Exquisitor at the Video Browser Showdown 2022

Khan, O.S.; Sharma, U.; Jónsson, B.Þ.; Koelma, D.C.; Rudinac, S.; Worring, M.; Zahálka, J.

DOI
10.1007/978-3-030-98355-0_47

Publication date
2022

Document Version
Final published version

Published in
MultiMedia Modeling

Exquisitor at the Video Browser Showdown 2022

Omar Shahbaz Khan[1,2(✉)], Ujjwal Sharma[2], Björn Þór Jónsson[1], Dennis C. Koelma[2], Stevan Rudinac[2], Marcel Worring[2], and Jan Zahálka[3]

[1] IT University of Copenhagen, Copenhagen, Denmark
omsh@itu.dk
[2] University of Amsterdam, Amsterdam, Netherlands
[3] Czech Technical University in Prague, Prague, Czech Republic

Abstract. Exquisitor is the state-of-the-art large-scale interactive learning approach for media exploration that utilizes user relevance feedback at its core and is capable of interacting with collections containing more than 100M multimedia items at sub-second latency. In this work, we propose improvements to Exquisitor that include new features extracted at shot level for semantic concepts, scenes and actions. In addition, we introduce extensions to the video summary interface providing a better overview of the shots. Finally, we replace a simple keyword search featured in the previous versions of the system with a semantic search based on modern contextual representations.

Keywords: Interactive learning · Video browsing · Multimodal representation learning · Semantic search

1 Introduction

The Video Browser Showdown (VBS) is a live interactive video retrieval challenge, in which researchers participate with their retrieval tools to solve interactive tasks. VBS holds significant importance to researchers developing exploration and search tools for multimedia collections, as it is an opportunity to test their techniques in a realistic setting. Furthermore, the event leads to better insight in interactive retrieval, thus inspiring new methods, systems and further research [7]. This year's edition expands the video collection from 7,475 video clips (∼1,000 h) [14] to 17,235 video clips (∼2,300 h) [13].

Exquisitor is a prototype interactive learning system for large-scale media exploration, and has participated in the last two editions of VBS, where it performed adequately [3,4]. It uses user relevance feedback at its core, which builds a semantic classifier on the fly to find relevant items for the information need represented in the tasks. In addition to the single classifier, Exquisitor can build more classifiers and merge their result sets using various relational operators [4], with each classifier also having options for applying metadata filters. These multiple classifiers allow Exquisitor to deal with task descriptions that have a temporal nature, which are common in VBS. The need is evident by

B. Þór Jónsson et al. (Eds.): MMM 2022, LNCS 13142, pp. 511–517, 2022.
https://doi.org/10.1007/978-3-030-98355-0_47

looking at systems that generally do well in VBS, which incorporate modules for addressing temporal queries [6,8,10,15].

Previously, Exquisitor relied on video segmentation provided with the VBS collection and represented each segment with semantic features extracted from a corresponding keyframe. The keyframes are presented to the user during the interactive learning sessions, whose relevance judgment is used to build the semantic classifier(s) deployed for producing new suggestions. To determine whether the correct video segment has been found, the user can browse the video in Exquisitor's video summary view to fully compare the video content with the task description. As the provided segments are of arbitrary length, the features extracted from a single keyframe may not include information about the entire segment, especially since Exquisitor uses a compression scheme that reduces the number of features used to describe a keyframe [17]. The video summary interface of Exquisitor is also affected by the pre-defined shot structure, which makes it difficult to get an overview of the video, especially those with longer shots.

Additionally, the keyword search for finding initial positives, in its current form, is too restrictive, as it only uses available ImageNet concepts as search terms. These concepts do not always match the descriptions of tasks and make it difficult for the user to determine the right concepts.

In this paper, we introduce improvements to Exquisitor with additional features at shot level based on semantic concepts, actions and scenes. Furthermore, we conduct our own shot boundary detection that ensures shots are between 1–10 s. In addition, we introduce extensions to the video summary interface providing a better overview of the shots. Lastly, the keyword search used to find initial positive items for the user relevance feedback process, has been replaced with a semantic search based on modern contextual representations.

2 Exquisitor

Exquisitor is the state-of-the-art large-scale interactive learning approach capable of handling collections with over 100 million images [5,11]. The system consists of a web-based user interface and an interactive learning server.

