Multimodal Multimedia Retrieval with vitrivr

Gasser, Ralph ; Rossetto, Luca ; Schuldt, Heiko

DOI: https://doi.org/10.1145/3323873.3326921

Multimodal Multimedia Retrieval with vitrivr

Ralph Gasser
ralph.gasser@unibas.ch
University of Basel

Luca Rossetto
luca.rossetto@unibas.ch
University of Basel

Heiko Schuldt
heiko.schuldt@unibas.ch
University of Basel

ABSTRACT

The steady growth of multimedia collections – both in terms of size and heterogeneity – necessitates systems that are able to conjointly deal with several types of media as well as large volumes of data. This is especially true when it comes to satisfying a particular information need, i.e., retrieving a particular object of interest from a large collection. Nevertheless, existing multimedia management and retrieval systems are mostly organized in silos and treat different media types separately. Hence, they are limited when it comes to crossing these silos for accessing objects. In this paper, we present *vitrivr*, a general-purpose content-based multimedia retrieval stack. In addition to the keyword search provided by most media management systems, *vitrivr* also exploits the object's content in order to facilitate different types of similarity search. This can be done within and, most importantly, across different media types giving rise to new, interesting use cases. To the best of our knowledge, the full *vitrivr* stack is unique in that it seamlessly integrates support for four different types of media, namely images, audio, videos, and 3D models.

CCS CONCEPTS

• **Information systems** → **Multimedia information systems**; **Multimedia and multimodal retrieval**; • **Human-centered computing** → *Open source software*; • **Computing methodologies** → *Visual content-based indexing and retrieval*;

KEYWORDS

Content-based Retrieval, Multimodal Retrieval, Multimedia Retrieval, Query-by-Sketch, Query-by-Example, Sketch-based Retrieval, Concept-map, 3D Model Retrieval, Music Retrieval

ACM Reference Format:
Ralph Gasser, Luca Rossetto, and Heiko Schuldt. 2019. Multimodal Multimedia Retrieval with vitrivr. In *International Conference on Multimedia Retrieval (ICMR '19), June 10–13, 2019, Ottawa, ON, Canada.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3323873.3326921

1 INTRODUCTION

As media collections grow and become more diverse, the quest for accessing the knowledge contained in these collections becomes more challenging. This is mainly due to the lack of proper tools for satisfying a particular information need. The classical approach of annotating media objects and retrieving them later based on this textual metadata has several shortcomings. Firstly, the sheer amount of data and the pace at which the steady growth continues makes the laborious task of prior annotation ever more daunting. Secondly, textual descriptions tend to be subjective due to experience, expertise, language, and culture of the author. And thirdly, it is difficult to describe certain aspects, such as temporal evolution in videos or the shape of a 3D model, in a way that enables others to retrieve the desired object later. Generally, it can be stated that it is impossible to anticipate all potential future queries at the time of annotation, necessitating retrieval approaches that operate directly on a multimedia document's content.

In fact, there is an entire branch of research dedicated to solving the challenges that arise from these shortcomings for the different modalities such as audio, video, and text. Interestingly, however, most of the existing systems are only fitted towards a specific type of media and there is actually very little research on the implications for data management and retrieval that arise when combining different media types into a single system. This is contrasted by the fact that multimodality becomes more important in different areas, such as the cultural heritage domain, where, in addition to the classical images and text, videos, audio transcripts and 3D models start to play a more important role in daily business.

In this paper, we present vitrivr – a scalable, content-based multimedia information retrieval stack [17]. The work described herein builds on previous efforts for video collections [7, 14]. We have integrated different, media type specific content-based retrieval techniques into a single system so as to contrive a solution that is capable of managing and searching mixed multimedia collections. Query techniques include, but are not limited to, content-based visual image and video retrieval, audio retrieval, 3D model retrieval, retrieval based on object and action classification, and classical, text-based methods. The contribution of the work described herein is twofold: Firstly, we introduce the architecture and the features of vitrivr. Secondly, we demonstrate the effectiveness of vitrivr and the advantages that come with a unified system for managing heterogeneous multimedia collections and for searching objects across media types.

The remainder of this paper is structured as follows: Section 2 briefly surveys related work. Section 3 gives an overview of vitrivr's system architecture. Section 4 outlines the retrieval techniques we showcase and Section 5 describes vitrivr at work. Section 6 concludes.

