Exquisitor at the Video Browser Showdown 2020

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

[1] IT University of Copenhagen, Denmark
[2] University of Amsterdam, Netherlands
[3] Czech Technical University in Prague, Czech Republic

Abstract. When browsing large video collections, human-in-the-loop systems are essential. The system should understand the semantic information need of the user and interactively help formulate queries to satisfy that information need based on data-driven methods. Full synergy between the interacting user and the system can only be obtained when the system learns from the user interactions while providing immediate response. Doing so with dynamically changing information needs for large scale multimodal collections is a challenging task. To push the boundary of current methods, we propose to apply the state of the art in interactive multimodal learning to the complex multimodal information needs posed by the Video Browser Showdown (VBS). To that end we adapt the Exquisitor system, a highly scalable interactive learning system. Exquisitor combines semantic features extracted from visual content and text to suggest relevant media items to the user, based on user relevance feedback on previously suggested items. In this paper, we briefly describe the Exquisitor system, and its first incarnation as a VBS entrant.

Keywords: Interactive learning · Video browsing · Scalability.

1 Introduction

The Video Browser Showdown (VBS) is a series of annual live competitions, where researchers are asked to study and develop methods to solve search-related tasks for a benchmark video collection. The VBS tasks, which are independent queries of three different flavours, are unknown to the researchers, who must prepare their systems and data representations for any potential task. At competition time, users of all systems are then given a few minutes to solve the tasks. Furthermore, depending on the task, the query may be gradually refined by adding information as time passes, to simulate real users with imperfect memories. While the systems taking part in previous VBS editions employ a variety of advanced search and retrieval techniques, a common observation is that they are highly interactive, requiring users to review and refine results of queries, resulting in a highly interactive process. Interactive multimodal learning has been proposed as an interactive method capable of satisfying users with uncertain information needs [15]. Given the format of VBS, it is of significant academic interest to apply interactive multimodal learning to VBS.

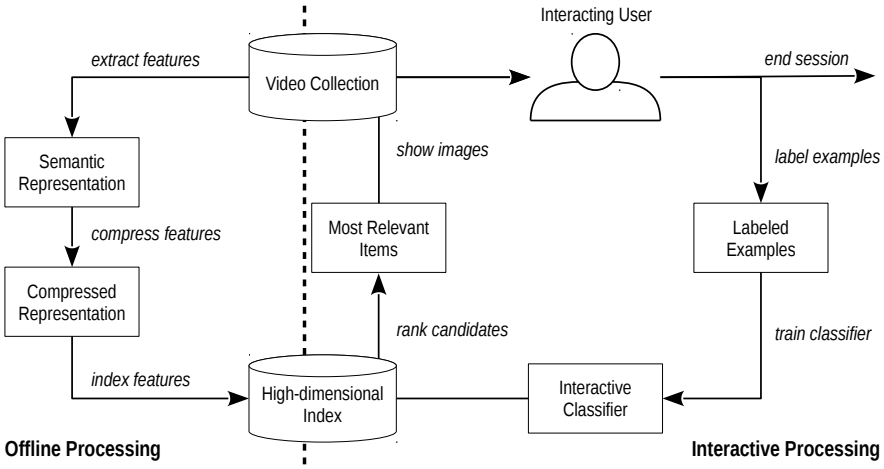

Fig. 1. Exquisitor's interactive learning pipeline. Initially, the video collection is processed to produce a compressed semantic representation, that is stored in a scalable high-dimensional index. In each round of the interactive learning process, the user is shown a set of potentially relevant videos. The user's judgments are then used to train a classifier, which in turn is used to retrieve a new set of videos to show to the user.

We have recently developed Exquisitor, a highly scalable interactive multimodal learning approach [5, 9]. Figure 1 illustrates the iterative feedback process employed by Exquisitor with video data. When a new task starts, the user is initially presented with a set of randomly selected video scenes from the collection and asked to give (positive or negative) feedback on (some of) the scenes. The feedback is used to build (and subsequently update) a classification model, which in turn is used to provide new suggestions; this iterative process continues as long as the user deems necessary. The Exquisitor system has been used to interactively explore the YFCC100M collection [9], and to compete in the Lifelog Search Challenge (LSC) 2019 [6], where it ranked 6th out of 9 competition entrants. A key feature that distinguishes Exquisitor from previous interactive learning systems is its scalability [5]; while the VBS video collection contains more than 1,000 hours of video, video suggestions can be retrieved in a fraction of a second in each interaction round. In this paper, we describe the adaptation of Exquisitor for participation in the Video Browser Showdown.

