EUROGRAPHICS Workshop on Graphics and Cultural Heritage (2017)
R. Sablatnig and B. Štular (Editors)

Short Paper

Cross-modal Content-based Retrieval for Digitized 2D and 3D Cultural Heritage Artifacts

Robert Gregor[1], Christof Mayrbrugger[1], Pavlos Mavridis[1], Benjamin Bustos[2], Tobias Schreck[1]

[1]Graz University of Technology, Austria
[2] University of Chile, Santiago de Chile

Abstract

Digitization of Cultural Heritage (CH) Objects is indispensable for many tasks including preservation, distributions and analysis of CH content. While digitization of 3D shape and appearance is progressing rapidly, much more digitized content is available in the form of 2D images, photographs, or sketches. A key functionality for exploring CH content is the ability to search for objects of interest. Search in CH repositories is often relying on meta-data of available objects. Also, methods for searching based on content in a given modality, e.g., using image or shape descriptors, are researched.

To date, few works have addressed the problem of content-based cross-modal search in both 2D and 3D object space without the requirement of meta data annotations of similar format and quality. We propose a cross-modal search approach relying on content-based similarity between 3D and 2D CH objects. Our approach converts a 3D query object into a 2D query image and then executes content-based search relying on visual descriptors. We describe our concept and show first results of our approach that were obtained on a pottery dataset. We also outline directions of future work.

1. Introduction

With advances in digitization, more and more Cultural Heritage (CH) content is becoming available, and in different modalities such as 2D image format, 3D object, textual transcriptions, among others. Much of this data for historic or pragmatic or licensing reasons, is distributed in different repositories. Content-based search allows users to find and explore CH content across distributed repositories even if the format of the available meta data is different. However, existing search methods support search only in a given modality. Cross-modal similarity search methods can enable linking related material across modality boundaries, and hence lead to new insights. For example, starting with a 3D digitized object of an artifact, an expert may be interested to find occurrences of similar shapes shown in existing, but very large 2D image repositories or in figures contained in domain publications or even on public web pages, where proper meta data annotation is usually absent. Multi-modal search can broaden the target search results, link across heterogeneous repositories, and hence support e.g., discovery of new relationships among objects, classifying material or adding textual descriptions found in figure captions.

Multimodal search can in principle be well supported by metadata approaches. In fact, much research has been published in the area. In practice, however there is often not enough descriptive metadata available to support search in heterogeneous repositories, e.g., because metadata may not be complete or not be comparable across different standards adopted in content curation. While content-based search methods for unimodal content exists, for multimodal content-based search we need to find a compatible representation by which we can compare content. Addressing the problem of 2D and 3D search, *view-based* approaches can provide such a compatible representation. We here contribute and apply a basic workflow for content-based retrieval of 2D images based on 3D queries. The approach relies on using standard 2D view features extracted from certain renderings of the query object and does not require extensive Machine Learning components. We present a first prototype, showing that standard view features extracted from textured renderings can retrieve matching 2D image content, providing robustness regarding image background. We present a preliminary study identifying appropriate view features and rendering parameters useful for cross-modal retrieval of 3D object views in 2D images. We also discuss our results in light of future research.

2. Related Work

As stated in the previous section, substantial research effort is focused on multi-modal retrieval. For example Europeana [PHSG17] essentially relies on harmonization of meta data across large collections of digitized CH content to also allow for multi-modal retrieval.

Wang et al. [WYW*16] provide a recent survey on research in the area of multi-modal and cross-modal Retrieval, that is backed by various machine learning approaches to establish cross-modal relationships within large collections of multimedia and text data.

DOI: 10.2312/gch.20171302

The approaches summarized there do not necessarily have to rely on the presence of extensive and harmonized meta data. Common ground among most of the approaches described there is that they are based on the idea that cross-modal relations can also be inferred by exploiting content-based similarity between data of the same modality as well as making use of the information that is implicitly provided in larger multi-modal collections of data, where, for a single topic or object, representations are available in multiple modalities simultaneously.

