Deep Learning-based Concept Detection in vitrivr

Luca Rossetto[1], Mahnaz Amiri Parian[1,2], Ralph Gasser[1], Ivan Giangreco[1], Silvan Heller[1], and Heiko Schuldt[1]

[1] Databases and Information Systems Research Group
Department of Mathematics and Computer Science, University of Basel, Switzerland
{firstname.lastname}@unibas.ch
[2] Numediart Institute, University of Mons, Belgium

Abstract. This paper presents the most recent additions to the *vitrivr* retrieval stack, which will be put to the test in the context of the 2019 Video Browser Showdown (VBS). The *vitrivr* stack has been extended by approaches for detecting, localizing, or describing concepts and actions in video scenes using various convolutional neural networks. Leveraging those additions, we have added support for searching the video collection based on semantic sketches. Furthermore, *vitrivr* offers new types of labels for text-based retrieval. In the same vein, we have also improved upon *vitrivr*'s pre-existing capabilities for extracting text from video through scene text recognition. Moreover, the user interface has received a major overhaul so as to make it more accessible to novice users, especially for query formulation and result exploration.

1 Introduction

In this paper, we present the latest iteration of the open source, content-based multimedia retrieval stack *vitrivr* [17]. The presented system is a continuation of previous versions, which have been participating in the Video Browser Showdown [3] for several years now, first under the name IMOTION [14, 16, 18], and since 2018 under its current name – *vitrivr* [13]. In the vein of the mainstreamification of Deep Learning and convolutional neural networks (CNNs), the focus of this year's iteration of the *vitrivr* system lies on augmenting the existing sketch-based retrieval capabilities with semantic concept detection, description, localization, and scene text detection, by making use of various available off-the-shelf tools, pre-trained models, and existing as well as custom training datasets. Moreover, we have made numerous changes to the user interface to improve the user experience particularly for novice users for both query formulation and result exploration.

The remainder of this paper is structured as follows: Section 2 provides a brief overview of the overall architecture of the *vitrivr* system stack and its pre-existing capabilities. Section 3 introduces the new functionality which has been added to *vitrivr* for this iteration of the competition. Section 4 concludes.

2 System Overview and Existing Capabilities

vitrivr is a content-based multimedia retrieval stack capable of retrieving from mixed media collections containing images, audio, video, and 3D data. In the context of the VBS competition, only the video retrieval capabilities are relevant. The *vitrivr* stack is comprised of three components: the database system ADAM$_{pro}$ [7], the retrieval engine Cineast [15], and the user interface vitrivr-ng. The user interface is browser-based and can be served either directly by Cineast or via an external web server. More details with respect to the architecture of the entire system can be found in [17].

3 New Functionality

In this section, we highlight the additions made to the *vitrivr* stack. They are primarily based on functionality provided by various neural network architectures. The first group of these additions generates various forms of textual output from a given video scene. These methods include scene text detection and extraction as well as scene labeling and captioning. The second group operates based on concept detection and localization, which is queried via a sketched input. Figure 1 depicts an overview of the different types of features, both new and pre-existing. In addition, the user interface was improved to increase the efficiency in query expression and result browsing in a competitive setting.

Fig. 1. Overview of the different types of features employed by the *vitrivr* system.

3.1 Scene Text Detection and Recognition

A straightforward way of searching for a particular scene is to use any visible text within that scene as a means for querying. Scene text detection and recognition involves the localization and transcription of textual objects in images. We perform the detection and recognition by leveraging and combining several different neural network-based concept detectors. The new end-to-end scene text detection and recognition module in Cineast is based on a combination of the work presented in [19] and [23] and implemented using the TensorFlow [1] Java API. The module is used during the off-line extraction phase to generate text labels for each keyframe extracted from a video. These labels can then be used on-line, i.e., at retrieval time, during the competition.

The first step during the extraction involves identifying potential text objects in a scene using EAST [23]. EAST leverages a fully connected CNN to generate bounding boxes for areas that contain text. In a second step, we extract the sub-images delimited by those bounding boxes and use a convolutional recurrent neural network (CRNN) – a combination of a CNN and a recurrent neural network, as proposed by [19] – to infer the text in the respective sub-image. We trained the CRNN network using the full MJSynth dataset [8], whereas the pre-trained model provided by the authors was used for EAST.

