Grand Challenge Proposal
2nd CASTLE Multimodal Analytics Challenge

Werner Bailer, Cathal Gurrin, Björn Þór Jónsson, Duc-Tien Dang-Nguyen, Luca Rossetto, Klaus Schoeffmann, Allie (Ly-Duyen) Tran

Motivation

The increasing availability of mobile computing devices equipped with a variety of sensors enables us to capture and quantify more and more aspects of the human condition. Automatically drawing insights from such captured data has the potential to open doors to a wide range of new applications, but this remains challenging, in particular when information needs to be combined from different points along the timeline.
The CASTLE challenge aims to act as a catalyst in the development of methods for multimodal understanding by providing a rich multimodal dataset that serves as a basis for a range of analysis tasks. Research on lifelog retrieval and analysis has so far focused on longitudinal data of a single person, or multi-person data over a short time range. The CASTLE challenge scales the problem to multi-person and multi-day data, aiming to model real-world problem settings more closely, and advancing the state of the art in multimodal understanding of human activity video data.

Dataset

The CASTLE dataset consists of time-aligned data streams, captured by multiple different sensors, including both stationary and head-mounted high-resolution cameras and microphones. The data was captured during four days by 11 participants in a quiet and secluded location. There are a total of 15 different video streams, with 10 egocentric and 5 static perspectives.
The final dataset consists of over 700 time-aligned 1-hour videos, starting on the hour and covering the time range from 08:00 to 21:00. The videos are recorded in UHD@50fps and include audio. They are accompanied by data streams from several sensors, including 6DoF IMU, GPS, and biometric data. A preview of the first 15 minutes of the dataset is available via https://www.youtube.com/watch?v=9Y9dXNHvIhk.

Challenge tasks

For its second edition, the CASTLE challenge will feature a variety of tasks, including object and event detection, retrieval, and question answering. Future instances of the challenge might expand on the range of task types. For the second edition, we will keep the following task types, introduced in the first edition:

- Event instance Search: given a textual description (in English), identify all timeframes where a specific event occurs. Events are to be identified by time range and video id.
- Object Instance Search: given a textual (in English) or visual (i.e., using an image) example of a physical object, find all occurrences of that object in any of the video streams (without the use of screen capture).
- Question Answering: given a question formulated in natural language (in English), find the answer to the question. Answers are to be provided in natural language and include references to sensor streams and time intervals to provide evidence.

Evaluation

The challenge will operate across two tracks: fully automatic and interactive (teams can participate in one of them, or both). For the fully-automatic track, participants receive queries in advance and produce their results by whatever process they see fit. Results will subsequently be submitted to the challenge organizers for evaluation.
In contrast, the interactive track will be evaluated in a live hybrid (both on-site and online) event during the conference, where participants have to solve tasks synchronously and interactively within a narrow time frame. This track will use a setting analogous to the Video Browser Showdown or the Lifelog Search Challenge. Should no room be available at the conference venue to host the on-site challenge participants, the interactive evaluation will be held exclusively online.
This dual-track setup combines two well-established challenge formats. Their combination is considerably rarer, but not without precedent. The closest analogue in setup is the IViSE[1] challenge, held at CVPR 2025. The evaluation mechanisms that will be used in the CASTLE challenge are analogous to those used by IViSE, which in turn bases its procedure on established campaigns such as TRECVID for the fully-automatic and the Video Browser Showdown and the Lifelog Search Challenge for the interactive track.