The remainder of the paper is organized as follows. In Sections 2 and 3, we briefly give background for interactive learning and the Video Browser Showdown, respectively. In Section 4, we then describe Exquisitor and its adaptation to VBS, before concluding in Section 5.

2 Interactive Learning

Interactive learning belongs to the family of human-in-the-loop learning approaches, eliciting data labels from the user and using that feedback to classify

the otherwise unannotated data on the fly. In contrast to supervised learning, no labels are required prior to the analysis. Interactive learning commonly uses a lightweight, fast classifier that learns online as the user inputs her feedback.

The two main learning strategies in interactive learning are active learning and user relevance feedback. The objective of *active learning* is to create the best classifier by eliciting labels on data most informative to the classifier, which often translates to the data points the classifier is the least confident about or those closest to the decision boundary [1, 4]. Conversely, *user relevance feedback* aims to satisfy the user, presenting items for which the classification model is the most confident [11]. While this latter strategy may require more interactions to achieve the same final quality of the classification model, users may obtain their desired insights earlier [15].

The increasing drive towards interactivity, personalized user experience, and higher-level semantic understanding, combined with recent advances in related scientific disciplines [12, 15, 16], have motivated us to re-visit user relevance feedback with our Exquisitor approach [5, 9].

3 The Video Browser Showdown

Involving users in the evaluation of retrieval processes has long been a challenge [7, 12, 14]. The majority of multimedia and computer vision benchmark competitions are held offline, allowing scientists to devote both significant computational power and time, which has helped solve difficult closed-world problems. Over the last two decades, however, international interactive search benchmarking events have emerged, where systems and their users must solve unknown and complex tasks within a limited time frame. From its inception in 2001, the TRECVID benchmark initiative included an interactive search task [14]. The VideOlympics [13] then started in 2008 and ran for five years, introducing the concept of live interactive video search benchmarking. The Video Browser Showdown (VBS) has been running since 2012 [8], and is now the premier live event, where participants must explore and search a collection of 1,000 hours of video [10]. A recent event series is the Lifelog Search Challenge (LSC), where a collection of lifelog image data must be explored [3]. While VBS and LSC represent only subsets of multimedia analytics applications, participation is important as it allows comparison with related state-of-the-art interactive systems.

The tasks in VBS have three different flavours. Visual Known-Item-Search (KIS) tasks present a randomly selected video clip to competitors, who must then identify the correct clip in the collection and submit it to the VBS server. Textual KIS tasks present a gradually evolving text description, which again has a specific matching scene in the collection. Finally, Ad-hoc Video Search (AVS) tasks ask for scenes matching a description; in this task judges evaluate the relevance of answers as they are submitted to the VBS server. The VBS competition has an expert session, where the teams use their own systems to solve all types of tasks, and a novice session, where conference participants, who have never seen the system, are asked to solve visual KIS and AVS tasks.