3.2 Captioning and Labeling

The ex post analysis of the approaches used most at the last iteration of VBS showed that –in particular with increasing collection sizes– concept-based searching is the preferred way of searching. Concept-based searching allows to search for semantic information included in a particular scene. We would like to use that information –that is, the produced labels and captions– to be able to perform a textual lookup during retrieval phase. With the VBS tasks, especially the KIS Textual task in mind, we have identified two objectives:

One objective is for *vitrivr* to be able to perform scene-based action recognition and to label the scenes accordingly. Examples of such action labels involve terms like "horse riding" or "rock climbing" if the scene depicts a person carrying out the respective action. In our implementation, labeling is mainly based on spatio-temporal information extracted from subsequent frames by a 3DCNN [9, 20]. More precisely, we employ spatio-temporal feature extraction by 3DConvNets proposed in [20]. This architecture[3] takes multiple consecutive frames as input and, in addition to spatial features, extracts motion features which together enable the recognition and classification of the action taking place in the current shot.

The second objective involves key frame-based captioning of a scene to describe it semantically, that is, describing on a high level what that scene depicts. An example could be a sentence such as "a white cow grazing on a meadow". We use CNN and LSTM networks to achieve this, as proposed by [21]. We apply

[3] https://github.com/hx173149/C3D-tensorflow

the proposed methods to the representative frame of every shot. The network producing the captions has been trained on [10]. Since this method generates multiple candidate captions, which might describe different aspects of the input image, we store the three most likely captions per segment.

3.3 Semantic Sketches

Inspired by the work presented in [6], we added a new type of sketch-based querying, which uses common semantic concepts rather than colors. We hope that this allows for a more intuitive search, in particular for novice users of the *vitrivr*system. The maps describing the localization of the concepts have been obtained using a DeepLab network [2] trained on three image datasets containing concept-instances from different contexts [4, 5, 22].

The obtained object maps are quantized into an $n \times n$ grid where the most extensive concept is used as a label for every cell. For every concept, a two-dimensional coordinate point has been pre-computed based on a 2D-embedding [11] of semantic distances [12] between the concepts. To generate a vector from the previously obtained grid, the 2D-coordinates per concept are simply concatenated for every cell in a pre-determined order, resulting in a vector of length $2n^2$. This leads to a compact representation, which still retains some notion of similarity between the different concepts as well as their spatial relation within the scene.

3.4 User Interface Improvements

The extensions of the user interface implemented for the VBS 2019 version of *vitrivr* primarily address the efficiency and speed of both query formulation and results exploration. In particular, the user interface has been improved with novice users in mind to ensure that they are able to use *vitrivr* right away and that the key functionality is accessible in a more intuitive fashion.

While the previous version of the user interface was merely an adaption of a general purpose UI aimed at content-based multimedia retrieval, the UI used in this iteration of the competition is geared more towards the competitive nature of the VBS setting. This is achieved by restructuring several workflows so as to reduce the number of clicks necessary to access the functionality relevant to the competition.

Moreover, we have substantially increased the number of results that are displayed in the interface in order to provide better browsing capabilities. To increase the browsing efficiency, the user interface has been extended by filters for certain visual characteristics (e.g., filter for colored shots or filter for black and white shots) or for excluding previously seen results from a previous query. Especially for the latter, such filter information can also be transmitted to other team members using their own instance of the interface, which enables a high degree of collaboration.

4 Conclusions

In this paper, we have presented recent additions to the *vitrivr* system in order to improve its video retrieval capabilities, especially in a competitive setting. The presented additions have focused on the use of Deep Learning techniques and the improvement of the user interface, in particular towards novice users. The *vitrivr* stack is released as open source software[4] under the MIT license.