2 RELATED WORK

Information retrieval in multimedia collections is an important topic in computer science with research focusing on many different domains, such as image and video retrieval [4, 16], music and audio retrieval [12, 24] or 3D model retrieval [23]. There also are many different evaluation campaigns where new systems and algorithms

ICMR '19, June 10–13, 2019, Ottawa, ON, Canada

ACM ISBN 978-1-4503-6765-3/19/06…$15.00
https://doi.org/10.1145/3323873.3326921

can compete, such as TRECVID[1] [1] or the Video Browser Showdown (VBS)[2] [22] for videos or the Music Information Retrieval Evaluation Exchange (MIREX)[3] for music.

However, there only seems to be little work on integrated solutions for content-based retrieval of different types of media. Most research focuses on a single modality such as the visual, the aural, or even further specialized subdomains, like for example, music, speech, or environmental sounds for audio. One example of an early general-purpose retrieval system is Query by Image Content (QBIC) [4]. The QBIC system allows for retrieval of images and videos based on image examples or sketches and it also considers object and background motion. However, it does not take any audio information into account. MUVIS [11] is a content-based multimedia retrieval and indexing framework for images and videos that builds upon this idea and closes the gap by adding support for audio, both standalone as well as interlaced with video. However, MUVIS is not publicly available and it limits itself to these three media types.

3 VITRIVR

vitrivr is the successor of the IMOTION system [16], which has originally been designed for multimedia retrieval in large video collections. Both IMOTION and vitrivr have been battle-tested on several occasions during different installments of the Video Browser Showdown (VBS) [3], which they won in 2017 [18] and 2019 [19], respectively. As part of our recent work, we have extended vitrivr [6] so as to support additional media types – namely images, audio, and 3D models – and query modes that support these new media types. The entire stack in its current version consists of the three layers as follows:

Vitrivr-NG A web-based user interface that facilitates query (re-)formulation, late score-fusion, and presentation of query results. A screenshot of the user interface is depicted in Figure 1.

Cineast A modular and extendable feature extraction and query processing engine. Cineast is responsible for generating the features during either the extraction (offline) or the query (online) process, for handing off the generated features to the storage engine, and for merging merging partial results generated through look-ups, i.e., fusion.

ADAM$_{pro}$ A polyglot database for multimedia retrieval that combines relational database functionality, full text indexing and search as well as nearest neighbor look-up used for vector space retrieval. ADAM$_{pro}$ offers both persistent on-disk storage and efficient look-up and operates efficiently on very large data sets.

On a very high level, the stack supports two major workflows. First, the *extraction workflow* (offline), which can be used to process different media objects, such as images, video or 3D model files, and to generate the relevant features using different *feature modules*. As a result of this workflow, the extracted features are persisted in ADAM$_{pro}$. Second, the *query workflow* (online), which is initiated by the users when they issue a query through the user interface. Once again, the relevant features are generated for that query and a sequence of database look-ups are being performed. The partial results are merged in a two-step fusion process which take place in Cineast and subsequently in Vitrivr-NG.

More details on the individual components and their inner workings can be found in [6, 7, 14]. The entire vitrivr stack is open source software and it can be downloaded from GitHub[4]. It also participated in Google Summer of Code (GSoC) 2016 and 2018, and many interesting (in most cases interdisciplinary) projects were realized on top of it.

4 MULTI-MODAL MULTIMEDIA RETRIEVAL

vitrivr supports several query modes, including Query-by-Sketch (QbS), Query-by-Example (QbE), relevance feedback, and textual queries. These query modes can be combined within a single query and can span several media modalities.

For the two-dimensional visual media types (images and videos), vitrivr leverages several low-level features, based on color, edge, and interest point (SIFT, SURF) information, which can be used for sketch- as well as example-based queries. In addition to visual sketches, vitrivr also supports semantic sketch queries, where colored regions represent a semantic concept previously identified by an automated object localizer [19].

In addition to two-dimensional media objects, vitrivr also supports retrieval of 3D models. In order to offer sketch-based querying for three dimensional content, vitrivr uses a series of projections of the 3D object onto circumscribing 2d planes. The outline of these projections can then be described using shape contour descriptors based on Zernike moments. These *light field descriptors* were proposed by [2]. This method can not only be applied to sketch input but also to photographs of a physical 3D object, which essentially bridges the gap between 2D and 3D representations. Similarity search with an example 3D model as query object is also possible. In this case we use shape descriptors based on the linear combination of spherical harmonics [9, 21].