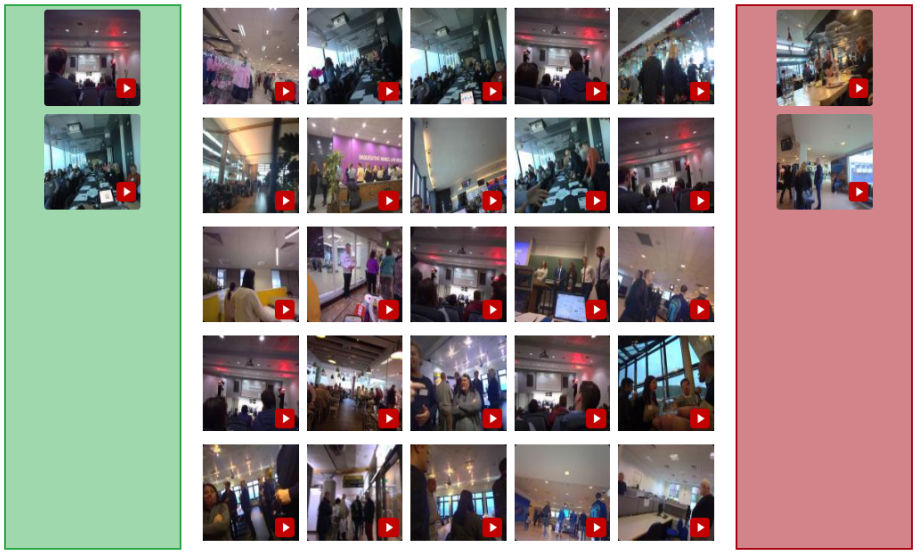

Fig. 2. Exquisitor's current user interface. The interface is browser-based and used primarily via mouse-based interaction. When hovering over a video, the user can choose to view the video in full, submit it to the VBS server, label it as a positive/negative example, or mark it as seen (using a 'next' button, the user can also mark all videos as seen and get a full screen of new videos). Positive (green column) and negative (red column) examples are immediately used to update the model.

4 Exquisitor

Exquisitor is a user relevance feedback approach capable of handling large scale collections in real time [5, 9]. The Exquisitor system used for VBS consists of three parts: (1) a web-based user interface for receiving and judging video suggestions; (2) an interactive learning server, which receives user judgments and produces a new round of suggestions; and (3) a web server which serves videos and video thumbnails. All three components run locally on the laptop of the VBS participants. In the following, we describe the first two parts of the system.

Exquisitor Interface: The current Exquisitor user interface is shown in Figure 2. In this initial incarnation, it is a pure interactive learning interface: the user is asked to label examples, which are subsequently used to learn the user's preference and suggest further examples. As the process to generate new suggestions is very efficient, however, new suggestions are retrieved each time the user identifies new positive or negative examples.

Exquisitor Server: Exquisitor has been developed to handle large-scale media collections, where each media item is described with feature vector data from

both visual and text modalities. The main components of the server are a) data representation and indexing, and b) the scoring process, described briefly below.

Each of the (just over) million scenes in the VBS collection is represented by a high-dimensional concept feature vector extracted from a selected keyframe. The high-dimensional feature vectors are compressed using an index-based compression method [16], where each feature vector is represented using the top 6 features of the modality and compressed into only three 64-bit integers. The compressed feature vectors are then indexed using the eCP high-dimensional indexing algorithm [2]. A set of representative vectors is chosen from the collection and each vector is assigned to the closest representative, thus forming clusters in the compressed high-dimensional space. To facilitate retrieval, the cluster representatives are recursively indexed to form an approximate cluster-based index.

Exquisitor uses a Linear SVM classifier learned from user interactions to score items in the compressed feature space. In each interaction round, the Linear SVM model yields a classification hyperplane, which is used to form a farthest neighbor query to the cluster-based index. The goal is to yield $k = 25$ suggestions, which can be presented to the user. The clusters farthest from the SVM hyperplane are selected and their contents scanned to yield the k furthest neighbors.

Solving VBS Tasks: In KIS tasks, the aim of positive and negative examples is to create a model that is good enough to bring the correct answer to the screen. If the user is satisfied that all videos displayed are neither useful as positive/negative examples nor the answer to the task, the user can use the 'next' button to continue browsing the results, similar to the typical 'query and browse' approach of many current VBS entrants. A submitted result is considered as a positive example, regardless of whether it is the correct result or not; once the correct result has been submitted the task is complete. For AVS tasks the process is identical, except that all videos on screen can be submitted at once using a special button, and the process only ends once time has expired.

5 Conclusions

This paper has outlined the adaptation of the Exquisitor system to the Video Browser Showdown, both in terms of the data used to represent the video collection and the interface changes made for video browsing. As a new entrant in the competition, our primary goal is to learn from our participation in the competition, aiming to understand both how well the interactive learning approach suits the different competition tasks, and how we can improve our preliminary interface to be better suited to the competitive environment.

