Overview of the First CASTLE Grand Challenge at ACM Multimedia 2025

Luca Rossetto
Dublin City University
Dublin, Ireland
luca.rossetto@dcu.ie

Werner Bailer
JOANNEUM RESEARCH
Graz, Austria
werner.bailer@joanneum.at

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

Duc-Tien Dang-Nguyen
University of Bergen
Bergen, Norway
ductien.dangnguyen@uib.no

Klaus Schöffmann
Klagenfurt University
Klagenfurt, Austria
klaus.schoeffmann@aau.at

Allie Tran
Dublin City University
Dublin, Ireland
allie.tran@dcu.ie

Abstract

The inaugural edition of the CASTLE grand challenge was held at ACM Multimedia 2025. The focus of the CASTLE challenge is to advance the state-of-the-art in analysis and understanding of multimodal data, especially centered around multistream ego- and exo-centric video. In this first instance of the challenge, participants had to solve three types of tasks – event instance search, object instance search, and question answering – in a fully automatic or interactive setting.

CCS Concepts

• **Information systems → Evaluation of retrieval results**; **Multimedia and multimodal retrieval**; *Question answering*; *Users and interactive retrieval*.

Keywords

Egocentric Video, Interactive Video Understanding, Multimodal Data Engineering, Video Question Answering

ACM Reference Format:
Luca Rossetto, Werner Bailer, Cathal Gurrin, Duc-Tien Dang-Nguyen, Klaus Schöffmann, and Allie Tran. 2025. Overview of the First CASTLE Grand Challenge at ACM Multimedia 2025. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25), October 27–31, 2025, Dublin, Ireland*. ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3746027.3760242

1 Introduction

With the rise of both wearable smart devices, such as smart glasses and augmented / mixed-reality head-sets, as well as the increasing developments in the area of humanoid robotics, egocentric first-person video enjoys increasing attention. Several datasets have been published in recent years, centered on egocentric video, recorded in a broad range of settings and contexts. A recent addition among those is the CASTLE 2024 Dataset [7], which contains more than 600 hours of high-resolution high-framerate video recorded over four days from ten egocentric and five exocentric perspectives at a common location. Based on this dataset, the inaugural edition of the CASTLE grand challenge, held at ACM Multimedia 2025, offers several tasks, evaluated in different settings, with the aim of advancing the current state-of-the-art in analysis and understanding of ego- and exocentric video.

This paper provides an overview of the first edition of the challenge. It is structured as follows: Section 2 provides an outline of the structure of the challenge and its tasks. Section 3 then gives a brief summary of the participants and their approaches. Finally, Section 4 offers some concluding remarks.

2 The Challenge

The challenge format draws inspiration from the Video Browser Showdown (VBS) [14] and the Lifelog Search Challenge (LSC) [10], as well as the TRECVID [1] campaigns. It features three types of tasks:

Event Instance Search Given a textual description (in English), participants must identify all time frames where a specific event occurs.

Object Instance Search Given a textual (in English) or visual (i.e., using an image) example of a physical object, participants must find all occurrences of that object across any of the video streams.

Question Answering Given a question in natural language (in English), participants must provide an answer. The response should be formulated in natural language and include references to relevant sensor streams and time intervals as supporting evidence.

Similarly to the IViSE [6] challenge, held initially at CVPR 2025, the 2025 CASTLE challenge features two different tracks for evaluation: fully automatic and interactive. For the fully automatic track, all queries are published beforehand and participants are able to use any means at their disposal to solve the tasks and submit their results to the challenge organizers.

In contrast, the interactive track, inspired by benchmarks like the VBS and the LSC, requires participants to solve queries within a time limit of a few minutes in a controlled setting. It is performed with the help of the Distributed Retrieval Evaluation Server (DRES) [8], which masters the presentation of tasks and the collection and evaluation of submissions. This interactive evaluation is conducted in a live session at ACM Multimedia 2025.

MM '25, Dublin, Ireland

ACM ISBN 979-8-4007-2035-2/2025/10
https://doi.org/10.1145/3746027.3760242

With these three initial task types, evaluated in two different scenarios, we already cover a large part of the space of relevant task configurations [4]. Future iterations of the challenge might introduce additional task types to further introduce coverage.

