A System for Interactive Multimedia Retrieval Evaluations

Luca Rossetto[2][0000−0002−5389−9465], Ralph Gasser[1][0000−0002−3016−1396], Loris Sauter[1][0000−0001−8046−0362], Abraham Bernstein[2][0000−0002−0128−4602], and Heiko Schuldt[1][0000−0001−9865−6371]

[1] Department of Mathematics and Computer Science
University of Basel, Basel, Switzerland
{firstname.lastname}@unibas.ch
[2] Department of Informatics, University of Zurich, Zurich, Switzerland
{lastname}@ifi.uzh.ch

Abstract. The evaluation of the performance of interactive multimedia retrieval systems is a methodologically non-trivial endeavour and requires specialized infrastructure. Current evaluation campaigns have so far relied on a local setting, where all retrieval systems needed to be evaluated at the same physical location at the same time. This constraint does not only complicate the organization and coordination but also limits the number of systems which can reasonably be evaluated within a set time frame. Travel restrictions might further limit the possibility for such evaluations. To address these problems, evaluations need to be conducted in a (geographically) distributed setting, which was so far not possible due to the lack of supporting infrastructure. In this paper, we present the *Distributed Retrieval Evaluation Server (DRES)*, an open-source evaluation system to facilitate evaluation campaigns for interactive multimedia retrieval systems in both traditional on-site as well as fully distributed settings which has already proven effective in a competitive evaluation.

Keywords: Interactive Multimedia Retrieval, Retrieval Evaluation

1 Introduction

Due to the continuous growth of multimedia collections in terms of their size and diversity, multimedia retrieval has evolved to a major discipline in the general field of "Big Data" research. Tools and techniques to efficiently store, manage, and search such data corpora have become more important, and a lot of research effort went into exploring techniques to extract features from media data, to store and manage large quantities of such data, and to efficiently index it so as to facilitate fast access even for collections beyond billions of entries [3,4,8].

Despite all these efforts, however, it has been shown repeatedly [7,9] that the task of finding a particular item in a large enough collection still is an interactive task that requires cooperation between a human actor and a system. This results

in the more general setting of *interactive retrieval*, in which users leverage end-to-end retrieval systems to explore media collections and to satisfy a particular information need, by refining queries and browsing through result sets.

Evaluating the performance of such systems is a far more difficult and complex undertaking than evaluating the algorithms used by them. Firstly, the human operator plays an important role in the overall combined human-system performance since the translation of an information need into a query is an inherently manual problem. Secondly, the task itself is more complex as its solution requires a sequence of steps involving a combination of techniques. One way of handling this complexity is by conducting evaluation campaigns such as the *Video Browser Showdown (VBS)* [12] for videos or the *Lifelog Search Challenge (LSC)* [5] for multimodal lifelog data. In both campaigns, teams from around the world gather once per year to compare their retrieval systems in a series of tasks. Each task formulates a particular information need, e.g., by depicting an example or by describing the desired object. The teams then have a predefined amount of time to find the item in question and to submit it to the *evaluation server*. Finding the correct item quickly is rewarded with a higher score, whereas wrong submissions or taking a lot of time are penalized. This setting incentivizes participants to continuously refine their systems in all aspects. It can be attributed to the success of such campaigns, that the evaluation setting has changed and adapted over the years. As systems become better, tasks need to become more challenging. With the increasingly complex techniques employed in systems, it is also no longer sufficient to simply rank these by their performance during an evaluation. Instead, one has to collect sufficient data so as to be able to explain why any one system performed better than another for a certain type of task, which requires specialized logging infrastructure.

The contribution of this paper is a demo of the *Distributed Retrieval Evaluation Server (DRES)*[3] – a modular and extendable open source system that generalizes not only the aforementioned, interactive evaluation setting conceptually but also enables a user to setup and hold various retrieval evaluations. DRES comes with a standardized API for logging, which can be used to collect metrics regarding the performance of individual systems. Since evaluating interactive retrieval systems in an on-site setting may not always be feasible, DRES is explicitly designed to support such evaluations in a distributed setting, where participants can reside in different locations. DRES has already been successfully used in multiple distributed retrieval evaluations outside of the larger international campaigns and is scheduled to replace the previously used *VBS Server*[4] from LSC 2020 and VBS 2021 onward. Its flexible architecture also enables its use in other retrieval evaluation campaigns.

The remainder of this paper is structured as follows: Section 2 briefly surveys related work. Section 3 introduces some of the concepts, gives a system overview and motivates some of the design decisions behind DRES. Finally, Section 4 provides some conclusion and outlook on future work.

[3] `https://dres.dev/`
[4] `https://github.com/klschoef/vbsserver/`

2 Related Work

Evaluating multimedia retrieval solutions in a competitive, challenge-focused setting has been an established practice for many years. The first such evaluation campaigns — such as the TREC Video Track [13] which later turned into TRECVID [1] or ImageCLEF [2], both established in 2001 and 2003 respectively — were set up as non-interactive evaluations. More evaluation campaigns in the multimedia domain have been started over the years – many of them in the context of the MediaEval benchmarking initiative, which has been active since 2008. However, none of these challenges have so far been evaluated interactively.