Information from the aural domain (videos and music) can be queried using audio segments that can either be recorded directly via the user interface, imported from an external data source or reused from a previously retrieved result. Audio comparison is then performed using fingerprinting techniques [5, 25] and other audio descriptors such as CENS and Harmonic Pitch Class Profiles [8, 13].

Textual queries can be used for both inherently textual information, such as video subtitles, song lyrics, or metadata, as well as information extracted from media that can be efficiently represented textually, such as object or action labels. For example, we employed several deep neural networks to extract concept and action labels for video scenes or images [19].

All of the aforementioned query modes can be combined into a single query. The user interface depicted in Figure 1 can be used to formulate such queries. The mechanism to do so is organized around so-called *query containers*. A query container can consist of one to many *query terms* that can be toggled on and off. Each term covers a certain media modality, such as, 2D visual, 3D example, audio example, textual query, etc. Figure 2 depicts a few query terms, as they are currently displayed in Vitrivr-NG. The individual

[1] https://trecvid.nist.gov
[2] http://www.videobrowsershowdown.org
[3] https://www.music-ir.org/mirex/wiki/MIREX_HOME
[4] See https://github.com/vitrivr/

Figure 1: Screenshot of Vitrivr-NG, the web-based user interface. On the left-hand part, users can (re-)formulate queries. Results are presented in the center of the user interface. The right-hand part can be used to refine the result set and apply offline filters.

terms within the same container are connected by a fuzzy AND operation. Multiple query containers can be created by the end-user, which will be connected by a fuzzy OR operation.

5 VITRIVR AT WORK

At the ICMR 2019 conference, participants will have the opportunity to search and explore a large multimodal dataset using vitrivr. The dataset contains approximately 200'000 media objects and is composed of videos, audio clips, images, and 3D models from various sources, including DeviantArt[5], Pixabay[6], Thingiverse[7], LibriVox[8], the Free Music Archive[9] as well as material from freely available video datasets [15, 20].

Conference participants can try to solve various prepared scenarios (search tasks). In order to solve these tasks, users will have to find specific media objects in the collection based on on some hint provided, such as a description or some reference object. For example, users might search for a video scene based on an audio clip or some depiction of the desired scene. Or they have to query for a 3D model by creating a photograph of a physical model and feeding it to the vitrivr system. The scenarios will focus on use cases that highlight multimodal aspects. Hence, vitrivr will highlight how, for instance, a 2D representation can be used to find a 3D model or how combining the aural and visual modality simplifies the task of retrieving a particular video scene.

Of course, visitors will also have the opportunity to use the system freehand and simply browse and search the collection irrespective of the predefined scenarios. Overall, this should give the user a very good impression of what vitrivr is capable of. We expect it to launch discussions on how a multimodal retrieval system such as vitrivr could be put to use in the context of research topics and practical applications.

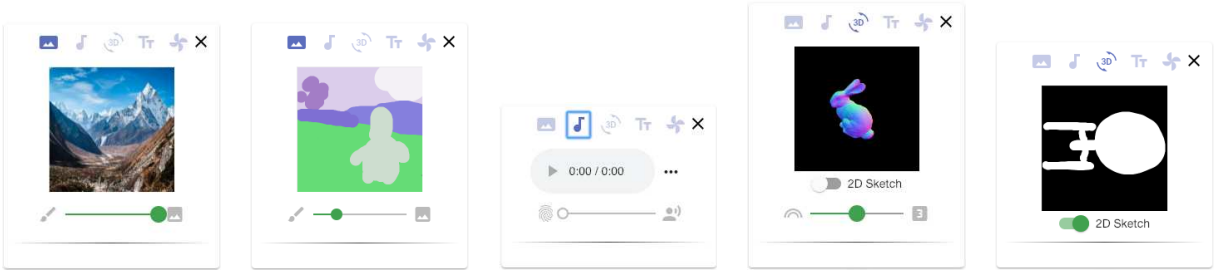
6 CONCLUSION

In this paper we have presented a new iteration of vitrivr, which has been demonstrated at ICMR 2019. vitrivr is a modular and extendable software stack that was originally designed for content-based video retrieval and has been enhanced to support multiple media types, namely audio, images, and 3D models. The new vitrivr version supports queries within and across the additional modalities that come with these new media types.