Acknowledgments: This work was supported by a PhD 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. Cohn, D.A., Ghahramani, Z., Jordan, M.I.: Active learning with statistical models. J. Artificial Intelligence Research **4**(1), 129–145 (1996)
2. Guðmundsson, G.T., Jónsson, B.T., Amsaleg, L.: A large-scale performance study of cluster-based high-dimensional indexing. In: Proc. Int. Workshop on Very-large-scale Multimedia Corpus, Mining and Retrieval (VLS-MCM). Firenze, Italy (2010)
3. Gurrin, C., Schoeffmann, K., Joho, H., Dang-Nguyen, D., Riegler, M., Piras, L. (eds.): Proceedings of the 2018 ACM Workshop on The Lifelog Search Challenge, LSC@ICMR 2018. Yokohama, Japan (2018)
4. Huijser, M.W., van Gemert, J.C.: Active decision boundary annotation with deep generative models. In: Proc. IEEE ICCV. pp. 5296–5305. Venice, Italy (2017)
5. Jónsson, B.Þ., Khan, O.S., Ragnarsdóttir, H., Þorleiksdóttir, Þ., Zahálka, J., Rudinac, S., Guðmundsson, G.Þ., Amsaleg, L., Worring, M.: Exquisitor: Interactive learning at large. arXiv:1904.08689 (2019)
6. Khan, O.S., Jónsson, B.Þ., Zahálka, J., Rudinac, S., Worring, M.: Exquisitor at the Lifelog Search Challenge 2019. In: Proceedings of the ACM Workshop on Lifelog Search Challenge, LSC@ICMR. pp. 7–11. Ottawa, ON, Canada (2019)
7. Larson, M., Arora, P., Demarty, C., Riegler, M., Bischke, B., Dellandréa, E., Lux, M., Porter, A., Jones, G.J.F. (eds.): Working Notes Proceedings of the MediaEval 2018 Workshop, CEUR Workshop Proceedings, vol. 2283. CEUR-WS.org, Sophia Antipolis, France (2018)
8. Lokoč, J., Bailer, W., Schoeffmann, K., Münzer, B., Awad, G.: On influential trends in interactive video retrieval: Video Browser Showdown 2015-2017. IEEE Trans. Multimedia **20**(12), 3361–3376 (2018)
9. Ragnarsdóttir, H., Þorleiksdóttir, Þ., Khan, O.S., Jónsson, B.Þ., Guðmundsson, G.Þ., Zahálka, J., Rudinac, S., Amsaleg, L., Worring, M.: Exquisitor: Breaking the interaction barrier for exploration of 100 million images. In: Proceedings of the ACM Multimedia Conference. Nice, France (2019)
10. Rossetto, L., Schuldt, H., Awad, G., Butt, A.A.: V3C - A research video collection. In: Proc. MultiMedia Modeling (MMM). pp. 349–360. Thessaloniki, Greece (2019)
11. Rui, Y., Huang, T.S., Mehrotra, S.: Content-based image retrieval with relevance feedback in MARS. In: Proc. ICIP. pp. 815–818. Santa Barbara, CA, USA (1997)
12. Schoeffmann, K., Bailer, W., Gurrin, C., Awad, G., Lokoč, J.: Interactive video search: Where is the user in the age of deep learning? In: Proc. ACM Multimedia. pp. 2101–2103. Seoul, Republic of Korea (2018)
13. Snoek, C.G.M., Worring, M., de Rooij, O., van de Sande, K.E.A., Yan, R., Hauptmann, A.G.: Videolympics: Real-time evaluation of multimedia retrieval systems. IEEE MultiMedia **15**(1), 86–91 (2008)
14. Thornley, C., Johnson, A.C., Smeaton, A.F., Lee, H.: The scholarly impact of TRECVID (2003-2009). Journal of the American Society for Information Science and Technology (JASIST) **62**(4), 613–627 (2011)
15. Zahálka, J., Worring, M.: Towards interactive, intelligent, and integrated multimedia analytics. In: Proc. of the IEEE Conference on Visual Analytics Science and Technology (VAST). pp. 3–12. Paris, France (2014)
16. Zahálka, J., Rudinac, S., Jónsson, B.T., Koelma, D.C., Worring, M.: Blackthorn: Large-scale interactive multimodal learning. IEEE Transactions on Multimedia **20**(3), 687–698 (2018)