, referred to as *query images*, provided by a user. In this context, an abstract representation is calculated for all images in a database as well as for the query image, and the images in the database having the representation that is most similar to the representation of the query image with respect to a similarity measure are provided to the user.

Early works investigate content-based image retrieval (CBIR), in which case the image representations are based on visual features such as colour histograms [15, 17]. The drawback of CBIR is the so called *semantic gap* [33], which describes the inability to extract semantics from the images based on the low-level content features [31]. It is attempted to be overcome this challenge by means of semantic-based image retrieval (SBIR), allowing for retrieval results that are similar to the query image on a conceptual level, which could be, for example, the depiction of the same object type [31]. SBIR matches semantic information that is available for images, e.g. text features like the image filename that is supposed to be meaningful, the Alternate Text in a Web page or the title of the Web page [5, 31].

All of the methods cited so far use representations based on hand-crafted features, which means that the performance is restricted by the suitability of the selected features for the task. Most recent techniques for image retrieval do not use hand-crafted features, but learn the features by deep learning methods, in particular, by means of convolutional neural networks (CNN) [22]. In order to use CNN for image retrieval, pairs or triplets of images are passed through two or three network branches having identical structures and sharing their weights. Each branch delivers a feature vector (*descriptor*) for its input image and the goal of training the CNN's free parameters is to produce similar descriptors for similar input images and dissimilar descriptors for dissimilar input images. A prerequisite for training is a mathematical formulation of both, the similarity of image descriptors as well as the similarity of images, in the form of an objective function (*loss*) to be minimized during training. In this context, many works investigated different variants of defining similarity: [26] defines the similarity of two input images, one showing a photograph of an object and the other one showing a sketch of an object, in a binary way indicating whether the depicted objects are similar or dissimilar. The image descriptors in that work are computed by means of a Siamese CNN, being a two branch neural network with shared weights, where the descriptors are assumed to be similar (dissimilar) in case of a small (large) Euclidean distance. Training of the network weights is performed by minimizing the contrastive loss [6] forcing the network to learn weights that similar descriptors are indeed close to each other after training.

An alternative to a Siamese CNN architecture is using a triplet architecture for descriptor learning like in [29]. Such an architecture allows to present an image to a network with both a similar example as well as a dissimilar example, where the actual similarity of images in [29] is defined beforehand by a human operator. Minimizing a triplet ranking loss, [29] train the network weights such that the Euclidean distances between image descriptors serve as a measure of the similarity of the associated images. As manual labelling of pairs or triplets of images with respect to their similarity as needed by [26, 29] is extremely time consuming, [32] defines similarity of images based on class labels.

The authors of [32] propose to create binary hash codes out of the class labels for all images, where the similarity of the images is defined by the Hamming distance of the binary vectors. Their image retrieval network is trained to produce these binary vectors by means of minimizing the surrogate loss [18] such that images with similar class labels obtain binary descriptors with a small Hamming distance and vice versa.

Deep learning-based image retrieval has also been applied in the field of indexing collections containing cultural heritage artefacts, where the focus mainly is on querying painting databases. In this context, it is not uncommon to investigate datasets with multiple modalities, i.e. with both images and some textual annotations, such as [12, 24]. In [11], context-aware descriptors are learned for the image retrieval of art paintings from [12], where similarity of images is defined by means of the cosine similarity of context-aware embeddings. During training, their ContextNet jointly learns a classifier and it learns to produce descriptors. The descriptors are learned such that the L1 loss of the context embedding generated out of the available information in a Knowledge Graph using node2vec [14] and of the context embedding learned by the network becomes small.

Even though there are works addressing image retrieval for fabrics [4, 8, 30], only [7] focuses on fabric image retrieval in the context of cultural heritage. This work can be seen as an extension of [7] that defines similarity on the basis of information derived

from a Knowledge Graph. In contrast to [12], we do not derive a context embedding from a graph but from annotations describing silk fabrics such as the production timespan or the production place of the object. Like in [7], these annotations are interpreted as class labels used to define the similarity of images by means of semantic similarity, i.e. the degree of equivalent class labels while facing the problem of missing annotations. Our contribution compared to [7] is a modification of the semantic similarity that contains a more intuitive interpretation of the unknown silk properties. Further, the semantic similarity is extended by a rule-based similarity and an additional colour similarity loss and a self similarity loss are combined with the semantic losses to produce not only semantically similar descriptors but also visually similar descriptors. The different proposed losses are evaluated on a much larger dataset compared to [7] and an additional evaluation by cultural heritage experts gives an impression of how useful the developed methodology actually is in application.

3 APPROACH

3.1 Knowledge Graph

We have collected museum records describing mostly silk textiles and fabrics and other objects made out of silk from the last several centuries. They all come from museums and collections around the world, for which we developed a crawling and harvesting software for their websites or APIs. For the full pipeline of data preprocessing and knowledge graph integration, we follow a traditional extraction, transform and load (ETL) procedure. In all cases, we start with choosing relevant museum and collection websites, downloading their data as described above, converting all this metadata into a common intermediate JSON format, which is, however, not semantically annotated at this point. All original fields and their labels are then represented in the form of arrays. In addition to that, all images of those museum records are downloaded as well.