When it comes to cross-modal retrieval between 2D and 3D representations a lot of research has been conducted towards sketch based retrieval, where 3D objects can be retrieved by using queries that are based on quick 2D sketches [LLL*15].

Apart from the goal of cross-modal retrieval there are a number of potentially related publications in the field of 3D Retrieval. For example Su et al. [SMKLm15] introduce a view-based 3D retrieval method that trains a convolutional neural network using multiple 2D projections of 3D objects, effectively using 2D retrieval methods to retrieve 3D data. Several other works [XXS*15, RCM*17] are using a similar multi-view approach or panoramic projections [PPTP10] to intermix the 2D and 3D modalities. These view-based techniques work particularly well when smaller parts of geometric information are actually missing. For example, inside cavities, it has been demonstrated that when combined with robust descriptors [SMKLm15, RCM*17, PPTP10], they can result in state of the art retrieval performance. Our work is to some extent inspired on some of the ideas presented there. However, our goal is to retrieve 2D images by the aid of 3D models, which has not been paid much attention in research so far.

A distinguishing characteristic of commonly used views for 2D-based 3D retrieval is, that they do not tend to rely on texture information. In a different context, Biasotti et. al. [BCFS14] proposed to exploit this information by also combining classical image descriptors - computed from textures of 3D objects with standard 3D shape descriptors to establish similarities between 3D objects. In our work however, we do not have the possibility to exploit any 3D shape information of the target images.

3. Retrieval Pipeline

Our Retrieval pipeline is designed to retrieve 2D images that are similar to provided examples of 3D models. While our current work is geared towards the retrieval of photos by the aid of textured 3D models, the pipeline could be adapted to e.g., the retrieval of digitized sketches with presumably little change on the conceptual level.

First, photo-realistic renderings of digitized 3D cultural heritage models are created. In contrast to well-known approaches for view-based 3D descriptors, such as [PPTP10], we do not rely on spherical or cylindrical 2D projections. Even though they provide good effectiveness for 3D to 3D retrieval, it would be a very difficult and in many cases ill-posed task to process multiple photos of the same CH object to extract a single spherical or cylindrical 2D projection. Similar to many earlier view-based 3D descriptors, we instead extract multiple views with a size 600 by 600 pixels, that are obtained using a standard perspective projection using fully textured 3D models. However, during our experiments, we found that commonly applied 3D pose estimation (e.g., by PCA) actually degraded the retrieval performance. Most digitized 3D CH objects already seem to be oriented correctly, with respect to the vertical axis. We also discarded the estimation of rotation around the vertical axis, as many 2D photos of solid CH artifacts tend to be taken from different directions. Also, for such CH objects, photos are almost always taken from a downward angle. Therefore, view generation directly along lateral, longitudinal or vertical axis, as done for traditional view based descriptors, is not appropriate and actually degrades retrieval performance in our case. Instead, a total of 8 views are generated, each at a downward angle of 30 degrees in fixed positions around the object.

Second, various image features are extracted from the generated views. We initially used a broad range of well established 2D image features, with implementations from the Lucene Image Retrieval Project [Lir] that are in part based on OpenCV [Bra00]. They can be separated into global features (Auto Color Correlogram, Color and Edge Directivity Descriptor (CEDD) [CB08], Edge Histogram, FCTH, Gabor Texture Features, Tamura Image Features) and local features (SIFT [Low04], SURF [BTV06]) encoded to a Bag of Features (BoF) global descriptor with dictionary sizes of 64, 96 and 128. In addition, the SIMPLE descriptor extracts local MPEG-7 Visual descriptors [AMDA*99] in the local neighborhood of interest points, either detected by SURF or selected randomly and encoded to a Bag of Features (BoF) with vocabulary (i.e. dictionary) sizes of 64 and 128. For the extraction of the BoF vocabulary, we tested either only rendered views, photos or both combined. Finally, we also tested max-normalized, weighted, late fusion of SURF, SIMPLE and CEDD descriptors. All BoF-based features were computed using hard quantization.