[1] see https://sites.google.com/view/ivise2025

Relevant technologies

The challenge builds on existing technologies for visual lifelog retrieval (see e.g. the summary paper of the Lifelog Search Challenge at ICMR[2], previous editions or related analysis papers[3]), for interactive video retrieval, such as the technologies used in the Video Browser Showdown,[4] as well as the wide area of research on automatic video retrieval (as benchmarked by TRECVID) and multimedia question answering and related tasks[5].
Datasets for question answering on “long” videos still address durations of several minutes,[6] while this challenge will require finding specific evidence or answering questions in a dataset of 15 parallel streams spanning over 4 days. This will require significant improvements in terms of scalability of the methods being used, pushing the limits of content embedding, indexing and retrieval methods.
In addition, the automatic track is still challenging with state-of-the-art language models, as complex contexts need to be understood to identify relevant information and extract factoids needed for question answering. The interactive track enables offloading these tasks requiring semantic understanding and reasoning to a human, but in turn poses research challenges of designing a UI that enables the user to identify, select and assess relevant information in hundreds of hours of video that have all been shot at the same location and are visually homogeneous.

Publicity and Continuation

We are proposing the second iteration of the CASTLE challenge, which centers around the CASTLE 2024 dataset. The CASTLE 2024 dataset[7] was recorded in 2024 and publicly released in 2025 as part of the ACM Multimedia Dataset Track. This dataset will serve as a proving ground for novel questions on multimodal human interaction data. The challenge is designed to reduce the barriers to entry as much as possible by using established query types already known from different campaigns. This only scratches the surface of what is feasible with this type of data. In future iterations of this challenge, we intend to expand the tasks to encompass more of the unique properties afforded by the dataset. In addition, initial design efforts for an expanded dataset based on the learnings of the creation of the first one are already underway.
We are committed to ensuring the long-term availability of the dataset, and will maintain an archive of challenge descriptions and tasks.

---

[2] Gurrin, Cathal, et al. "Introduction to the 8th Annual Lifelog Search Challenge, LSC'25." Proceedings of the 2025 International Conference on Multimedia Retrieval. 2025.
[3] Tran, Ly-Duyen, et al. "The State-of-the-Art in Lifelog Retrieval: A Review of Progress at the ACM Lifelog Search Challenge Workshop 2022-2024." IEEE Access 13 (2025): 216340-216363.
[4] Vadicamo, Lucia, et al. "Evaluating Performance and Trends in Interactive Video Retrieval: Insights from the 12th VBS Competition." *IEEE Access* (2024).
[5] Wang, Jiaqi, et al. "A Comprehensive Review of Multimodal Large Language Models: Performance and Challenges Across Different Tasks." *arXiv preprint arXiv:2408.01319* (2024).
[6] Rawal, Ruchit, et al. "CinePile: A Long Video Question Answering Dataset and Benchmark." *Synthetic Data for Computer Vision Workshop@ CVPR 2024*.
[7] Rossetto, Luca, et al. "The CASTLE 2024 dataset: Advancing the Art of Multimodal Understanding." *Proceedings of the 33rd ACM International Conference on Multimedia*. 2025.

Organizers

Werner Bailer (werner.bailer@joanneum.at)

Werner Bailer is a Key Researcher at JOANNEUM RESEARCH's Intelligent Vision Applications group (Graz, Austria), with research interests in video content analysis, processing and retrieval, machine learning, and modeling of metadata of audiovisual content. He serves regularly as a reviewer for journals (e.g., IEEE TIP, IEEE TCSVT, and ACM ToMM) and conferences (e.g., ACM MM, IEEE ICME, ACM ICMR, MMM). Since 2012, he is co-organizer of the Video Browser Showdown interactive video search challenge and has recently co-chaired workshops at ACM TVX, ICML, and ICCV.

Cathal Gurrin (cathal.gurrin@dcu.ie)

Cathal Gurrin is a full Professor at the School of Computing, at Dublin City University and deputy director of the national Adapt centre for Digital Content Technologies. Gurrin leads a group of researchers dedicated to developing assistive technologies using wearable sensors and data analytics. He is the founder and co-organiser of the Lifelog Search Challenge at ICMR, the NTCIR-Lifelog participation workshop and the ImageCLEF lifelog task. He has been the general chair of many high-ranking computer science conferences, such as ECIR 2011 & MMM 2014, ACM ICMR 2020 and he has been the general co-chair of MB2016 & MMM2017, CBMI 2019, MMM 2022, ECIR 2023, ACM ICMR 2024, ACM Multimedia 2025 and the upcoming ACM Web 2027 conference.