Next, we convert these JSON files to create a knowledge graph using the Resource Description Framework (RDF) as data model. We mostly instantiate a subset of the CIDOC-CRM ontology using domain-experts defined mappings between each original museum field and the most appropriate properties and classes from the CIDOC-CRM ontology. To give a simple example: most museums have a field for describing the production time of a (silk) object, but it is often called differently. Furthermore, the museums come from all over the world and we have to deal with different languages for both the field names and the values. Therefore a mapping is created for, e.g, a field named "Date" (in the case of the Metropolitan Museum of Arts) and the class `E12_Production` with the property `P4_has_time-span` and another class `E52_Time-Span`. A similar mapping rule will be written for the field named "date_text" (in the case of the Victoria and Albert Museum) and for the field named "Datación" (in the case of the Red Digital de Colecciones de Museos de España). The values for those fields will also be harmonized and the knowledge graph makes use of the Getty AAT thesaurus[1] for naming the time periods.

This conversion process involves at the same time a semantic enrichment process which is mostly based on a domain-expert designed thesaurus for silk textiles. If a string value for a material, technique or motif depiction can be matched with one of several labels in this thesaurus full of textile related concepts, the value gets replaced by a unique identifier of this concept, which is a fixed and language independent number after a base URI. The matching itself is string based and triggered when the field value is simple as opposed to complete paragraphs of texts. Not all of these concepts are necessarily linked, however, as they have been defined before actual data from the museums has been downloaded. Finally, we use additional controlled vocabularies such as Geonames[2] to link and normalise location values and use a more elaborate parsing mechanism to properly detect the production dates and time spans.

The resulting knowledge graph is finally loaded in a triple store while the images are separately uploaded onto a media server. The knowledge graph contains descriptions of 36210 unique objects illustrated by 74527 images (Figure 1).

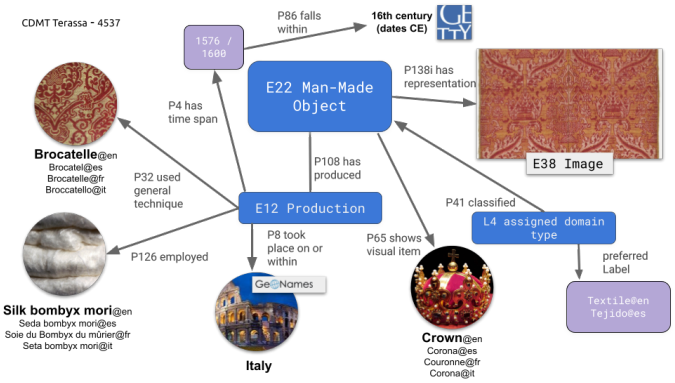

Figure 1: Excerpt of the knowledge graph: a textile object coming from the CDMT Terasse museum which has been produced in Italy in the 16th century, with the Brocatelle technique, using silk bombyx mori as material and showing the motif of a crown.

3.2 Domain Experts Rules for Image Similarity

Training of the method for image-based retrieval requires pairs of images (x_n, x_o) for which it is known whether they are similar or not, as it will be described in Section 3.3. One way of obtaining these data is to provide rules defining sets of images that should be similar and sets of images that should be dissimilar based on the content of the knowledge graph.

Such rules have been formulated by cultural heritage experts on the basis of an analysis of early image retrieval results. The rules are used to formulate SPARQL queries to the European silk textile knowledge graph. The results of these queries are transformed into a set T_{CE} of image pairs (x_n, x_o) with known similarity status. This state is binary, i.e. a pair can be similar according to the rules formulated by the experts or not. An overview of the rules that are used is given in Table 1.

These rules correspond to different aspects of similarity. Rule 1 corresponds to self-similarity but is based on real images; for the

[1] https://www.getty.edu/research/tools/vocabularies/aat/

[2] https://www.geonames.org/

Nr.	Rule: Two images are supposed to be similar if ...
1	they belong to the same record in the knowledge graph, i.e. if they show the same fabric
2	they both correspond to objects in the Garín dataset and the production material is “graph paper” or the technique is “gouache sobre papel” or “gouache sobre papel milimetrado”
3	for both corresponding fabrics, the information about production material or production technique is “pile-on-pile velvet”
4	for both corresponding fabrics, the information about production material or production technique is “ciselé velvet”
5	for both corresponding fabrics, the information about production material or production technique is "ciselé velvet" and the sub-depiction is “pomegranate”
6	anywhere in the corresponding records, “plain fabric” is mentioned
7	for both corresponding fabrics, the colour feature vectors belong to the same colour cluster among clusters identified to be relevant for defining similarity because the corresponding objects were found to be similar according to the cultural heritage experts. These clusters are clusters 9 and 11 (saturated red), cluster 5 (blue), cluster 22 (blue damasks) and cluster 27 (green damasks), see also Section 3.3.2.

Table 1: Rules defined by the cultural heritage experts to define pairs of similar images. Nr: number of the rules.

image pairs affected by it, the loss in eq. 11 is equivalent to the one in eq. 8 (see Section 3.3.2). Rules 2-6 consider semantic properties of silk fabrics and can be seen as variants of semantic similarity. However, they only consider one or two semantic properties and disregard all of the others, and a binary concept of similarity is used. Finally, rule 7 considers the colour distribution and, thus, an aspect of visual similarity. This rule has been designed based on the results of a cluster analysis of colour feature vectors as follows. First, a large set of images has been exported from the knowledge graph. From these images, colour feature vectors $h(x_n)$ were computed in the way described below (Section 3.3.2) in the context of the colour similarity loss. After that, k-means clustering using $k = 30$ clusters was carried out. Some clusters, identified by their cluster indices in Tab. 1, were found to contain images which should be considered similar.

