Multimedia Retrieval in Mixed Reality: Leveraging Live Queries for Immersive Experiences

Rahel Arnold, Heiko Schuldt — University of Basel, Switzerland

rahel.arnold@unibas.ch, heiko.schuldt@unibas.ch

Contribution

We merge Mixed Reality (MR) and multimedia retrieval in our $(MR)^2$ framework. Combining MR, object detection, and multimedia retrieval, $(MR)^2$ enhances user interactions in MR environments, allowing seamless searches through the innovative live query feature.

Mixed Reality

Mixed Reality (MR) technology combines real-world elements with digital content in real-time to create an immersive environment where physical and virtual objects interact. The experience is facilitated through headsets with transparent glasses or cameras and smartphones.

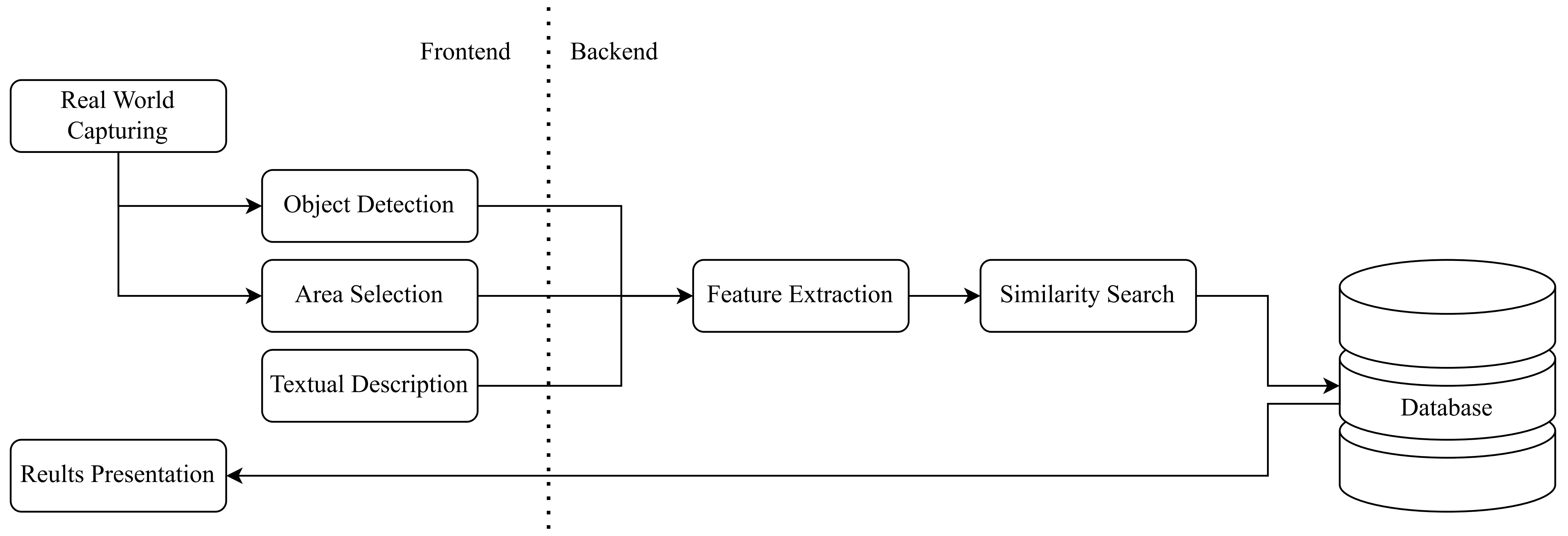
Multimedia Retrieval

Multimedia retrieval involves searching and retrieving multimedia content from large databases. It is used in image and video search engines, content-based recommendation systems or archives to retrieve documents most similar to a given query.

Mixed Reality Multimedia Retrieval Framework $(MR)^2$

$(MR)^2$ blends natural and virtual worlds, prioritising immersion, real-time interaction, user-centric design, and the integration of cutting-edge machine learning models.

The frontend is crafted for capturing user interactions and performing computations on the device. It supports three query modalities: object detection, area selection and text input. Furthermore, it provides the presentation of query results. The backend handles query data, processes inputs with a machine learning model, and conducts similarity searches on the extracted feature vectors.