Exquisitor Interface: Figure 1 shows the user interface of Exquisitor which allows the user to provide feedback to the systems suggestions based on the current classifier. Initially, arbitrary suggestions are presented. When a user hovers over a keyframe 4 buttons appear that allow, labeling it positive, negative, submitting the segment to the VBS evaluation server or ignoring it. The user can get new suggestions from the current model by pressing the "Next" button marking all current segments as viewed and at any point during the interactive session view these already seen segments through the "History" button. Additionally, at this screen the user can also apply metadata filters such as video level filters (categories and tags), and keyframe level filters (number of faces present, dominant color and amount of text present). In addition, the user can directly search for a specific video using its id, which opens the video summary view.

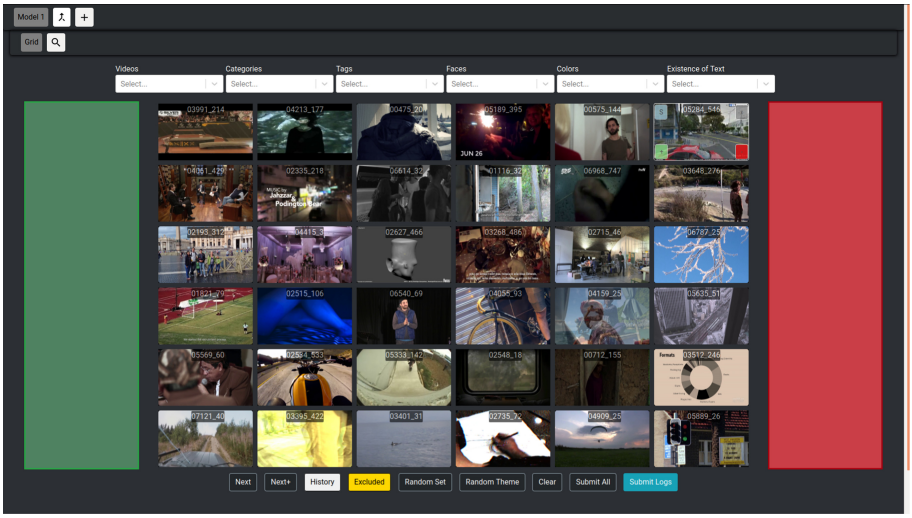

Fig. 1. Exquisitor's interface for building semantic classifiers.

At the top of the screen the user can add more classifiers, which opens a new similar screen to Fig. 1 in a new tab, to define multiple classifiers for different temporal concepts. The returned items from the classifiers can be merged using the merge view, which are ranked using classifier ranking operations [4]. The user can view the merge results to find the relevant segment or continue improving the classifiers. Depending on the collection size, the interface can directly retrieve the videos using local paths, but if space requirement is higher than the available local storage, a web-server is used to serve the videos instead.

Exquisitor Server: Exquisitor combines cluster-based indexing with an efficient compressed data representation containing the most important features extracted from each shot per modality [1,17]. Given Exquisitor's scalability, the expanded VBS dataset is of no concern to the core relevance feedback performance. The underlying model used for interactive multimodal learning in Exquisitor is a linear SVM, which is trained based on user relevance judgments on presented video segments. The hyperplane formed by the SVM is used for extracting k-farthest clusters from the index, as the intention is to present the user with the items for which the classifier is most confident. From this point a late fusion step in the form of rank aggregation is performed to get the top r suggestions to present the user with. Additionally, the server allows multiple users and handle keyword search requests.

Interactive Learning and VBS: There are three different task types presented at VBS: Textual and Visual Known-Item-Search (KIS), and Ad-hoc Video Search (AVS). The goal of the KIS task is to find a single video segment in one

specific video. The description is provided textually or through a visual snippet with audio. For the AVS task the goal is to find as many video segments matching the textual description in any video.

Within Exquisitor, semantic classifiers are built through user relevance feedback to bring forth the correct segment(s). The tasks have a time limit—KIS tasks end after 1 correct submission from the team, whereas AVS tasks allows submissions until the time has run out.