Originally, the domain experts have also defined one dissimilarity rule (i.e. a negative rule example). However, this rule did not produce a sufficient number of examples to be useful and only positive rules (i.e. pairs of images considered to be similar) have been considered. Consequently, as there are no dissimilar pairs, the loss in eq. 11 can only be used in combination with other loss functions in training. However, the principle could be expanded by additional rules to produce dissimilar pairs in the future.

3.3 Image-Based Retrieval

The goal of image retrieval is to use images as input to search for records in the knowledge graph. The core of the method is a convolutional neural network (CNN) [21] that converts images into feature vectors (*descriptors*) so that the descriptors of similar image pairs have a small Euclidean distance and descriptors of dissimilar pairs have a large one. Using the CNN to compute descriptors for all images in the knowledge graph and using a k-d tree as a spatial index, image retrieval itself becomes a k nearest neighbour (*knn*) search in the k-d tree [2]. The prerequisite of our method is a knowledge graph with records containing both images and annotations in one or more semantic variables. In this work, the knowledge graph containing images of silk fabrics with annotations in the five variables *production timespan*, *production place*, *production material*, *production technique* and *subject depicted* is used for that purpose. This section describes the CNN used to compute descriptors, focusing on the training procedure, which leverages the contents of the knowledge graph to generate training samples automatically without any human intervention.

3.3.1 CNN architecture. The CNN architecture used for image retrieval is based on [7]. Using an RGB image x scaled to 224 x 224 pixels as an input, a ResNet-152 [16] backbone is applied to generate a 2048-dimensional feature vector. This is followed by two fully connected (FC) layers with ReLU (Rectified Linear Unit) activations [25] of 1028 and 128 dimensions, respectively. The output of the last layer is normalized to unit length, resulting in the 128-dimensional feature vector $f(x)$ which represents the input image.

3.3.2 Training. For the ResNet-152 backbone, the pre-trained parameters from [16] are used and they remain constant during training. Thus, the only parameters that are determined in training are those of the two FC layers of the network. Training is based on standard stochastic minibatch gradient descent (SGD) using backpropagation for computing gradients. The training procedure requires pairs of images (x_n, x_o) for which it is known whether they are similar or not. It is based on the assumption that descriptors for similar image pairs should have a small Euclidean distance, whereas for dissimilar images this distance should be large [3, 13]. The loss function $E(\mathbf{x}, w)$ minimized in training to determine the parameters w of the network using the data $\mathbf{x}$ available for training, which can be derived automatically from the contents of the knowledge graph, is the weighted sum of four loss terms:

$$E(\mathbf{x}, w) = \alpha_t \cdot E_t(\mathbf{x}, w) + \alpha_s \cdot E_s(\mathbf{x}, w) + \alpha_c \cdot E_c(\mathbf{x}, w) + \alpha_r \cdot E_r(\mathbf{x}, w). \quad (1)$$

The four loss terms (E_t, E_s, E_c, E_r) in eq. 1 correspond to different definitions of similarity and are explained in the subsequent paragraphs. The weights $(\alpha_t, \alpha_s, \alpha_c, \alpha_r)$, which have to sum to 1, can be modified to define different similarity scenarios. Compared to [7], the innovation of our method lies in an improved formulation of the semantic similarity loss E_t and the integration of the other three loss terms.

E_t: Semantic Similarity. This loss term considers two images (x_n, x_o) to be similar if the semantic information associated with them is similar. Thus, "similarity" becomes a gradual concept: the more annotations are shared, the more similar a pair of images is considered to be. This definition of *semantic similarity* Y_s of a pair

of images (x_n, x_o) has to consider the fact that a sample might have annotations for a subset of the considered variables only:

$$Y_s\left(x_n, x_o\right) = \frac{1}{M} \cdot \sum_{m=1}^{M} v_m \cdot d_m\left(x_n, x_o\right) \cdot \pi_m^n \cdot \pi_m^o. \tag{2}$$

In eq. 2, m is the index of a semantic variable and M denotes the number of variables considered. The binary variables π_m^iindicate whether for the image $i \in \{n, o\}$ the annotation for variable m is available ($\pi_m^i = 1$) or not ($\pi_m^i = 0$). Thus, the term

$$u\left(x_n, x_o\right) = 1 - \frac{1}{M} \cdot \sum_{m=1}^{M} \pi_m^n \cdot \pi_m^o, \tag{3}$$

the percentage of variables for which there is no annotation in at least one of the images (x_n, x_o), expresses the level of uncertainty of the similarity. The weight v_m of a variable m can be used to give more or less importance to certain variables. In accordance with cultural heritage experts, these weights were set to 0.30, 0.25, 0.20, 0.15, 0.10 for the variables *subject depicted*, *production material*, *production place*, *production technique* and *production timespan*, respectively, i.e. the depicted subject was considered to be most relevant. Finally, the function $d_m\left(x_n, x_o\right)$ computes the level of agreement between the annotations of (x_n, x_o) for variable m:

$$d_m\left(x_n, x_o\right) = \frac{1}{max\left(K_n, K_o, \varepsilon\right)} \cdot \sum_{k=1}^{K} \delta(l_{mk}(x_n) = l_{mk}(x_o)), \tag{4}$$

where $l_{mk}(x_i)$ is an indicator variable with $l_{mk} = 1$ if for variable m, the class label k applies to image x_i with $i \in \{n, o\}$, K is the number of class labels for variable m, and $\delta(\cdot)$ denotes the Kronecker delta which returns 1 if the argument is true and 0 otherwise. K_i, $i \in \{n, o\}$, is the sum of all values $l_{mk}(x_i)$ for the image x_i and ε is a small constant to avoid division by zero. For most semantic variables m, $K_i = 1$, i.e. the class labels are mutually exclusive. However, for some classes, multiple class labels are permitted for a sample, e.g. a sample may consist of multiple *production materials*.