Third, for each view of a query model, the distance to the photos is computed. For global feature types, specific distances were used (mostly L1 and L2) whereas L2 was computed for all of the BoF-based descriptors. For late-fusion, we experimented with interactively selected weights and conducted our final assessment with the following weighting of the distances with a dictionary size of 128:

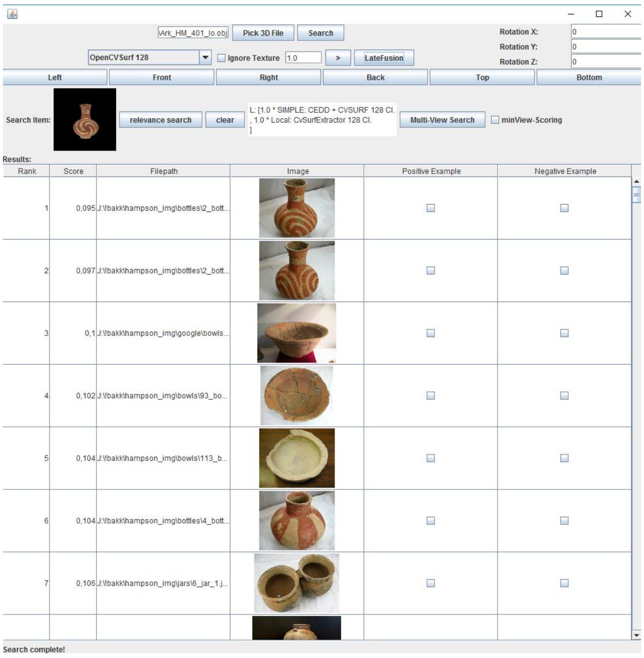
- $0.5 \cdot dist_{SURF} + 0.5 \cdot dist_{SIMPLE}$
- $0.1 \cdot dist_{CEDD} + 0.6 \cdot dist_{SURF} + 0.3 \cdot dist_{randomSIMPLE}$

When using multiple views for retrieval, the final ranking is obtained by relying on the minimum distances between one of the views and a result candidate. Initially we also considered basing the ranking of results on the average distance for views, however, this is very likely to work only in case of mostly highly rotationally symmetric objects.

4. First Results

We tested our approach on the well-known Hampson data set [Ham]. We selected a subset of 171 photos referring to objects from 6 different classes from the Hampson Museum website: *bottles*, *bowls*, *effigies*, *jars*, *lithies* and *others*.

Furthermore, a total of 26 additional photos of CH artifacts found on the web has been added to the classes *bottles*, *bowls*, *effigies* and *jars*. Finally, 10 images, consisting of random photos, gradients and



\Ark_HM_401_lo.obj

Rotation X: 0
Rotation Y: 0
Rotation Z: 0

OpenCVSurf 128 ☐ Ignore Texture 1.0 LateFusion

Search Item:

L: [1.0 * SIMPLE: CEDD + CVSURF 128 Cl.
, 1.0 * Local: CvSurfExtractor 128 Cl.
]

☐ minView-Scoring

Results:

Rank	Score	Filepath	Image	Positive Example	Negative Example
1	0,095	J:\!bakk\hampson_img\bottles\2_bott...		☐	☐
2	0,097	J:\!bakk\hampson_img\bottles\2_bott...		☐	☐
3	0,1	J:\!bakk\hampson_img\google\bowls...		☐	☐
4	0,102	J:\!bakk\hampson_img\bowls\93_bo...		☐	☐
5	0,104	J:\!bakk\hampson_img\bowls\113_b...		☐	☐
6	0,104	J:\!bakk\hampson_img\bottles\4_bott...		☐	☐
7	0,106	J:\!bakk\hampson_img\jars\6_jar_1.j...		☐	☐

Search complete!