To the best of our knowledge, we have developed the first integrated, content-based multimedia retrieval solution, building further on the idea laid out in [10, 11]. Thereby, we have created the foundation for future work in the multimedia retrieval domain, as vitrivr can be considered a framework to design, implement, and test new retrieval techniques and to adapt them for specific use cases or concrete requirements.

In future work, we plan to add new features and build further on existing ones – e.g., our new semantic sketch functionality. We also

[5]https://www.deviantart.com/
[6]https://pixabay.com/
[7]https://www.thingiverse.com/
[8]https://librivox.org/
[9]http://freemusicarchive.org/



(a) QbE for video/image (b) QbS for video/image (c) QbE for audio (d) QbE for 3D models (e) QbS for 3D models

Figure 2: Illustration of how the query terms present themselves in the user interface. We have query terms for QbE (2a) and QbS (2b) for videos and images, QbE for audio (2c) and QbE (2d) and QbS (2e) for 3D models. In every case, the user can either select or create a reference object that is later used for lookup.

plan to address more complex composite object types that inherently combine different modalities (e.g., a music album that consists of several audio objects and an image representing the album cover) and to extend the multimodal search to these composite objects (e.g., to search for an audio object based on a sketch depicting the cover of an album). Moreover, we intend to continue to test vitrivr in competitive settings such as the *Video Browser Showdown* (VBS) or the more recent *Lifelog Search Challenge* (LSC). We strongly believe that the multimodal nature of vitrivr and its versatility, which we have demonstrated on many different occasions, could prove to be of great benefit in these diverse settings.

ACKNOWLEDGEMENTS

This work was partly supported by the Swiss National Science Foundation, project IMOTION (20CH21_151571).

REFERENCES

[1] George Awad, Asad Butt, Keith Curtis, Yooyoung Lee, Jonathan Fiscus, Afzal Godil, David Joy, Andrew Delgado, Alan F. Smeaton, Yvette Graham, Wessel Kraaij, Georges Quénot, Joao Magalhaes, David Semedo, and Saverio Blasi. 2018. TRECVID 2018: Benchmarking Video Activity Detection, Video Captioning and Matching, Video Storytelling Linking and Video Search. In *Proceedings of TRECVID 2018*. NIST, USA.

[2] Ding-Yun Chen, Xiao-Pei Tian, Yu-Te Shen, and Ming Ouhyoung. 2003. On Visual Similarity based 3D Model Retrieval. In *Computer Graphics Forum*, Vol. 22. Wiley Online Library, 223–232.

[3] Claudiu Cobârzan, Klaus Schoeffmann, Werner Bailer, Wolfgang Hürst, Adam Blažek, Jakub Lokoč, Stefanos Vrochidis, Kai Uwe Barthel, and Luca Rossetto. 2017. Interactive video search Tools: a Detailed Analysis of the Video Browser Showdown 2015. *Multimedia Tools and Applications* 76, 4 (2017), 5539–5571.

[4] Myron Flickner, Harpreet Sawhney, Wayne Niblack, Jonathan Ashley, Qian Huang, Byron Dom, Monika Gorkani, Jim Hafner, Denis Lee, Dragutin Petkovic, et al. 1995. Query by Image and Video Content: The QBIC System. *Computer* 28 (1995), 23–32.

[5] Jonathan T Foote. 1997. Content-based Retrieval of Music and Audio. In *Proc. SPIE 3229, Multimedia Storage and Archiving Systems II*. 138–147. https://doi.org/10.1117/12.290336

[6] Ralph Gasser, Luca Rossetto, and Heiko Schuldt. 2019. Towards an All-Purpose Content-Based Multimedia Information Retrieval System. *arXiv preprint arXiv:1902.03878* (2019).

[7] Ivan Giangreco and Heiko Schuldt. 2016. ADAM$_{pro}$: Database Support for Big Multimedia Retrieval. *Datenbank-Spektrum* 16, 1 (2016), 17–26. https://doi.org/10.1007/s13222-015-0209-y

[8] Emilia Gómez. 2006. *Tonal Description of Music Audio Signals*. Doctoral Dissertation. Universitat Pompeu Fabra, Barcelona. https://doi.org/10.1086/508205

[9] Michael Kazhdan, Thomas Funkhouser, and Szymon Rusinkiewicz. 2003. Rotation Invariant Spherical Harmonic Representation of 3D Shape Descriptors. In *Eurographics Symposium on Geometry Processing*, Vol. 43. 156–164.