The loss term E_t is based on the the *triplet loss* of [7, 27]:

$$E_t(\mathbf{x}, w) = \frac{1}{N_t} \cdot \sum_{n_1=1}^{N_t} max\left(M(x_a^{n_1}, x_{ps}^{n_1}, x_{ng}^{n_1}) + \Delta_{a,ps,w}^{n_1} - \Delta_{a,ng,w}^{n_1}, 0\right). \tag{5}$$

The sum in eq. 5 is taken over N_t triplets of images $x_a^{n_1}, x_{ps}^{n_1}, x_{ng}^{n_1}$, where n_1 is the index of a triplet and each triplet consists of an anchor sample $x_a^{n_1}$, a positive sample $x_{ps}^{n_1}$ (i.e., a sample considered to be similar to $x_a^{n_1}$), and a negative sample $x_{ng}^{n_1}$ (a sample considered to be dissimilar from $x_a^{n_1}$). The term

$$\Delta_{a,i,w}^{n_1} = ||f_w(x_i^{n_1}) - f_w(x_a^{n_1})||_2 \tag{6}$$

denotes the Euclidean distance of the feature vectors $f_w(x_i^{n_1})$ computed for the image x_i, $i \in \{ps, ng\}$, of triplet n_1 and the feature vector $f_w(x_a^{n_1})$ of the anchor pixel. $M\left(x_a^{n_1}, x_{ps}^{n_1}, x_{ng}^{n_1}\right)$ is a margin:

$$M\left(x_a^{n_1}, x_{ps}^{n_1}, x_{ng}^{n_1}\right) = Y_s(x_a^{n_1}, x_{ps}^{n_1}) - \left(Y_s(x_a^{n_1}, x_{ng}^{n_1}) + u(x_a^{n_1}, x_{ng}^{n_1})\right) > !0 \tag{7}$$

Minimizing this loss forces the learned descriptors of x_a and x_{ps} to be close together in feature space and the descriptor of x_{ng} to have a larger distance from x_a than x_{ps}. The restriction expressed by $M\left(x_a^{n_1}, x_{ps}^{n_1}, x_{ng}^{n_1}\right) > !0$ in eq. 7 is used to select triplets. For each minibatch consisting of a set S of images with annotations, all possible triplets are considered as potential training triplets. For each such triplet, the margin is computed, and all the N_t triplets fulfilling the constraint are used to compute the loss and, consequently, to update the parameters. The constraint implies that the similarity $Y_s(x_a, x_{ps})$ of the anchor and the positive sample has to to be larger than the sum of the similarity $Y_s(x_a, x_{ng})$ of the anchor and the negative sample and the potential positive similarity according to the unknown properties expressed by the uncertainty term $u(x_a, x_{ng}$.

E_s: Self-Similarity. This loss function considers a visual aspect of similarity: an image should be considered similar to a synthetically adapted version of itself. It should help the CNN to learn that images of the same fabric that were captured, e.g. from different perspectives should be considered to be very similar. For every image x_n in a minibatch consisting of N_{MB} images, a synthetic image x_n' is generated by applying random rotations by 90^o, a random horizontal and vertical flipping, and the cropping of a window containing a random percentage $b_{crop} \in [0.7, 1]$ of the pixels of x_n. Furthermore, a random zero mean Gaussian noise with a standard deviation $\sigma_G = 0.1$ is added to the grey values. The loss is forces the Euclidean distance between the feature vectors generated by the CNN for an image x_n and its synthetic partner x_n' to be close to zero:

$$E_s(\mathbf{x}, w) = \frac{1}{N_{MB}} \cdot \sum_{n=1}^{N_{MB}} ||f_w(x_n) - f_w(x_n')||_2. \tag{8}$$

E_c: Colour Similarity. This loss takes into account a visual aspect of similarity: two fabrics should be considered similar if the corresponding images have a similar colour distribution. To avoid dependencies on the intensity, the images to be compared are transformed into the *HSV* (hue H, saturation S, value V) colour space, with $H \in [0, 1]$ and $S \in [0, 1]$. In order to compensate for the periodic definition of H, which is usually interpreted as an angle, H and S are considered to be polar coordinates and used to determine Cartesian coordinates (x^c, y^c), both in the interval $[0, r]$:

$$\begin{aligned} x^c(H, S) &= \frac{r}{2} + \frac{r}{2} \cdot S \cdot cos(2 \cdot \pi \cdot H) \\ y^c(H, S) &= \frac{r}{2} + \frac{r}{2} \cdot S \cdot sin(2 \cdot \pi \cdot H), \end{aligned} \tag{9}$$

where r defines the scale of the transformation. In this Cartesian coordinate system, a 2D grid of r x r cells and grid size 1 is defined. A 2D histogram is determined by assigning each transformed point to the grid cell in which it is situated and counting the number of points per grid cell. The histogram obtained for an input image x_n is converted into a *colour feature vector* $h(x_n)$ having r^2 components by stacking the columns of the 2D histogram on top of each other; it represents the colour distribution of x_n. Unless otherwise noted, we used r=5 in all experiments involving the colour loss, i.e. each colour vector had 25 elements. Using the N_{MB} images of a minibatch, $N_c = N_{MB} \cdot (N_{MB} - 1)/2$ pairs of images $(x_1^{n_2}, x_2^{n_2})$ can be generated, where n_2 is the index of a pair, and the colour feature vectors $h(x_1^{n_2})$ and $h(x_2^{n_2})$ can be computed. Using the symbol Δ^{n_2} to denote the Euclidean distances of the feature vectors $f_w(x_1^{n_2})$ and $f_w(x_2^{n_2})$ delivered by the CNN for the two images of pair n_2 and $\rho^{n_2} \in$

