Interactive video retrieval in the age of deep learning

Jakub Lokoč
lokoc@ksi.mff.cuni.cz
Charles University
Prague, Czech Republic

Klaus Schoeffmann
ks@itec.aau.at
Klagenfurt University
Klagenfurt, Austria

Werner Bailer
werner.bailer@joanneum.at
Joanneum Research
Graz, Austria

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

Cathal Gurrin
cathal.gurrin@dcu.ie
Dublin City University
Dublin, Ireland

ABSTRACT

We present a tutorial focusing on video retrieval tasks, where state-of-the-art deep learning approaches still benefit from interactive decisions of users. The tutorial covers general introduction to the interactive video retrieval research area, state-of-the-art video retrieval systems, evaluation campaigns and recently observed results. Moreover, a significant part of the tutorial is dedicated to a practical exercise with three selected state-of-the-art systems in the form of an interactive video retrieval competition. Participants of this tutorial will gain a practical experience and also a general insight of the interactive video retrieval topic, which is a good start to focus their research on unsolved challenges in this area.

KEYWORDS

Interactive video retrieval, deep learning, evaluation campaigns

ACM Reference Format:
Jakub Lokoč, Klaus Schoeffmann, Werner Bailer, Luca Rossetto, and Cathal Gurrin. 2019. Interactive video retrieval in the age of deep learning. In *International Conference on Multimedia Retrieval (ICMR '19), June 10–13, 2019, Ottawa, ON, Canada.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3323873.3326588

1 MOTIVATION FOR THE TOPIC

With the rise of available large annotated collections (e.g., ImageNet [2]) and computation resources (e.g., over 125 TFLOPS per second for deep learning with NVIDIA Volta[1]), new machine learning models gradually increase precision of automatic annotation tasks. Yet, many video retrieval scenarios focusing on recall remain challenging, as evident from recent TRECVID [1] and Video Browser Showdown (VBS) [6, 7] publications. Let us note that tasks focusing on recall remain challenging also in related domains and competitions, for example at the Lifelog Search Challenge [4].

[1]https://www.nvidia.com/en-us/data-center/volta-gpu-architecture

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

ACM ISBN 978-1-4503-6765-3/19/06…$15.00
https://doi.org/10.1145/3323873.3326588

One of the reasons for this problem is clearly the discrepancy between the search intent of a user and the provided search features of a video retrieval tool. For example, consider the situation when a user needs to find videos showing some '*suspicious activities*'. How would they formulate this search intent as a text query? Such browsing-type of queries are only possible with interactive search tools that provide a combination of automatic retrieval and interactive search. Moreover, consider the fuzzy but also quite specific query "*Find the situation of a person sitting in a car, with the steering wheel visible, for the moment when the car starts moving.*" – a similar query was part of the VBS2019 competition. Such a query can be accurately solved with an interactive video retrieval tool, while a typical text-based retrieval engine with state-of-the-art concept detection would still face problems with many false hits. In many real life scenarios we even have more troubles translating our – sometimes very specific but also personal – memories of a scene or a moment to the provided semantic concepts of a textual retrieval engine.

This tutorial is based on a tutorial at ACM MM 2018 title "Where is the user in the age of deep learning?" [13], which has been presented to an engaged group of 50-60 people in Seoul, Korea, in October 2018. However, the current version of this tutorial has been extended by an interactive section that lets users experience the setting of interactive video retrieval benchmarks. The most recent results from the Video Browser Showdown 2019 are included too.

2 COURSE DESCRIPTION

The tutorial is planned as a *half-day* event divided into four parts. After a general introduction to the topic and three selected existing systems, the second part is planned as an entertaining interactive video retrieval competition organized for tutorial participants. Once participants gain a direct experience with three video retrieval systems and tasks, the tutorial participants will get an overview of interactive multimedia evaluation campaigns and observed results from recent events.