An early example for an interactive retrieval evaluation campaign was VideOlympics [14] from 2008, which had tasks similar to TRECVID's Ad-Hoc Video Search but took place live in front of an audience. The Video Browser Showdown (VBS) [12] campaign was started in 2012 [11] and has since been held annually in conjunction with the International Conference on MultiMedia Modelling, making it the longest running interactive multimedia retrieval campaign to date, relying on an ever increasing video collection [10]. The tasks evaluated during VBS have undergone some changes over the years. As of 2020, there were three types of tasks: (1) a *Visual Known-Item Search* (Visual KIS) task, where participants have to find an unique video segment of roughly 20 seconds in length from within a pre-defined dataset, (2) a *Textual Known-Item Search* (Textual KIS) task where an unique video segment must be found based on a precise textual description, and (3) an *Ad-Hoc Video Search* (AVS) task, where participants are required to find as many video segments as possible that match a rough textual description. This last task type is similar to the challenge posed by VideOlympics, but employs human judges rather than a pre-determined ground truth, since increasing dataset sizes made exhaustive pre-labelling of the data impractical. All task types are solved by *experts*, which are usually the developers of the retrieval systems. Visual KIS and AVS tasks are additionally solved by *novices* who are selected from the conference audience and have no prior experience with any particular system, in order to assess the usability of the retrieval systems for non-specialists.

Inspired by the VBS, the Lifelog Search Challenge (LSC) started in 2018 [6] as a workshop at the ACM International Conference on Multimedia Retrieval, where it is since held annually [5]. The challenge is similar to the Textual KIS task, but uses lifelog data consisting of image sequences as a retrieval target, which were captured by a wearable camera and annotated with various meta-information.

3 DRES: System Overview

The following describes the inner workings of the system, its architecture and interaction models as well as certain considerations made during its design in order to support present and future requirements.

3.1 Capabilities

DRES is designed to meet the requirements posed by interactive retrieval campaigns both presently and in the foreseeable future and to be easily extendable should new requirements arise. An *evaluation* in DRES consists of multiple *retrieval tasks*. Each task is based on a defined *media collection*, which can contain any type of media such as images, videos or audio. In order to support a wide range of evaluation settings, *retrieval tasks* can be configured by an evaluation coordinator to meet any given need. Such tasks are composed of a set of *hints* and a set of *targets*. The *target* can either be predefined as one or many media objects, a temporal range within a media object, or not specified at all. In the latter case, submissions are forwarded to a judgement mechanism, to have the correctness of a submitted result determined by an external (human) judgement. The *hints* are presented to the participants during a task and they can consist of text, images, or videos as well as any combination thereof. Hints can be arranged on a timeline such that the displayed information changes over time.

The flexible data model of DRES enables evaluation coordinators to build tasks of various types, such as the aforementioned *Textual KIS* type. Such a task's hints are textual and typically, they come with three hints each starting 0, 60 and 120 seconds into the task, where each hint is replaced by the next one and the last hint remains active until the end of the task. The target of such a task is then simply a temporal range within a video item.

Further configurations enable evaluation coordinators to specify if a submission preview is to be shown while the task is still running or if a submission needs to specify a temporal range in addition to a media item. The modular and flexible design would enable evaluation coordinators to expand upon the currently known Textual KIS task by, e.g. adding a doodle of the description for the second half of the task duration as a visual aid. A user management component keeps track of all participants of an evaluation and ensures that all submitted solutions to a task are attributed to the correct team and member. A dedicated component collects interaction- as well as result-logs submitted by the participating systems [9] which can be used to gain additional insights into the system behaviours and search strategies.

3.2 Architecture

Architecturally, DRES can be divided into two primary components: a back-end and a front-end. The back-end manages the *evaluation configurations* and the individual *evaluations*, i.e., instances of a specific configuration, as well as the required multimedia collections and user data. It communicates with the users through an interactive shell as well as a RESTful API, which adheres to the OpenAPI standard. The RESTful API is primarily used by the front-end, however, there is a public API that can be used to submit results and report system metrics such as user/system interactions and excerpts of query results.

The front-end runs in a web-browser in order to minimize software requirements on the user's side. The primary purpose of the front-end is to present the

tasks to the participants during an active evaluation and to let the organizers manage the sequence of tasks. It also provides functionality for evaluation and task configuration as well as the management of users and media collections.

The back-end is primarily structured around a component called *RunExecutor*, which coordinates an arbitrary number of *RunManager*s that in turn are responsible for an individual evaluation. Each RunManager is initialized with an evaluation configuration that specifies the tasks of the evaluation and information about the participants trying to solve them. This configuration serves as a sort of template for the individual instances (i.e., a run) of an actual evaluation, which are then managed by a RunManager. The system currently supports *synchronous* runs, meaning all participants get the same task at the same time. *Asynchronous* runs where participants solve the same tasks but not necessarily all at the same time are planned for future expansions.