Figure 1: *Our graphical tool used for finding initial parameter choices and feature combinations. In the query shown here, a single frontal view of a bottle was used with a fused combination of CEDD, SURF and SIMPLE features. In comparison to the feature type configurations that provide better precision and recall according to the benchmark, this result appears visually plausible to some extent, but seems to be much more sensitive to the overall color texture of the object.*

colors have been added as an additional *distraction* class. Note that this was done mostly to support the initial setup of parameters during interactive testing the pipeline with a Graphical User Interface (see Figure 1). Our dataset can be found at [Dat]

While all query objects and photos were used for the dictionary extraction. For obtaining first precision and recall results, we did not query the database with all objects from the collection but instead selected as subset of 2 -3 visually distinctive representatives as query objects for all classes except *distractions* and *ephigies*.

In order to quickly obtain initial results, we had to reduce the amount of computational effort required for computing the various feature, dictionary, fusing and view combinations. Also, during early testing it appeared, that the the geometrical and textural variance within the effigy class was too large when compared to the remainder of the dataset to yield meaningful results. Figure 2 shows our initial measurements. We first tested all of the mentioned features with views along the vertical, longitudinal and lateral axis with no upward or downward angle (top row in Figure 2). As the SURF feature outperformed all other feature combinations by a large amount, we limited the feature types to SURF variations in a second run (bottom row in Figure). This time, we used only views with a downward angle of about 30 degrees. Besides this, we also experimented with a dictionary size of 96. For both test runs, we tested with dictionaries extracted from only 2D photos (see Figure 2), 3D views and both combined.

5. Discussion, Applications & Future Work

Concerning the choice of training data for the dictionary, we observe that if more than the first results should be retrieved, it appears beneficial to only use features extracted from rendered views for training. When using only features extracted from photos for dictionary extraction, overall the first result tends to be slightly more precise over our test data set.

Considering our Precision and Recall scores, it appears that there is room for improvement. Especially when taking into account scores obtained by e.g. [BCFS14] for multi-modal 3D to 3D retrieval also using the Hampson dataset for evaluation. However, in our case, the task at hand is different and inherently much more difficult. We do not use any 3D data for the result candidates. Hence, a multi-modal approach can actually not be applied here. We can not rely on 3D shape information that is clearly separated from the texture of an object, instead it is inherently difficult to separate the object from its background.

Unfortunately, using multiple views appears to degrade the retrieval performance (see bottom row in Figure 2). Switching from directly frontal views to views from a downward angle improves the performance, for both, single and multi-view retrieval. This effect is strongest, when only rendered views are used for the dictionary extraction.

When it comes to using multiple views, a possible explanation for the degraded performance could be, that virtually all of the objects in the Hampson dataset are rotationally symmetric. Hence the actual perspective from which an object is observed has little effect in terms of generating false negatives. When it comes to bottles and jars, they are often distinguished by handle grips. When using the minimum distance from multiple views, this is a source for false positives in the ranking. Also, we observed that the visual variance within certain classes of the Hampson dataset appears to be larger than the smallest visual dissimilarities between certain objects of different classes (see Figure 3).

The bad performance of fused features might be explained by the fact that they do not rely on multiple modalities in contrast to Biasotti et al. [BCFS14]. Even there a combination of multi-modal features that includes very robust 3D shape descriptors (e.g. spectral features) only slightly outperforms the much earlier spherical harmonics descriptor. Apparently, fusion of features is actually a very hard problem that requires careful data and task specific adjustment of weighting and processing parameters.

More experiments need to be conducted when using the full Hampson dataset for queries. Also, the current view extraction could be improved in several ways. E.g., in many images, larger parts of objects are actually clipped from the photograph, so that in many cases, not the entire object is depicted, which could also be simulated in the rendered views. Relevance Feedback could be used to steer the weighting of individual features in the fused approach.