Björn Þór Jónsson (bjorn@ru.is)

Björn Þór Jónsson is a full Professor in the Department of Computer Science at Reykjavik University, Iceland. His research work focuses primarily on the performance and scalability of content based multimedia analytics and retrieval. He has recently served as general chair for the SISAP 2025, MMM 2022, ICMR 2020 conferences, as technical program chair for CBMI 2024 and MMM 2024, and as co-organiser of interactive retrieval workshops and competitions, among other organising roles.

Duc-Tien Dang-Nguyen (ductien.dangnguyen@uib.no)

Duc-Tien Dang-Nguyen is a professor at the University of Bergen. His main area of expertise is on lifelogging, multimedia forensics, multimedia verification, and multimedia retrieval. Dang-Nguyen is co-organiser of the Lifelog Search Challenge, the NTCIR Lifelog Task and the ImageClef Lifelog tasks. He is general co-chair of MMM 2023, CBMI 2025; TPC of MMM 2022, ICMR 2024, ACM MM 2025, and incoming MMM 2026.

Luca Rossetto (luca.rossetto@dcu.ie)

Luca Rossetto is an Assistant Professor at the School of Computing at Dublin City University. His research focuses on managing, analyzing, and retrieving multi-modal data. He is one of the core developers of the open-source multimedia retrieval engine 'vitrivr' and

co-creator of the ‘Distributed Retrieval Evaluation Server’ used for interactive multimedia evaluations in different areas. Luca is a member of ACM and SIGMM and a regular reviewer for international multimedia conferences such as ACM MM, ACM MM Asia, ACM ICMR, MMM, and others, as well as journals including IEEE Transactions on Multimedia, Multimedia Systems, and Multimedia Tools and Applications.

Klaus Schoeffmann (klaus.schoeffmann@aau.at)

Klaus Schoeffmann is an Associate Professor at the Institute of Information Technology (ITEC) at Klagenfurt University, Austria. His research focuses on video content understanding (including medical/surgery videos), deep learning, computer vision, multimedia retrieval, and interactive multimedia. He is co-founder and co-organizer of the annual Video Browser Showdown (VBS) and the annual Lifelog Search Challenge (LSC). He is a member of the IEEE and the ACM, a regular reviewer for international conferences and journals in the field of multimedia and medical imaging. Klaus Schoeffmann has been the program co-chair of MMM 2021, CBMI 2021, ACM ICMR 2020, MMM2018, CMBI 2013, the demo & video co-chair of ACMMM2020, open-source software competition chair of ACMMM2019, the general co-chair of MMM2012, a general co-chair of ACM ICMR 2024, CBMI2025, and ACMMM2025, and a program co-chair of ACM MMAsia 2025.

Allie (Ly-Duyen) Tran (allie.tran@dcu.ie)

Allie (Ly-Duyen) Tran is a postgraduate researcher at the School of Computing at Dublin City University, specializing in question answering over personal data with a focus on multimodal data understanding, exploration, and interaction. An active member of the lifelog research community, she has contributed to several international benchmarking activities and regularly reviews for conferences such as ACM ICMR, MMM, and CBMI, as well as the journal Multimedia Tools and Applications. She is also involved in organizing major multimedia and retrieval conferences, serving as Interactive Arts Chair for ACM-MM 2025, Local Chair for CBMI 2025 and ECIR 2023, as well as co-organizing NTCIR-Lifelog Task 2025 and AIQAM 2024 (ACM-ICMR).

Timeline

- 09.03.2026: Challenge announcement, publication of website
- 09.03.2026: Registration Opens
- 09.03.2026: Query release
- 11.06.2026: Fully-automated Solution / Paper submission deadline
- 30.07.2026: Notification to Authors
- 06.08.2026: Camera-ready deadline