3 Improving Segment Representations

Exquisitor uses features extracted from the keyframes of video segments. Since video segments are of arbitrary size, more information is lost for longer segments as the keyframe may not fully cover it. Furthermore, as the information is based on entities, queries related to actions are difficult to represent with the classifier.

The length of provided video segments can be longer than 30 s but also shorter than 1 s. The longer segments may contain the relevant sequence which the representative keyframe is not highlighting [12]. The shorter segments cause an issue of segment overload, where no major change is happening in multiple continuous short segments, which would benefit from merging them together and improving the browsing experience.

Inspired by [16], we process the videos with our own video shot boundary detection method to provide less coarse segments, while still being slightly over segmented for addressing subtle changes. A post process step is performed on the segments that ensures a lower and upper bound of 1 s and 10 s respectively, to avoid excessively short or long segments. The first 7,475 videos in the collection corresponds to the previous years' collection which had 1,087,657 segments. With our approach this is reduced to 992,455 segments, a reduction of 8.7%. For the entire collection with 17,235 videos, this results in 2,285,514 segments. We extract semantic concepts [9], actions [2] and scenes [18] from the segments. These are treated as separate modalities within Exquisitor and are combined through rank aggregation during the retrieval process.

4 UI Extensions

Video Summary: The new shots are easier to view and process for a user. However, viewing is not always ideal as tasks are time dependent. Therefore, we now show up to 5 uniformly sampled thumbnails from the shot, so the user can rapidly determine whether it is relevant or not. Figure 2 shows the new video summary view. In addition to the shot thumbnails, the video player has been updated to show the entire video file instead of the shot file.

Keyword Search: The keyword search feature of Exquisitor is used to find initial positive examples for the user relevance feedback process. Originally it used a mapping of 12,988 assigned semantic concepts from ImageNet [9]. However,

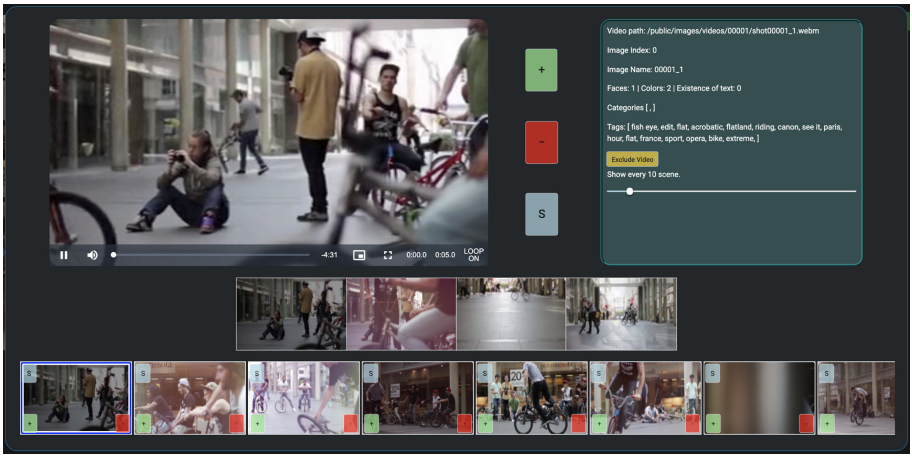

Fig. 2. The new video summary view with every shot having up to 5 thumbnails.

these concepts may not necessarily align with the user's vocabulary making it difficult to find the right concept. Restricting users to a limited set of search terms may result in users spending more time locating the appropriate search terms instead of actually searching. To avoid this, we replace the current keyword search with a semantic search that accepts natural language queries, using all available features, allowing users to search without being constrained by the limited set of search terms available from metadata.

5 Conclusions

This work presents a series of incremental improvements to Exquisitor. We extract features based on semantic concepts, actions and scenes at shot level. These features are used within Exquisitor as separate modalities and are combined through rank aggregation. We use our own using a shot boundary detection method that ensures they have an upper and lower bound on the duration. Furthermore, we replace the keyword search with semantic search allowing the use of natural language queries. These changes permit Exquisitor to more efficiently explore video collections while being transparent to the user.

Acknowledgments. This work was supported by a Ph.D. grant from the IT University of Copenhagen and by the European Regional Development Fund (project Robotics for Industry 4.0, CZ.02.1.01/0.0/0.0/15 003/0000470).