$[-1, 1]$ to denote the normalized cross correlation coefficient of the corresponding colour feature vectors $h(x_1^{n_2})$ and $h(x_2^{n_2})$, the colour similarity loss is formulated as:

$$E_c(\mathbf{x}, w) = \frac{1}{N_c} \cdot \sum_{n_2=1}^{N_c} max\left(0, |\Delta^{n_2} - \left(1 - \rho^{n_2}\right)|\right). \quad (10)$$

For image pairs having a similar colour distribution, i.e. a value of ρ^{n_2} close to 1, this loss will force the Euclidean distance to be close to 0, i.e. the feature vectors to be similar. The smaller the correlation coefficient, the more the Euclidean distance will be pushed away from 0; for $\rho^{n_2} = -1$, the distance will be pushed to 2, the maximum possible value because of the normalization (section 3.3.1).

E_r: Similarity Rules. The last loss function is based on the rules for defining sets of images that should be similar and sets of images that should be dissimilar described in section 3.2 (cf. Tab. 1. Assuming that a minibatch contains N_r such pairs $(x_1^{n_3}, x_2^{n_3}) \in T_{CE}$ (cf. section 2), where n_3 is an index of such a pair, and denoting the Euclidean distances of the feature vectors $f_w(x_1^{n_3})$ and $f_w(x_2^{n_3})$ by Δ^{n_3}, a standard loss to train the CNN to produce similar descriptors for similar images and dissimilar descriptors for dissimilar images can be applied:

$$E_r(\mathbf{x}, w) = \frac{1}{N_r} \cdot \sum_{n_3=1}^{N_r} \delta_s^{n_3} \cdot \Delta^{n_3} + (1 - \delta_s^{n_3}) \cdot max\left(2 - \Delta^{n_3}, 0\right). \quad (11)$$

In eq. 11, the variable $\delta_s^{n_3}$ indicates whether the pair $(x_1^{n_3}, x_2^{n_3}) \in T_{CE}$ is similar ($\delta_s^{n_3} = 1$) or not ($\delta_s^{n_3} = 0$). For pairs which are similar according to the rules defined in section 3.2, only the first term is active, and the loss will try to minimize the Euclidean distance of the descriptors of the two images. For dissimilar pairs, only the second term is active, and the loss will try to push the Euclidean distance close to the maximum possible distance of 2.

Minibatch generation. The training data $\mathbf{x}$ consist of images with annotations exported from the knowledge graph. In each training iteration, N_{MB} images are randomly selected from these data to form a minibatch (we used $N_{MB} = 150$ in training). If $\alpha_r \neq 0$, i.e. if the rule-based loss E_r (equ. 11) is used, 50% of the samples in the minibatch are drawn from the subset of images found to be affected by one of the rules described in Section 3.2 to ensure that the number N_r of pairs considered in E_r is sufficiently high. Note that the loss function terms are based on a comparison of different numbers of images. If the semantic similarity loss E_t (eq. 2) is used, all triplets fulfilling the constraint expressed by eq. 7 will be considered. If colour similarity E_c (eq. 10) is used, all possible pairs of images will be considered. For the self-similarity loss E_s (eq. 8), every image in the minibatch and a synthetically modified version of it will be considered. Finally, if the loss E_r (eq. 11) is to be used, all pairs of images in the minibatch affected by one of the rules will be considered. Note that the formulation of the total loss E_{total} (eq. 1) is flexible w.r.t. the combination of the loss terms. However, at least one of the two terms E_t and E_c has to be considered, because E_s and E_r do not contribute for dissimilar pairs, in the first case by design and in the second case because the rules in Tab. 1 do not define any dissimilar pairs.

One iteration of SGD starts by extracting a minibatch from the training data and defining the required sets of image pairs and triplets in the way just described. Afterwards, all images are propagated through the network, and the loss E_{total} is computed and back-propagated through the network to compute the gradient of the loss with respect to the unknown parameters w. Finally, these gradients are used to update the parameter values.

4 EVALUATION

The evaluation was carried out in three steps and all test data was exported from the knowledge graph. The test data consisted of 25,825 images with annotations in at least one of five semantic variables mentioned in Section 3.3 and an additional set of records affected by at least one of the rules described in Section 3.2. The first step involved a set of experiments for finding the optimal set of hyperparameters training and classification. These experiments were based on semantic similarity only. As we involved domain experts in the development of definitions of similarity, we also wanted to make sure we do not only evaluate the model with regards to a general similarity. Based on the different defined types of similarity, 5 different scenarios have been created together with the cultural heritage domain experts of our project in the second step:

- **Scenario A**: Semantic similarity and self-similarity.
- **Scenario B**: Colour similarity and self-similarity. Only scenario with exclusively visual similarities.
- **Scenario C**: Augmentation of semantic similarity with the rules defined by cultural heritage domain experts.
- **Scenario D**: Augmentation of colour similarity with the rules defined by cultural heritage domain experts.
- **Scenario E**: Combination of all concepts of similarity, which is meant to be a compromise between semantic and visual aspects of similarity.