For the automatic track, the scoring is based on the fraction of correct entries in the submission file. The correct set of results is an open set of the object/event instance search tasks, and the submissions are manually evaluated. In contrast, there is a defined correct answer for question answering tasks, but in order to be robust to various ways of expressing the result, the answer is also manually checked. In the interactive track, question answering submissions are judged in real-time by human judges, and the score depends on the time of submission, using the KIS scoring function employed by VBS [2]. The evaluation of the instance search tasks follows the evaluation of AVS tasks in VBS, which takes into account both the ratio of correct and incorrect submissions (precision), as well as the ratio of correct submissions per team with respect to all correct submissions across all teams (recall). In contrast to VBS, no pre-defined shot segmentation or segment aggregation are used.

3 Participants in 2025

The initial instance of the challenge attracted three participating teams, one of which withdrew their paper shortly before publication. Of the remaining two teams, one [3] participates only in the interactive track of the challenge, while the other one [12] participates in both tracks. Both teams base their approaches on systems that have previously successfully participated in LSC or VBS. Specifically, [3] presents an extension of Exquisitor [9] which previously successfully participated in both LSC and VBS, whereas [12] uses methods from multiple LSC systems[11, 13] for their automatic submissions and an extension of Eagle [5], also a previous LSC participant, for their interactive part.

4 Conclusion

Despite a much larger expression of interest, the initial CASTLE challenge received only three submissions, one of which was withdrawn at the last moment. The discrepancy between the initial expression of interest and the final set of participants, both of whom have prior experience in interactive video and lifelog retrieval leads us to believe that the barrier of entry for the challenge might have been too high and that the tasks might have been too difficult for new teams. For future instances of the challenge, we will endeavor to lower the barrier of entry to give more teams an opportunity to participate.

Acknowledgments

This work was partially funded by the Swiss National Science Foundation through Project "MediaGraph" (Grant Number 202125) and Research Ireland under grant number 13/RC/2106_P2 at the ADAPT SFI Research Centre at Dublin City University.

References

[1] George Awad, Jonathan Fiscus, Afzal Godil, Lukas Diduch, Yvette Graham, and Georges Quénot. 2024. TRECVID 2024 - Evaluating video search, captioning, and activity recognition. In *Proceedings of TRECVID 2024*. NIST, USA.

[2] Silvan Heller, Viktor Gsteiger, Werner Bailer, Cathal Gurrin, Björn Þór Jónsson, Jakub Lokoc, Andreas Leibetseder, Frantisek Mejzlík, Ladislav Peska, Luca Rossetto, Konstantin Schall, Klaus Schoeffmann, Heiko Schuldt, Florian Spiess, Ly-Duyen Tran, Lucia Vadicamo, Patrik Veselý, Stefanos Vrochidis, and Jiaxin Wu. 2022. Interactive video retrieval evaluation at a distance: comparing sixteen interactive video search systems in a remote setting at the 10th Video Browser Showdown. *International Journal of Multimedia Information Retrieval* 11, 1 (2022), 1–18. doi:10.1007/s13735-021-00225-2

[3] Omar Shahbaz Khan, Ujjwal Sharma, Gonçalo Marcelino, Aaron Duane, Stevan Rudinac, Marcel Worring, and Björn Þór Jónsson. 2025. An Interactive Retrieval System for Multi-Stream Collections: multiXview at CASTLE 2025 Interactive Grand Challenge. In *Proceedings of the 33rd ACM International Conference on Multimedia* (Dublin, Ireland). Association for Computing Machinery, New York, NY, USA.

[4] Jakub Lokoc, Werner Bailer, Kai Uwe Barthel, Cathal Gurrin, Silvan Heller, Björn Þór Jónsson, Ladislav Peska, Luca Rossetto, Klaus Schoeffmann, Lucia Vadicamo, Stefanos Vrochidis, and Jiaxin Wu. 2022. A Task Category Space for User-Centric Comparative Multimedia Search Evaluations. In *MultiMedia Modeling - 28th International Conference, MMM 2022, Phu Quoc, Vietnam, June 6-10, 2022, Proceedings, Part I (Lecture Notes in Computer Science, Vol. 13141)*, Björn Þór Jónsson, Cathal Gurrin, Minh-Triet Tran, Duc-Tien Dang-Nguyen, Anita Min-Chun Hu, Huynh Thi Thanh Binh, and Benoit Huet (Eds.). Springer, 193–204. doi:10.1007/978-3-030-98358-1_16

[5] Thang-Long Nguyen-Ho, Onanong Kongmeesub, Minh-Triet Tran, Dongyun Nie, Graham Healy, and Cathal Gurrin. 2024. Eagle: Eyegaze-assisted guidance and learning evaluation for lifeloging retrieval. In *Proceedings of the 7th Annual ACM Workshop on the Lifelog Search Challenge*. 18–23.