References

1. Guðmundsson, G.Þ., Jónsson, B.Þ., Amsaleg, L.: A large-scale performance study of cluster-based high-dimensional indexing. In: Proceedings of International Workshop on Very-large-scale Multimedia Corpus, Mining and Retrieval (VLS-MCM), Firenze, Italy (2010)
2. Hara, K., Kataoka, H., Satoh, Y.: Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet? In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6546–6555 (2018)
3. Jónsson, B.Þ, Khan, O.S., Koelma, D.C., Rudinac, S., Worring, M., Zahálka, J.: Exquisitor at the video browser showdown 2020. In: Ro, Y.M., et al. (eds.) MMM 2020. LNCS, vol. 11962, pp. 796–802. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-37734-2_72
4. Khan, O.S., et al.: Exquisitor at the video browser showdown 2021: relationships between semantic classifiers. In: Lokoč, J., et al. (eds.) MMM 2021. LNCS, vol. 12573, pp. 410–416. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-67835-7_37
5. Khan, O.S., et al.: Interactive learning for multimedia at large. In: Jose, J.M., et al. (eds.) ECIR 2020. LNCS, vol. 12035, pp. 495–510. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-45439-5_33
6. Kratochvíl, M., Veselý, P., Mejzlík, F., Lokoč, J.: SOM-Hunter: video browsing with relevance-to-SOM feedback loop. In: Ro, Y.M., et al. (eds.) MMM 2020. LNCS, vol. 11962, pp. 790–795. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-37734-2_71
7. Lokoč, J., et al.: Interactive search or sequential browsing? A detailed analysis of the video browser showdown 2018. ACM TOMM **15**(1), 1–18 (2019)
8. Lokoč, J., Kovalčík, G., Souček, T.: VIRET at video browser showdown 2020. In: Ro, Y.M., et al. (eds.) MMM 2020. LNCS, vol. 11962, pp. 784–789. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-37734-2_70
9. Mettes, P., Koelma, D.C., Snoek, C.G.: The ImageNet shuffle: reorganized pre-training for video event detection. In: Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval, ICMR 2016, pp. 175–182. Association for Computing Machinery, New York (2016)
10. Nguyen, P.A., Wu, J., Ngo, C.-W., Francis, D., Huet, B.: VIREO @ video browser showdown 2020. In: Ro, Y.M., et al. (eds.) MMM 2020. LNCS, vol. 11962, pp. 772–777. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-37734-2_68
11. Ragnarsdóttir, H., et al.: Exquisitor: breaking the interaction barrier for exploration of 100 million images. In: Proceedings of ACM Multimedia, Nice, France (2019)
12. Rossetto, L., Giangreco, I., Gasser, R., Schuldt, H.: Competitive video retrieval with vitrivr at the video browser showdown 2018-final notes. arXiv preprint arXiv:1805.02371 (2018)
13. Rossetto, L., Schoeffmann, K., Bernstein, A.: Insights on the V3C2 dataset. CoRR arXiv:2105.01475 (2021)
14. Rossetto, L., Schuldt, H., Awad, G., Butt, A.A.: V3C – a research video collection. In: Kompatsiaris, I., Huet, B., Mezaris, V., Gurrin, C., Cheng, W.-H., Vrochidis, S. (eds.) MMM 2019. LNCS, vol. 11295, pp. 349–360. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-05710-7_29

15. Sauter, L., Amiri Parian, M., Gasser, R., Heller, S., Rossetto, L., Schuldt, H.: Combining Boolean and multimedia retrieval in vitrivr for large-scale video search. In: Ro, Y.M., et al. (eds.) MMM 2020. LNCS, vol. 11962, pp. 760–765. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-37734-2_66
16. Yuan, J., et al.: Tsinghua University at TRECVID 2004: shot boundary detection and high-level feature extraction. In: TRECVID. Citeseer (2004)
17. Zahálka, J., Rudinac, S., Jónsson, B.Þ., Koelma, D.C., Worring, M.: Blackthorn: large-scale interactive multimodal learning. IEEE TMM **20**(3), 687–698 (2018)
18. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: a 10 million image database for scene recognition. IEEE Trans. Pattern Analy. Mach. Intell. **40**, 1452–1464 (2017)