2.1 Interactive video retrieval

The first part of the tutorial presents an overview of challenging video retrieval tasks and motivates for interactive means of retrieval. Different approaches to integrate users into the search process are discussed in connection with popular deep learning approaches providing effective ranking of searched scenes. More specifically,

the following three state-of-the-art tools are considered for the tutorial:

- *VIRET* [8, 9] is a frame-based multi-modal video retrieval system developed at Charles University, Czech Republic. VIRET regularly (and successfully) participates at international interactive multimedia retrieval campaigns. The system enables specification of multi-modal temporal queries targeting a sequence of (different) frames. The results are presented as representative frames and video summaries, with an option to quickly inspect temporal context of each displayed frame. The system employs deep convolutional neural networks (currently NasNet [15]) for automatic frame annotation, face/text bounding box detection and embedding of frames to a vector representation.
- *vitrivr* [10, 11] is a content-based multimedia retrieval stack developed at University of Basel, Switzerland. The system supports several query- and media types and successfully participated at VBS for several years. It has served as a platform for experimentation with various retrieval approaches, feature representations as well as query formulation and result presentation strategies. Its most recent addition is a semantic-sketch query mode, which enables users to intuitively describe a scene both spatially as well as semantically.
- *diveXplore* [14] is a distributed interactive video exploration system developed at Klagenfurt University, Austria. The system provides a set of components enabling interactive video retrieval for different types of tasks and search scenarios. Characteristic diveExplore functionalities are interactive browsing in pre-computed feature maps and support for collaborative retrieval. The good performance of the system at the Video Browser Showdown and Lifelog Search Challenge demonstrates that interactive browsing is a competitive search strategy with respect to query-oriented systems.

2.2 Live evaluation

The second part of the tutorial is intended as a practical exercise in the form of a lightweight installment of the Video Browser Showdown competition [6, 7] for the tutorial participants. Following the success of novice user sessions from the Video Browser Showdown, we believe that the participants quickly gain an experience with the topic, tools, tasks and related problems. The tools can be also used to demonstrate the practical effectiveness of involving deep learning approaches. After a closer introduction of the tools by the organizers, two types of tasks are planned for the participants – visual known-item search and ad-hoc search tasks. Visual known-item search tasks will be presented on a data projector by playing the searched scene in the loop, while ad-hoc search tasks will be presented as a short textual description. The actual score of the participants will be shown on the VBS server.

2.3 Evaluation campaigns

The third part will provide an overview of existing evaluation campaigns, such as the VBS [6, 7], the LSC [3] or TRECVID [1], outline their tasks, goals, commonalities and differences and discuss their evaluation strategies. The choice of evaluation strategies is not only influenced by aspects such as repeatability and the reuse of assessments, but also impacted by the setting of the evaluation campaign, i.e., whether the competition is live in front of the audience (as e.g. in VBS) or an offline process (as e.g. in TRECVID). The history of selected evaluation campaigns is briefly described, and examples of tasks from TRECVID, VBS and LSC are reviewed, in order to illustrate specific evaluation goals and task settings.

Task design also addresses the question *who* performs the task. While it is usually the developers of the teams who participate in evaluation campaigns, "novice" sessions, in which members of the audience use the tools, provide valuable insights into the complexity and usability of the tools. In many application areas, tools are likely to be used by domain experts rather than retrieval experts, thus this condition models real situations.

This part will also discuss the datasets [5, 12] currently used for these campaigns. It will discuss the selection and preparation of large-scale datasets and methods of generating ground truth. The aspects of dataset generation are put in relation to task types, e.g., concerning the effort for creating ground truth, covering the complete dataset and the reusability of annotations in other settings.

2.4 Recent results and future directions

The last part of the tutorial focuses on observed results at the Video Browser Showdown [6, 7] and Lifelog Search Challenge [3]. For both competitions, we present evaluated tasks, metrics, scoring and constraints. We discuss recent successful information retrieval trends observed at the competitions and their impact on the settings of evaluated tasks. For example, the Video Browser Showdown 2019 has revealed that state-of-the-art AVS and OCR models are already very efficient and effective for the current settings of visual known-item search tasks. Hence, the task presentation could be changed towards a more realistic setting. We also report our log analysis and show that many teams have found the searched item but did not register it on the display. Hence, result presentation is essential for efficient retrieval and browsing. We conclude that there is still no clear dominant search strategy for the considered types of tasks.