We already started to prepare a larger benchmark for cross-modal 3D to 2D retrieval based on other collections of digitized

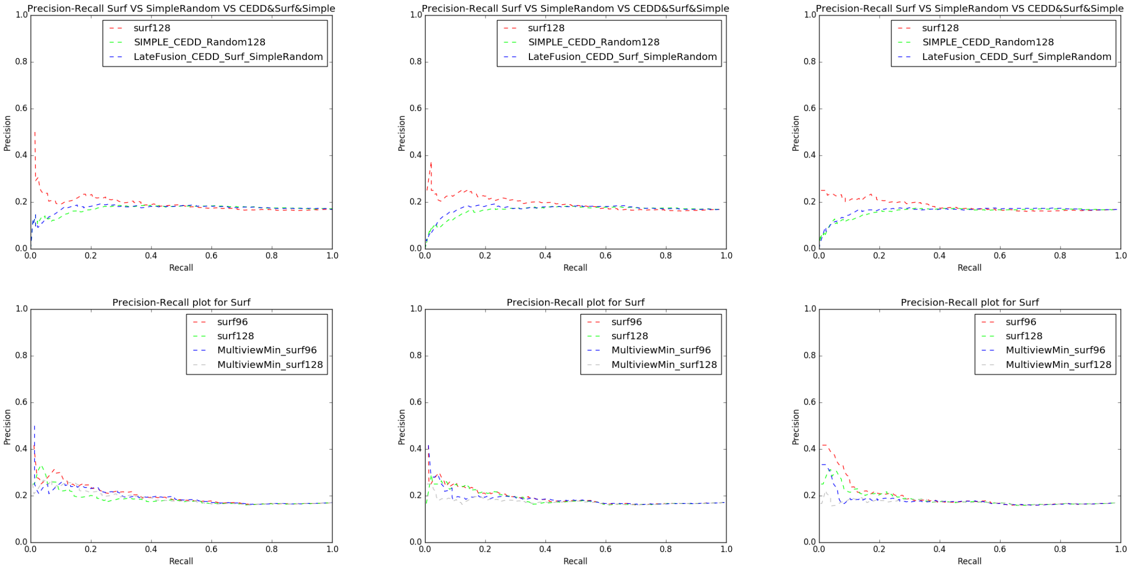


Figure 2: *Results obtained with various feature types, view extractions and dictionary variants. Top row: Initial results with frontal views Bottom row: Refined results with single/multiple views from 30 degrees above Left column: Dictionary extracted from images only, Right column: Images extracted from rendered views only, Middle columns: Dictionary based on photos and views.*

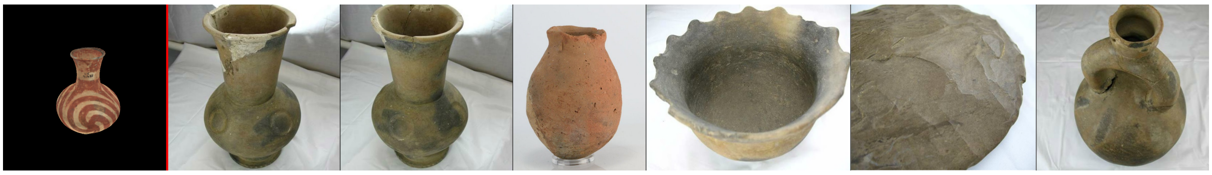


Figure 3: *Results for an exemplary query with a 3D model of class bottle (left) using SURF with a dictionary size of 128, trained on photos only using a single view at 30 degree downward angle. Classes of result items: bottles (1st, 2nd and 6th position), jars (3rd and 4th), lithies (5th position). Note that the similarity across two different classes (e.g. between jar and bottle) than the similarity between two objects of the same class (e.g. the bottles at 2nd and 6th position).*

CH collections. Clearly the goal here is to create a benchmark with much better class separation.

Concerning the feature encoding, there are also multiple ways for further improvement. Many recent publications in the field of 2D Computer Vision indicate, that retrieval performance can be improved by choosing either soft quantization to compute the BoFs or to switch to either Vectors of locally aggregated descriptors (VLAD) or Fisher Vectors. Also engineered features might be replaced with fully Machine Learning based approaches such as a CNN. However, we consider the amount of training data that is readily available to be insufficient.