As part of the second step, a purely technical evaluation has been performed based on five-fold cross validation and performing a k-nearest neighbour classification based on the optimal hyperparameters identified in step 1. This part of the evaluation focused on the ability to find images having similar semantic properties. The average accuracies and F1 scores of step 2 can be seen in Tables 2 and 3. As can be seen in these results, the overall F1 scores and accuracies are relatively similar, with Scenario E being altogether the best case.

The third step relied on these five scenarios, but the evaluation was performed by cultural heritage experts through an interactive analysis of the results. This type of expert evaluation is very time consuming, therefore only a limited amount of test data has been chosen and a fixed split into training and test data was used. 100 images were selected to be retrieved as test images, for which the k = 10 most similar images should be retrieved by the image retrieval tool. All remaining samples were used for training. Images of objects that contribute to the test set were excluded from training. This is especially important as one object can have several images.

The evaluation criteria used by the domain expert were based on the following concepts:

- *Pattern*: This concept is about decorative motives, for example flowers or birds. Therefore it is related to aspects of semantic similarity, as some records have explicit textual metadata descriptions about those.

- *Colour*: The perception of the colour of an image is relatively easy for most users. This term represents a visual type of similarity here.
- *Appearance*: The domain experts use this term for a concept of a generic evaluation of the outward form of the silk fabric in the image. This includes shape, the geometric form, but also colour again. The domain experts consider this to be a characteristic that can also be easily perceived by a typical user.

If a pair of images matches at least two criteria it was considered a meaningful pair, otherwise not. A graphical representation of the top-k-scores and the percentage of meaningful images for values k between 1 and 10 can be seen in Figure 2 and 3. For Step 3, we can see Scenario B performing by far the best, with Scenario E actually being mostly second best, but with a significant distance.

Based on these results two best scenarios have been chosen: Scenario E led to the highest F1 scores and overall accuracies based on semantic similarity, whereas Scenario B proved to be the best one according to the evaluations by the domain experts.

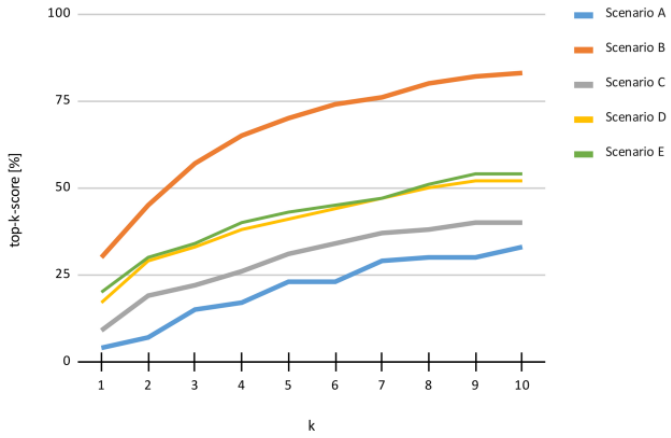

Figure 2: Top-k-scores as a function of k for all evaluated scenarios. The score gives the percentage of query images for which there was at least one meaningful result among the k most similar images delivered by the image retrieval module.

The investigated scenarios for similarity are based on different definitions of the loss function; they indicate that the consideration of the additional loss terms beyond those used in Scenario A do indeed contribute to a better performance if the focus of retrieval is on semantic aspects of similarity, whereas the new colour loss is essential for retrieving meaningful results according to the evaluation by domain experts. A full ablation study considering the contributions of all loss terms is beyond the scope of this study. Note that the method described in [7] is very similar to Scenario A; the difference is in the use of an improved version of the semantic loss and in the self-similarity loss.

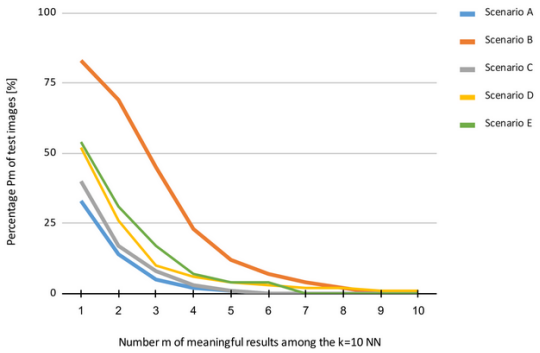

Figure 3: Percentage P_m [%] of query images for which the image retrieval module delivered at least m meaningful images among the k=10 nearest neighbours for Scenarios A-E.

5 AN EXPLORATORY SEARCH ENGINE FOR FINDING SIMILAR OBJECTS

The knowledge graph that we used to train the models is accessible via a RESTful API that has been developed using the grlc framework and the SPARQL Transformers library [23]. A web based application has been developed using this API to provide an exploratory search engine for silk textiles. It offers a user-friendly interface with facetted search to apply filters corresponding to the different properties of the silk textiles, like the material or technique being used or the production place and time [10].

We have integrated in this exploratory search engine the image-based retrieval module described in this paper. More precisely, we have integrated the two scenarios B and E described above under two buttons named "visually similar images" and "objects with similar properties". The user can upload any image of his choice (preferably depicting a silk textile) and invoke one of these two methods to retrieve up to 20 similar objects from the knowledge graph. Similarly, when browsing the knowledge graph, the user can request what are the similar objects (either visually or semantically) with respect to the one being viewed (Figure 4).