3 ORGANIZER BIOGRAPHIES

3.1 Jakub Lokoč

Jakub Lokoč is an Assistant Professor with the Department of Software Engineering, Faculty of Mathematics and Physics, Charles University. His research interests include metric indexing, multimedia databases, video retrieval, known-item search, and similarity modeling. He is a co-organizer of the Video Browser Showdown competition (VBS).

3.2 Klaus Schoeffmann

Klaus Schoeffmann is an associate professor at the Institute of Information Technology (ITEC) at Klagenfurt University, Austria. His research focuses on medical multimedia systems, video understanding, and interactive multimedia. He has co-authored more than 110 publications on various topics in multimedia and he has co-organised several international conferences, workshops, and special sessions in the field of multimedia. Furthermore, he is co-founder of the Video Browser Showdown (VBS) – an international live evaluation competition of interactive video search.

3.3 Werner Bailer

Werner Bailer is a key researcher at DIGITAL – Institute for Information and Communication Technologies at JOANNEUM RESEARCH in Graz, Austria. He received a degree in Media Technology and Design in 2002 for his diploma thesis on motion estimation and segmentation for film/video standards conversion. His research interests include digital film restoration, audiovisual content analysis and retrieval as well as multimedia metadata. He regularly contributes to standardization in MPEG and to EBU working groups, has co-organized Video Browser Showdown since 2012 and contributed to the TRECVID and MediaEval benchmarks.

3.4 Luca Rossetto

Luca Rossetto is a postdoctoral researcher at the Databases and Information Systems Group at the University of Basel. His research focus is on content-based multimedia retrieval with focus on video. He is one of the core developers of the open-source multimedia retrieval engine vitrivr.

3.5 Cathal Gurrin

Cathal Gurrin is an associate professor at Dublin City University and a principal co-investigator at the Insight Centre for Data Analytics. His research focuses on the gathering, indexing and retrieval of various forms of multimedia data, with a specific focus on lifelog and personal data. He is a founding member of the teams running the NTCIR-Lifelog, ImageCLEF Lifelog and LSC comparative benchmarking exercises.

ACKNOWLEDGMENTS

This paper has been supported by Czech Science Foundation (GAČR) project Nr. 19-22071Y and Science Foundation Ireland (SFI) under grant Nr. SFI/12/RC/2289.

REFERENCES

[1] George Awad, Asad Butt, Jonathan Fiscus, Martial Michel, David Joy, Wessel Kraaij, Alan F. Smeaton, Georges Quénot, Maria Eskevich, Roeland Ordelman, Gareth J. F. Jones, and Benoit Huet. 2017. TRECVID 2017: Evaluating Ad-hoc and Instance Video Search, Events Detection, Video Captioning and Hyperlinking. In *Proceedings of TRECVID 2017*. NIST, USA.

[2] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 248–255. https://doi.org/10.1109/CVPR.2009.5206848

[3] Cathal Gurrin, Klaus Schoeffmann, Hideo Joho, Andreas Leibetseder, Liting Zhou, Aaron Duane, Duc-Tien Dang-Nguyen, Michael Riegler, Luca Piras, Minh-Triet Tran, Jakub Lokoč, and Wolfgang Hürst. 2019. Comparing Approaches to Interactive Lifelog Search at the Lifelog Search Challenge (LSC2018). *ITE Transactions on Media Technology and Applications* 7 (04/2019 2019), 46–59. https://doi.org/10.3169/mta.7.46

[4] Cathal Gurrin, Klaus Schoeffmann, Hideo Joho, Andreas Leibetseder, Liting Zhou, Aaron Duane, Duc-Tien Dang-Nguyen, Michael Riegler, Luca Piras, Minh-Triet Tran, Jakub Lokoč, and Wolfgang Hürst. 2019. [Invited papers] Comparing Approaches to Interactive Lifelog Search at the Lifelog Search Challenge (LSC2018). *ITE Transactions on Media Technology and Applications* 7, 2 (2019), 46–59. https://doi.org/10.3169/mta.7.46

[5] Cathal Gurrin, Klaus Schoeffmann, Hideo Joho, Bernd Munzer, Rami Albatal, Frank Hopfgartner, Liting Zhou, and Duc-Tien Dang-Nguyen. 2019. A Test Collection for Interactive Lifelog Retrieval. In *MultiMedia Modeling*, Ioannis Kompatsiaris, Benoit Huet, Vasileios Mezaris, Cathal Gurrin, Wen-Huang Cheng, and Stefanos Vrochidis (Eds.). Springer International Publishing, Cham, 312–324.