[10] Patrick M. Kelly, Michael Cannon, and Donald R. Hush. 1995. Query by image example: the CANDID approach.

[11] Serkan Kiranyaz, Kerem Caglar, Esin Guldogan, Olcay Guldogan, and Moncef Gabbouj. 2003. MUVIS: a Content-based Multimedia Indexing and Retrieval Framework. In *Proceedings of the Seventh International Symposium on Signal Processing and Its Applications (ISSPA)*, Vol. 1. 1–8.

[12] Goujun Lu. 2001. Indexing and retrieval of Audio: A Survey. *Multimedia Tools and Applications* 15, 3 (2001), 269–290.

[13] Meinard Muller, Frank Kurth, and Michael Clausen. 2005. Chroma-based Statistical Audio Features for Audio Matching. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2005*. IEEE, New Paltz, NY, USA, 275–278. https://doi.org/10.1109/ASPAA.2005.1540223

[14] Luca Rossetto, Ivan Giangreco, and Heiko Schuldt. 2014. Cineast: a Multi-feature Sketch-based Video Retrieval Engine. In *2014 IEEE International Symposium on Multimedia*. IEEE, Taichung, Taiwan, 18–23.

[15] Luca Rossetto, Ivan Giangreco, and Heiko Schuldt. 2015. OSVC-Open Short Video Collection 1.0. *Technical Report* CS-2015-002 (2015).

[16] Luca Rossetto, Ivan Giangreco, Heiko Schuldt, Stéphane Dupont, Omar Seddati, Metin Sezgin, and Yusuf Sahillioğlu. 2015. IMOTION – a Content-based Video Retrieval Engine. In *International Conference on Multimedia Modeling*. Springer, 255–260.

[17] Luca Rossetto, Ivan Giangreco, Claudiu Tănase, and Heiko Schuldt. 2016. vitrivr: A Flexible Retrieval Stack Supporting Multiple Query Modes for Searching in Multimedia Collections. In *Proceedings of the 2016 ACM on Multimedia Conference*. ACM, Amsterdam, the Netherlands, 1183–1186.

[18] Luca Rossetto, Ivan Giangreco, Claudiu Tănase, Heiko Schuldt, Stéphane Dupont, and Omar Seddati. 2017. Enhanced Retrieval and Browsing in the IMOTION System. In *International Conference on Multimedia Modeling*. Springer, 469–474.

[19] Luca Rossetto, Mahnaz Amiri Parian, Ralph Gasser, Ivan Giangreco, Silvan Heller, and Heiko Schuldt. 2019. Deep Learning-Based Concept Detection in vitrivr. In *International Conference on Multimedia Modeling*. Springer, 616–621.

[20] Luca Rossetto, Heiko Schuldt, George Awad, and Asad A Butt. 2019. V3C–A Research Video Collection. In *International Conference on Multimedia Modeling*. Springer, 349–360.

[21] Dietmar Saupe and Dejan V. Vranić. 2001. 3D Model Retrieval with Spherical Harmonics and Moments. In *Proceedings of the 23rd DAGM-Symposium on Pattern Recognition*, Vol. 2191. Springer, 392–397. https://doi.org/10.1007/3-540-45404-7

[22] Klaus Schoeffmann, David Ahlström, Werner Bailer, Claudiu Cobârzan, Frank Hopfgartner, Kevin McGuinness, Cathal Gurrin, Christian Frisson, Duy-Dinh Le, Manfred Del Fabro, Hongliang Bai, and Wolfgang Weiss. 2014. The Video Browser Showdown: a Live Evaluation of Interactive Video Search Tools. *International Journal of Multimedia Information Retrieval* 3, 2 (2014), 113–127.

[23] Johan WH Tangelder and Remco C Veltkamp. 2004. A survey of Content Based 3D Shape Retrieval Methods. In *Shape Modeling Applications, 2004. Proceedings*. IEEE, 145–156.

[24] Rainer Typke, Frans Wiering, and Remco C Veltkamp. 2005. A Survey of Music Information Retrieval Systems. In *Proceedings of the 6th International Conference on Music Information Retrieval*. Queen Mary, University of London, 153–160.

[25] Avery Wang. 2006. The Shazam Music Recognition Service. *Commun. ACM* 49, 8 (2006), 44–48.