6 CONCLUSION

In this paper, we have presented an image retrieval module that considers different aspect of similarity between cultural heritage objects that silk textiles are. One of our contribution is to use a knowledge graph in order to convert domain-expert similarity rules into queries that generate vast amount of training data. The design and the evaluation of the image retrieval models benefit from the knowledge of domain experts. The code of the image retrieval method is available under Github[3].

While exploring different scenarios, we observe that the simplest visual only similarity provides the best accuracy: At least one meaningful image was retrieved per query image in 83% of all cases. The semantic similarity proves also to be useful for domain experts who appreciate to switch from one to the other and observe the

[3]https://github.com/silknow/image-retrieval

Variable	α_s	*Material*	*Production Place*	*Technique*	*Timespan*	*Depiction*	**Average**
Scenario A	1/2	**78.2** / 73.6	44.4	**61.8**	54.0	88.0	65.3 / 64.4
Scenario B	0	77.1 / 72.6	40.6	57.3	52.5	89.6	63.4 / 62.5
Scenario C	1/3	77.9 / 73.3	43.3	60.2	**54.4**	**90.1**	65.2 / 64.3
Scenario D	0	77.9 / 73.2	43.3	61.1	54.1	89.4	65.2 / 64.2
Scenario E	1/4	**78.2** / 73.4	44.1	61.6	53.8	89.1	65.4 / 64.4
SIR_LR_4	1	78.2 / 73.9	**44.6**	61.6	**55.3**	88.9	**65.7 / 64.9**

Table 2: Overall accuracies [%] per variable for the different scenarios of similarity as well as the best performing experiment of test step 1 (SIR_LR_4). The highest score per variable is highlighted in bold font. The second column contains the weight α_s of the loss function term related to semantic similarity and, thus, indicates whether semantic similarity is considered (α_s >0) or not (α_s = 0); the last column gives average values over all variables. In case of the variable Production Material, the first value refers to the classification results based on a binary classification procedure; the second value refers to the results including the most probable class of samples assigned to the background for all classes.

Variable	α_s	*Material*	*Production Place*	*Technique*	*Timespan*	*Depiction*	**Average**
Scenario A	1/2	**28.3** / 29.6	**27.0**	**57.8**	42.8	63.1	43.8 / 44.1
Scenario B	0	25.1 / 26.7	22.8	52.8	41.2	62.0	40.8 / 41.1
Scenario C	1/3	27.7 / 29.0	26.3	56.1	**43.4**	**66.3**	44.0 / 44.1
Scenario D	0	28.1 / 29.3	26.0	57.0	43.0	63.3	43.5 / 43.7
Scenario E	1/4	**29.6** / 29.7	26.7	57.3	42.9	65.6	**44.4 / 44.4**
SIR_LR_4	1	29.2 / 30.2	26.8	56.9	43.0	58.0	42.8 / 43.0

Table 3: Average F1-Scores [%] per variable for different scenarios of similarity as well as the best performing experiment of test step 1 (SIR_LR_4). For more details, see the caption of Table 2

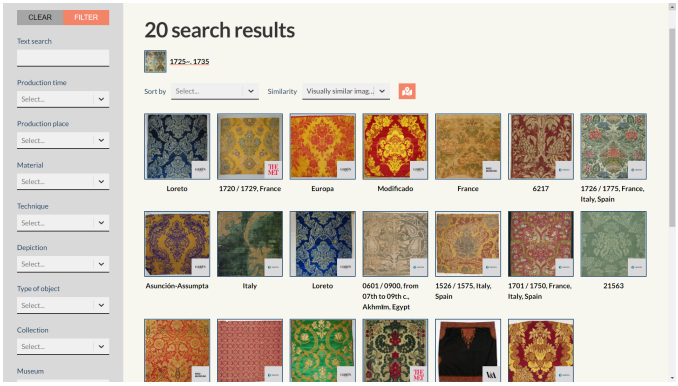

Figure 4: Objects that are visually similar with respect to an object produced in 1725-1735 in France using the embroidery technique and coming from the Art Institute of Chicago (ARTIC) museum.

differences. The integration of this module in a user friendly interface, an exploratory search engine, enables to conduct additional human evaluations.

ACKNOWLEDGMENT

This work has been partially supported by the European Union's Horizon 2020 research and innovation program within the SILKNOW (grant agreement No. 769504).

REFERENCES

[1] Matthew Addis, Mike Boniface, Simon Goodall, Paul Grimwood, Sanghee Kim, Paul Lewis, Kirk Martinez, and Alison Stevenson. 2003. SCULPTEUR: Towards a New Paradigm for Multimedia Museum Information Handling. In *Semantic Web ISWC (31/08/03)*. 582 –596. https://eprints.soton.ac.uk/258593/ Event Dates: September.

[2] J.L. Bentley. 1975. Multidimensional Binary Search Trees Used for Associative Searching. *Commun. ACM* 18, 9 (1975), 509–517. https://doi.org/10.1145/361002.361007

[3] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah. 1994. Signature verification using a "Siamese" time delay neural network. In *Advances in Neural Information Processing Systems (NIPS)*. 737–744.

[4] Yen-Wei Chen, Shota Sobue, and Xinyin Huang. 2009. KANSEI based clothing fabric image retrieval. In *International Workshop on Computational Color Imaging*. Springer, 71–80.

[5] Zheng Chen, Liu Wenyin, Feng Zhang, Mingjing Li, and Hongjiang Zhang. 2001. Web mining for web image retrieval. *Journal of the American Society for Information Science and Technology* 52, 10 (2001), 831–839.