[6] Jakub Lokoc, Gregor Kovalcík, Bernd Münzer, Klaus Schöffmann, Werner Bailer, Ralph Gasser, Stefanos Vrochidis, Phuong Anh Nguyen, Sitapa Rujikietgumjorn, and Kai Uwe Barthel. 2019. Interactive Search or Sequential Browsing? A Detailed Analysis of the Video Browser Showdown 2018. *TOMCCAP* 15, 1 (2019), 29:1–29:18.

[7] Jakub Lokoč, Werner Bailer, Klaus Schoeffmann, Bernd Münzer, and George Awad. 2018. On Influential Trends in Interactive Video Retrieval: Video Browser Showdown 2015-2017. *IEEE Trans. Multimedia* 20, 12 (2018), 3361–3376. https://doi.org/10.1109/TMM.2018.2830110

[8] Jakub Lokoč, Gregor Kovalčík, Tomáš Souček, Jaroslav Moravec, and Přemysl Čech. 2019. VIRET: A Video Retrieval Tool for Interactive Known-item Search. In *International Conference on Multimedia Retrieval (ICMR '19), June 10–13, 2019, Ottawa, ON, Canada*. 1–5. https://doi.org/10.1145/3323873.3325034

[9] Jakub Lokoč, Tomáš Souček, and Gregor Kovalčík. 2018. Using an Interactive Video Retrieval Tool for LifeLog Data. In *Proceedings of the 2018 ACM Workshop on The Lifelog Search Challenge, LSC@ICMR 2018, Yokohama, Japan, June 11, 2018*. 15–19. https://doi.org/10.1145/3210539.3210543

[10] Luca Rossetto, Ivan Giangreco, Claudiu Tanase, 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, 1183–1186.

[11] Luca Rossetto, Mahnaz Amiri Parian, Ralph Gasser, Ivan Giangreco, Silvan Heller, and Heiko Schuldt. 2019. Deep Learning-Based Concept Detection in vitrivr. In *MultiMedia Modeling - 25th International Conference, MMM 2019, Thessaloniki, Greece, January 8-11, 2019, Proceedings, Part II*. 616–621. https://doi.org/10.1007/978-3-030-05716-9_55

[12] Luca Rossetto, Heiko Schuldt, George Awad, and Asad A. Butt. 2019. V3C - A Research Video Collection. In *MultiMedia Modeling - 25th International Conference, MMM 2019, Thessaloniki, Greece, January 8-11, 2019, Proceedings, Part I*. 349–360. https://doi.org/10.1007/978-3-030-05710-7_29

[13] Klaus Schoeffmann, Werner Bailer, Cathal Gurrin, George Awad, and Jakub Lokoč. 2018. Interactive Video Search: Where is the User in the Age of Deep Learning?. In *Proceedings of the 26th ACM International Conference on Multimedia (MM '18)*. ACM, New York, NY, USA, 2101–2103. https://doi.org/10.1145/3240508.3241473

[14] Klaus Schoeffmann, Bernd Münzer, Andreas Leibetseder, Jürgen Primus, and Sabrina Kletz. 2019. Autopiloting Feature Maps: The Deep Interactive Video Exploration (diveXplore) System at VBS2019. In *MultiMedia Modeling - 25th International Conference, MMM 2019, Thessaloniki, Greece, January 8-11, 2019, Proceedings, Part II*. 585–590. https://doi.org/10.1007/978-3-030-05716-9_50

[15] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V. Le. 2017. Learning Transferable Architectures for Scalable Image Recognition. *CoRR* abs/1707.07012 (2017). arXiv:1707.07012 http://arxiv.org/abs/1707.07012