References

1. Martín Abadi, Paul Barham, Jianmin Chen, et al. Tensorflow: A system for large-scale machine learning. In *Proceedings of the USENIX Symposium on Operating Systems Designand Implementation (OSDI)*, volume 16, pages 265–283, Savannah, GA, USA, 2016. USENIX.
2. Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the 15th European Conference on Computer Vision (ECCV)*, page to appear, Munich, Germany, 2018.
3. Claudiu Cobârzan, Klaus Schoeffmann, Werner Bailer, Wolfgang Hürst, Adam Blažek, Jakub Lokoč, Stefanos Vrochidis, Kai Uwe Barthel, and Luca Rossetto. Interactive Video Search Tools: a detailed Analysis of the Video Browser Showdown 2015. *Multimedia Tools and Applications (MTAP)*, 76(4):5539–5571, 2017.
4. Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3213–3223, Las Vegas, NV, USA, 2016. IEEE.
5. Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of computer vision (IJCV)*, 111(1):98–136, 2015.
6. Ryosuke Furuta, Naoto Inoue, and Toshihiko Yamasaki. Efficient and interactive spatial-semantic image retrieval. In *Proceedings of the International Conference on Multimedia Modeling (MMM)*, pages 190–202, Bangkok, Thailand, 2018. Springer.
7. Ivan Giangreco and Heiko Schuldt. ADAM$_{pro}$: Database Support for Big Multimedia Retrieval. *Datenbank-Spektrum*, 16(1):17–26, 2016.
8. Max Jaderberg, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Synthetic data and artificial neural networks for natural scene text recognition. *arXiv preprint arXiv:1406.2227*, pages 1–10, 2014.
9. Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1725–1732, Columbus,OH,USA, 2014. IEEE.
10. Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[4] https://github.com/vitrivr

11. Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research (JMLR)*, 9(Nov):2579–2605, 2008.
12. Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, pages 1–12, 2013.
13. Luca Rossetto, Ivan Giangreco, Ralph Gasser, and Heiko Schuldt. Competitive video retrieval with vitrivr. In *Proceedings of the International Conference on Multimedia Modeling (MMM)*, pages 403–406, Bangkok, Thailand, 2018. Springer.
14. Luca Rossetto, Ivan Giangreco, Silvan Heller, Claudiu Tănase, Heiko Schuldt, Stéphane Dupont, Omar Seddati, Metin Sezgin, Ozan Can Altıok, and Yusuf Sahillioğlu. IMOTION – Searching for Video Sequences using Multi-shot Sketch Queries. In *Proceedings of the International Conference on Multimedia Modeling (MMM)*, pages 377–382, Miami, FL, USA, January 2016. Springer.
15. Luca Rossetto, Ivan Giangreco, and Heiko Schuldt. Cineast: a Multi-feature Sketch-based Video Retrieval Engine. In *Proceedings of the International Symposium on Multimedia (ISM)*, pages 18–23, Taichung, Taiwan, December 2014. IEEE.
16. Luca Rossetto, Ivan Giangreco, Heiko Schuldt, Stéphane Dupont, Omar Seddati, Metin Sezgin, and Yusuf Sahillioğlu. IMOTION - A Content-based Video Retrieval Engine. In *Proceedings of the International Conference on Multimedia Modeling (MMM)*, pages 255–260, Sydney, Australia, January 2015. Springer.
17. Luca Rossetto, Ivan Giangreco, Claudiu Tănase, and Heiko Schuldt. vitrivr: A Flexible Retrieval Stack Supporting Multiple Query Modes for Searching in Multimedia Collections. In *Proceedings of the ACM Conference on Multimedia Conference (ACM MM)*, pages 1183–1186, Amsterdam, The Netherlands, October 2016. ACM.
18. Luca Rossetto, Ivan Giangreco, Claudiu Tănase, Heiko Schuldt, Stéphane Dupont, and Omar Seddati. Enhanced Retrieval and Browsing in the IMOTION System. In *Proceedings of the International Conference on Multimedia Modeling (MMM)*, pages 469–474, Reykjavik, Iceland, January 2017. Springer.
19. Baoguang Shi, Xiang Bai, and Cong Yao. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 39(11):2298–2304, 2017.
20. Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 4489–4497, Santiago, Chile, 2015. IEEE.
21. Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 39(4):652–663, 2017.
22. Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, page 4, Honolulu, HI, USA, 2017. IEEE.
23. Xinyu Zhou, Cong Yao, He Wen, Yuzhi Wang, Shuchang Zhou, Weiran He, and Jiajun Liang. East: an efficient and accurate scene text detector. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2642–2651, Honolulu, HI, USA, 2017. IEEE.