[6] Sumit Chopra, Raia Hadsell, and Yann LeCun. 2005. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, Vol. 1. IEEE, 539–546.

[7] D. Clermont, M. Dorozynski, D. Wittich, and F. Rottensteiner. 2020. Assessing the semantic similarity of images of silk fabrics using convolutional neural networks. In *ISPRS Annals of the Photogrammetry, Remote Sensing & Spatial Information Sciences*, Vol. V-2. 641–648. https://doi.org/10.5194/isprs-annals-V-2-2020-641-2020

[8] Daiguo Deng, Ruomei Wang, Hefeng Wu, Huayong He, Qi Li, and Xiaonan Luo. 2018. Learning deep similarity models with focus ranking for fabric image retrieval. *Image and Vision computing* 70 (2018), 11–20.

[9] M. Doerr. 2005. The CIDOC CRM, an Ontological Approach to Schema Heterogeneity. In *Semantic Interoperability and Integration*.

[10] Thibault Ehrhart, Pasquale Lisena, and Raphaël Troncy. 2021. KG Explorer: a Customisable Exploration Tool for Knowledge Graphs. In 6^{th} *International Workshop on Visualization and Interaction for Ontologies and Linked Data (VOILA)*. Online.

[11] Noa Garcia, Benjamin Renoust, and Yuta Nakashima. 2020. ContextNet: representation and exploration for painting classification and retrieval in context. *International Journal of Multimedia Information Retrieval* 9, 1 (2020), 17–30.

[12] Noa Garcia and George Vogiatzis. 2018. How to read paintings: semantic art understanding with multi-modal retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*. 0–0.

[13] A. Gordo, J. Almazán, J. Revaud, and D. Larlus. 2016. Deep Image Retrieval: Learning Global Representations for Image Search. In *European Conference on Computer Vision (ECCV)*. 241–257.

[14] Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*. 855–864.

[15] Venkat N Gudivada and Vijay V Raghavan. 1995. Content based image retrieval systems. *Computer* 28, 9 (1995), 18–22.

[16] K. He, X. Zhang, S. Ren, and J. Sun. 2016. Identity mappings in deep residual networks. In *European Conference on Computer Vision (ECCV)*. 630–645.

[17] Anil K Jain and Aditya Vailaya. 1996. Image retrieval using color and shape. *Pattern recognition* 29, 8 (1996), 1233–1244.

[18] Thorsten Joachims. 2002. Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. 133–142.

[19] Inês Koch, Cristina Ribeiro, and Carla Lopes. 2020. *ArchOnto, a CIDOC-CRM-Based Linked Data Model for the Portuguese Archives*. 133–146. https://doi.org/10.1007/978-3-030-54956-5_10

[20] Mikko Koho, Esko Ikkala, Petri Leskinen, Minna Tamper, Jouni Tuominen, and Eero Hyvönen. 2021. WarSampo Knowledge Graph: Finland in the Second World War as Linked Open Data. *Semantic Web – Interoperability, Usability, Applicability* 12, 2 (January 2021), 265–278. https://doi.org/10.3233/SW-200392

[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton. 2012. ImageNet classification with deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems (NIPS)*. 1097–1105.

[22] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. 1989. Backpropagation applied to handwritten zip code recognition. *Neural computation* 1, 4 (1989), 541–551.

[23] Pasquale Lisena, Albert Meroño-Peñuela, Tobias Kuhn, and Raphaël Troncy. 2019. Easy Web API Development with SPARQL Transformer. In *18th International Semantic Web Conference (ISWC), In-use Track*. Auckland, New Zealand.

[24] Hui Mao, Ming Cheung, and James She. 2017. Deepart: Learning joint representations of visual arts. In *Proceedings of the 25th ACM international conference on Multimedia*. ACM, 1183–1191.

[25] V. Nair and G. E. Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *27th International Conference on Machine Learning (ICML)*. 807–814.

[26] Yonggang Qi, Yi-Zhe Song, Honggang Zhang, and Jun Liu. 2016. Sketch-based image retrieval via siamese convolutional neural network. In *2016 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2460–2464.

[27] F. Schroff, D. Kalenichenko, and J. Philbin. 2015. A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 815–823.

[28] Amit Singhal et al. 2001. Modern information retrieval: A brief overview. *IEEE Data Eng. Bull.* 24, 4 (2001), 35–43.

[29] Jiang Wang, Yang Song, Thomas Leung, Chuck Rosenberg, Jingbin Wang, James Philbin, Bo Chen, and Ying Wu. 2014. Learning fine-grained image similarity with deep ranking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1386–1393.

[30] Jun Xiang, Ning Zhang, Ruru Pan, and Weidong Gao. 2019. Fabric image retrieval system using hierarchical search based on deep convolutional neural network. *Ieee Access* 7 (2019), 35405–35417.

[31] Hsin-Chang Yang and Chung-Hong Lee. 2008. Image semantics discovery from web pages for semantic-based image retrieval using self-organizing maps. *Expert Systems with Applications* 34, 1 (2008), 266–279.

[32] Fang Zhao, Yongzhen Huang, Liang Wang, and Tieniu Tan. 2015. Deep semantic ranking based hashing for multi-label image retrieval. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1556–1564.

[33] Xiang Sean Zhou and Thomas S Huang. 2003. Relevance feedback in image retrieval: A comprehensive review. *Multimedia systems* 8, 6 (2003), 536–544.