3.3 Demonstration

During the demonstration, visitors will not only be able to see the participant facing side of the system, which they might already have seen during an evaluation such as VBS or LSC, but be able to experience the entire workflow from setup of an evaluation, configuration of tasks and running of the campaign itself.

4 Conclusion and Outlook

In this paper, we introduced DRES, the Distributed Retrieval Evaluation Server, an open-source system that can be used to setup and host evaluation sessions for interactive multimedia retrieval solutions for both on-site as well as distributed settings. Its flexible data model and modular architecture enables it to support all types of evaluation tasks currently in use by established evaluation campaigns such as VBS and LSC as well as further constellations, which might become relevant in the future. The support for distributed evaluation settings opens up new avenues to advance the state of interactive multimedia retrieval by eliminating the spatial restrictions as well as reducing the organizational and financial overhead of holding an evaluation in one common location. First experiences gathered with distributed evaluations are promising and we expect that such settings will become a powerful augmentation to the established, localized ones. Currently, DRES supports *synchronous* evaluations, where all participants solve the same task at the same time. In future versions, we aim to also support *asynchronous* evaluations, where participants can solve the same tasks independently of one other hence offering spatial as well as temporal distribution. This would further reduce the burden placed on participants, especially in larger distributed settings, which might stretch across multiple time zones.

Acknowledgements

This work was partly supported by the Hasler Foundation in the context of the project City-Stories (contract no. 17055).

References

1. Awad, G., Butt, A., Curtis, K., Lee, Y., Fiscus, J., Godil, A., Delgado, A., Zhang, J., Godard, E., Diduch, L., Smeaton, A.F., Graham, Y., Kraaij, W., Quénot, G.: Trecvid 2019: An evaluation campaign to benchmark video activity detection, video captioning and matching, and video search & retrieval. In: Proceedings of TRECVID 2019. NIST, USA (2019)
2. Clough, P., Sanderson, M.: The clef 2003 cross language image retrieval track. In: Workshop of the Cross-Language Evaluation Forum for European Languages. pp. 581–593. Springer (2003)
3. Gasser, R., Rossetto, L., Heller, S., Schuldt, H.: Cottontail DB: An Open Source Database System for Multimedia Retrieval and Analysis. In: Proceedings of the 28th ACM International Conference on Multimedia (MM '20). ACM, Seattle, WA, USA (Oct 2020)
4. Giangreco, I., Schuldt, H.: Adam pro: Database support for big multimedia retrieval. Datenbank-Spektrum **16**(1), 17–26 (2016)
5. Gurrin, C., Le, T.K., Ninh, V.T., Dang-Nguyen, D.T., Jónsson, B.Þ., Lokoš, J., Hürst, W., Tran, M.T., Schoeffmann, K.: Introduction to the third annual lifelog search challenge (lsc'20). In: Proceedings of the 2020 International Conference on Multimedia Retrieval. pp. 584–585 (2020)
6. Gurrin, C., Schoeffmann, K., Joho, H., Leibetseder, A., Zhou, L., Duane, A., Dang-Nguyen, D.T., Riegler, M., Piras, L., Tran, M.T., et al.: [invited papers] comparing approaches to interactive lifelog search at the lifelog search challenge (lsc2018). ITE Transactions on Media Technology and Applications **7**(2), 46–59 (2019)
7. Lokoč, J., Kovalčík, G., Münzer, B., Schöffmann, K., Bailer, W., Gasser, R., Vrochidis, S., Nguyen, P.A., Rujikietgumjorn, S., Barthel, K.U.: Interactive search or sequential browsing? a detailed analysis of the video browser showdown 2018. ACM Trans. Multimedia Comput. Commun. Appl. **15**(1) (Feb 2019)
8. Pouyanfar, S., Yang, Y., Chen, S.C., Shyu, M.L., Iyengar, S.S.: Multimedia big data analytics: A survey. ACM Computing Surveys **51**(1) (Jan 2018)
9. Rossetto, L., Gasser, R., Lokoc, J., Bailer, W., Schoeffmann, K., Muenzer, B., Soucek, T., Nguyen, P.A., Bolettieri, P., Leibetseder, A., et al.: Interactive video retrieval in the age of deep learning - detailed evaluation of vbs 2019. IEEE Transactions on Multimedia pp. 1–1 (2020)
10. Rossetto, L., Schuldt, H., Awad, G., Butt, A.A.: V3c–a research video collection. In: International Conference on Multimedia Modeling. pp. 349–360. Springer (2019)
11. Schoeffmann, K.: A user-centric media retrieval competition: The video browser showdown 2012-2014. IEEE MultiMedia **21**(4), 8–13 (2014)
12. Schoeffmann, K.: Video browser showdown 2012-2019: A review. In: 2019 International Conference on Content-Based Multimedia Indexing (CBMI). pp. 1–4. IEEE (2019)
13. Smeaton, A.F., Over, P., Taban, R.: The trec-2001 video track report. In: TREC (2001)
14. Snoek, C.G., Worring, M., de Rooij, O., van de Sande, K.E., Yan, R., Hauptmann, A.G.: Videolympics: Real-time evaluation of multimedia retrieval systems. IEEE MultiMedia **15**(1), 86–91 (2008)