6. Conclusion

We proposed a simple, yet easy to use approach for cross-modal content-based retrieval of CH photographs by queries based on 3D model examples that while taking some ideas from existing view based 3D retrieval approaches, rely on completely different features types. By testing a large range of feature types and parameters, we were able to identify an approach that provides promising first results. We outlined directions for further improvement of the approach and also provided examples of how such a system can be of practical benefit in the CH domain.

References

[AMDA*99] Abdel-Mottaleb M., Dimitrova N., Agnihori L., Dagtas S., Jeannin S., Krishnamachari S., McGee T., Vaithilingam G.: MPEG-7: a content description standard beyond compression. *42nd Midwest Symposium on Circuits and Systems 2*, August (1999), 770–777 vol. 2. 2

[BCFS14] Biasotti S., Cerri A., Falcidieno B., Spagnuolo M.: Similarity Assessment for the Analysis of 3D Artefacts. *Eurographics Workshop on Graphics and Cultural Heritage* (2014). 2, 3

[Bra00] Bradski G.: OpenCV Library. *Dr. Dobb's Journal of Software Tools* (2000). 2

[BTV06] Bay H., Tuytelaars T., Van Gool L.: SURF: Speeded up robust features. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 3951 LNCS* (2006), 404–417. 2

[CB08] Chatzichristofis S. A., Boutalis Y. S.: CEDD: Color and edge directivity descriptor: A compact descriptor for image indexing and retrieval. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 5008 LNCS* (2008), 312–322. 2

[Dat] The dataset used for our experiments. https://puck.cgv.tugraz.at/s/YkDo4mQ0oxnA798. Accessed: 2017-09-05. 3

[Ham] The virtual Hampson museum. http://hampson.cast.uark.edu/. Accessed: 2017-08-04. 2

[Lir] LIRE - Open Source Visual Information Retrieval. http://www.lire-project.net/. Accessed: 2017-09-05. 2

[LLL*15] Li B., Lu Y., Li C., Godil A., Schreck T., Aono M., Burtscher M., Chen Q., Chowdhury N. K., Fang B., Fu H., Furuya T., Li H., Liu J., Johan H., Kosaka R., Koyanagi H., Ohbuchi R., Tatsuma A., Wan Y., Zhang C., Zou C.: A comparison of 3D shape retrieval methods based on a large-scale benchmark supporting multimodal queries. *Computer Vision and Image Understanding 131* (2015), 1–27. 2

[Low04] Lowe D. G.: Distinctive Image Features from Scale-invariant Keypoints. *International Journal of Computer Vision 60*, 2 (2004), 91–110. 2

[PHSG17] Petras V., Hill T., Stiller J., Gäde M.: Europeana - a Search Engine for Digitised Cultural Heritage Material. *Datenbank-Spektrum 17*, 1 (2017), 41–46. 1

[PPTP10] Papadakis P., Pratikakis I., Theoharis T., Perantonis S.: Panorama: A 3D shape descriptor based on panoramic views for unsupervised 3d object retrieval. *International Journal of Computer Vision 89*, 2-3 (2010), 177–192. 2

[RCM*17] Rashwan H. A., Chambon S., Morin G., Gurdjos P., Charvillat V.: Towards Recognizing 3D Models Using A Single Image. In *Eurographics Workshop on 3D Object Retrieval* (2017), The Eurographics Association. 2

[SMKLm15] Su H., Maji S., Kalogerakis E., Learned-miller E.: Multi-view Convolutional Neural Networks for 3D Shape Recognition. *IEEE ICCV* (2015), 945–953. 2

[WYW*16] Wang K., Yin Q., Wang W., Wu S., Wang L.: A Comprehensive Survey on Cross-modal Retrieval. *IEEE ICCV* (2016), 1–20. 1

[XXS*15] Xie Z., Xu K., Shan W., Liu L., Xiong Y., Huang H.: Projective feature learning for 3d shapes with multi-view depth images. *Computer Graphics Forum 34*, 7 (2015), 1–11. 2