[6] Luca Rossetto, George Awad, Werner Bailer, Cathal Gurrin, Björn Þór Jónsson, Jakub Lokoc, Stevan Rudinac, and Klaus Schoeffmann. 2025. Overview of the 1st International Workshop on Interactive Video Search and Exploration. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2025, Nashville, TN, USA, June 11-15, 2025*. Computer Vision Foundation / IEEE, 3673–3678. https://openaccess.thecvf.com/content/CVPR2025W/IViSE/html/Rossetto_Overview_of_the_1st_International_Workshop_on_Interactive_Video_Search_CVPRW_2025_paper.html

[7] Luca Rossetto, Werner Bailer, Duc-Tien Dang-Nguyen, Graham Healy, Björn Þór Jónsson, Onanong Kongmeesub, Hoang-Bao Le, Stevan Rudinac, Klaus Schöffmann, Florian Spiess, Allie Tran, Minh-Triet Tran, Quang-Linh Tran, and Cathal Gurrin. 2025. The CASTLE 2024 Dataset: Advancing the Art of Multimodal Understanding. *CoRR* abs/2503.17116 (2025). arXiv:2503.17116 doi:10.48550/ARXIV.2503.17116

[8] Loris Sauter, Ralph Gasser, Heiko Schuldt, Abraham Bernstein, and Luca Rossetto. 2025. Performance Evaluation in Multimedia Retrieval. *ACM Trans. Multim. Comput. Commun. Appl.* 21, 1 (2025), 23:1–23:23. doi:10.1145/3678881

[9] Ujjwal Sharma, Omar Shahbaz Khan, Stevan Rudinac, and Björn Þór Jónsson. 2025. Exquisitor at the Video Browser Showdown 2025: Unifying Conversational Search and User Relevance Feedback. In *MultiMedia Modeling - 31st International Conference on Multimedia Modeling, MMM 2025, Nara, Japan, January 8-10, 2025, Proceedings, Part V (Lecture Notes in Computer Science, Vol. 15524)*, Ichiro Ide, Ioannis Kompatsiaris, Changsheng Xu, Keiji Yanai, Wei-Ta Chu, Naoko Nitta, Michael Riegler, and Toshihiko Yamasaki (Eds.). Springer, 264–271. doi:10.1007/978-981-96-2074-6_31

[10] Allie Tran, Werner Bailer, Duc-Tien Dang-Nguyen, Graham Healy, Steve Hodges, Björn Þór Jónsson, Luca Rossetto, Klaus Schoeffmann, Minh-Triet Tran, Lucia Vadicamo, and Cathal Gurrin. 2025. The State-of-the-Art in Lifelog Retrieval: A Review of Progress at the ACM Lifelog Search Challenge Workshop 2022-24. *CoRR* abs/2506.06743 (2025). arXiv:2506.06743 doi:10.48550/ARXIV.2506.06743

[11] Ly-Duyen Tran, Nguyen Thanh Binh, Cathal Gurrin, and Liting Zhou. 2024. MyEachtraX: Lifelog Question Answering on Mobile. In *Proceedings of the 7th Annual ACM Workshop on the Lifelog Search Challenge, LSC 2024, Phuket, Thailand, 10 June 2024*. ACM, 93–98. doi:10.1145/3643489.3661128

[12] Quang-Linh Tran, Hoang-Bao Le, Thang-Long Nguyen-Ho, Graham Healy, Liting Zhou, and Allie Tran. 2025. Extending Lifelog Retrieval to Multi-stream Video Retrieval at the CASTLE Challenge 2025. In *Proceedings of the 33rd ACM International Conference on Multimedia* (Dublin, Ireland). Association for Computing Machinery, New York, NY, USA.

[13] Quang-Linh Tran, Binh Nguyen, Gareth JF Jones, and Cathal Gurrin. 2024. MemoriEase 2.0: A Conversational Lifelog Retrieve System for LSC'24. In *Proceedings of the 7th Annual ACM Workshop on the Lifelog Search Challenge*. 12–17.

[14] Lucia Vadicamo, Rahel Arnold, Werner Bailer, Fabio Carrara, Cathal Gurrin, Nico Hezel, Xinghan Li, Jakub Lokoc, Sebastian Lubos, Zhixin Ma, Nicola Messina, Thao-Nhu Nguyen, Ladislav Peska, Luca Rossetto, Loris Sauter, Klaus Schöffmann, Florian Spiess, Minh-Triet Tran, and Stefanos Vrochidis. 2024. Evaluating Performance and Trends in Interactive Video Retrieval: Insights From the 12th VBS Competition. *IEEE Access* 12 (2024), 79342–79366. doi:10.1109/ACCESS.2024.3405638