Object Detection

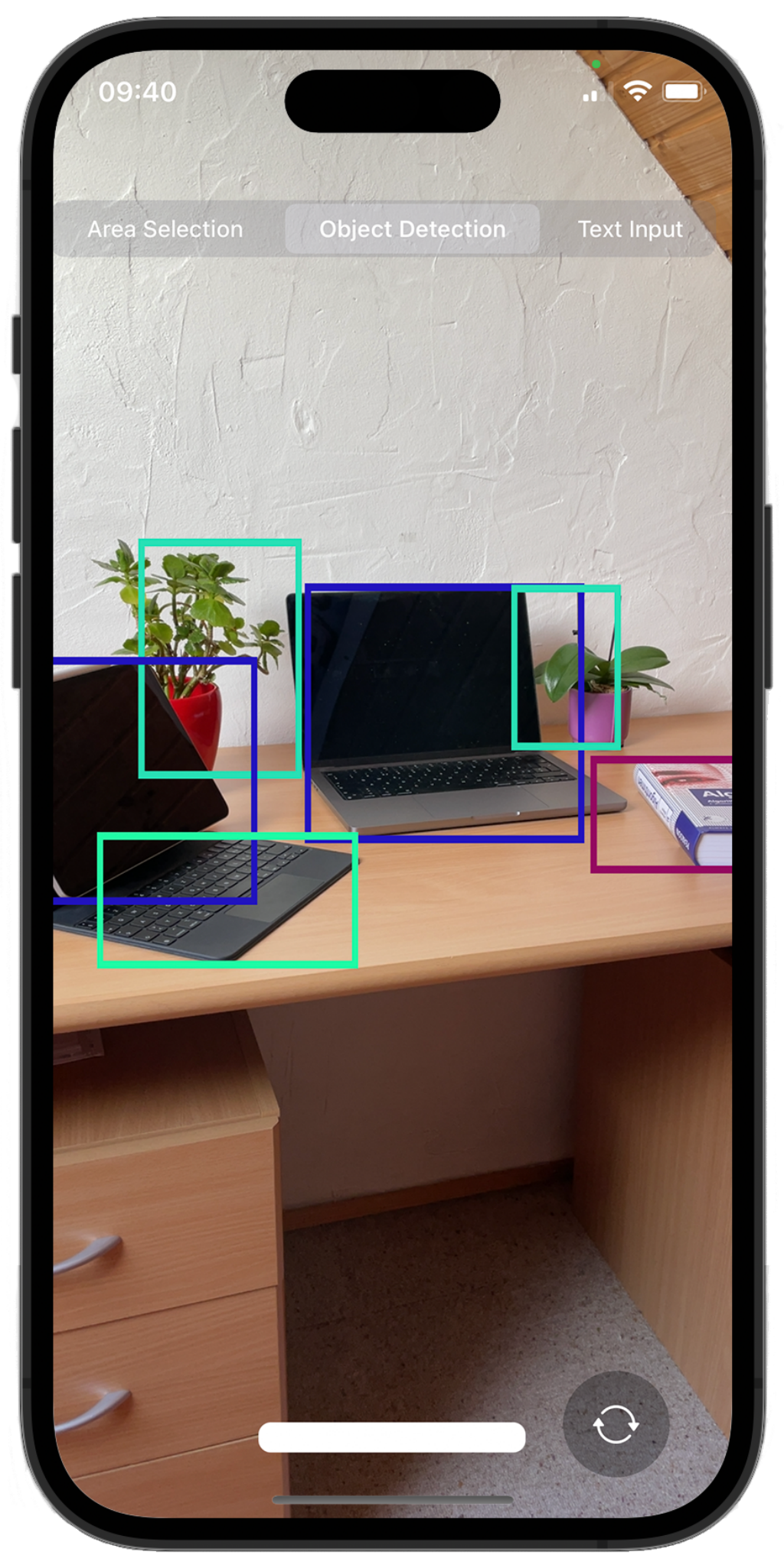
Specialised object detection is crucial in MR environments. YOLO [1] efficiently predicts objects with a grid system, bounding boxes, and advanced neural networks. Leveraging artificial intelligence (AI), YOLOv8, developed by Ultzralytics, enhances its ability to detect objects quickly and accurately, making it a valuable technology in various applications.

Visual-Text Co-Embedding

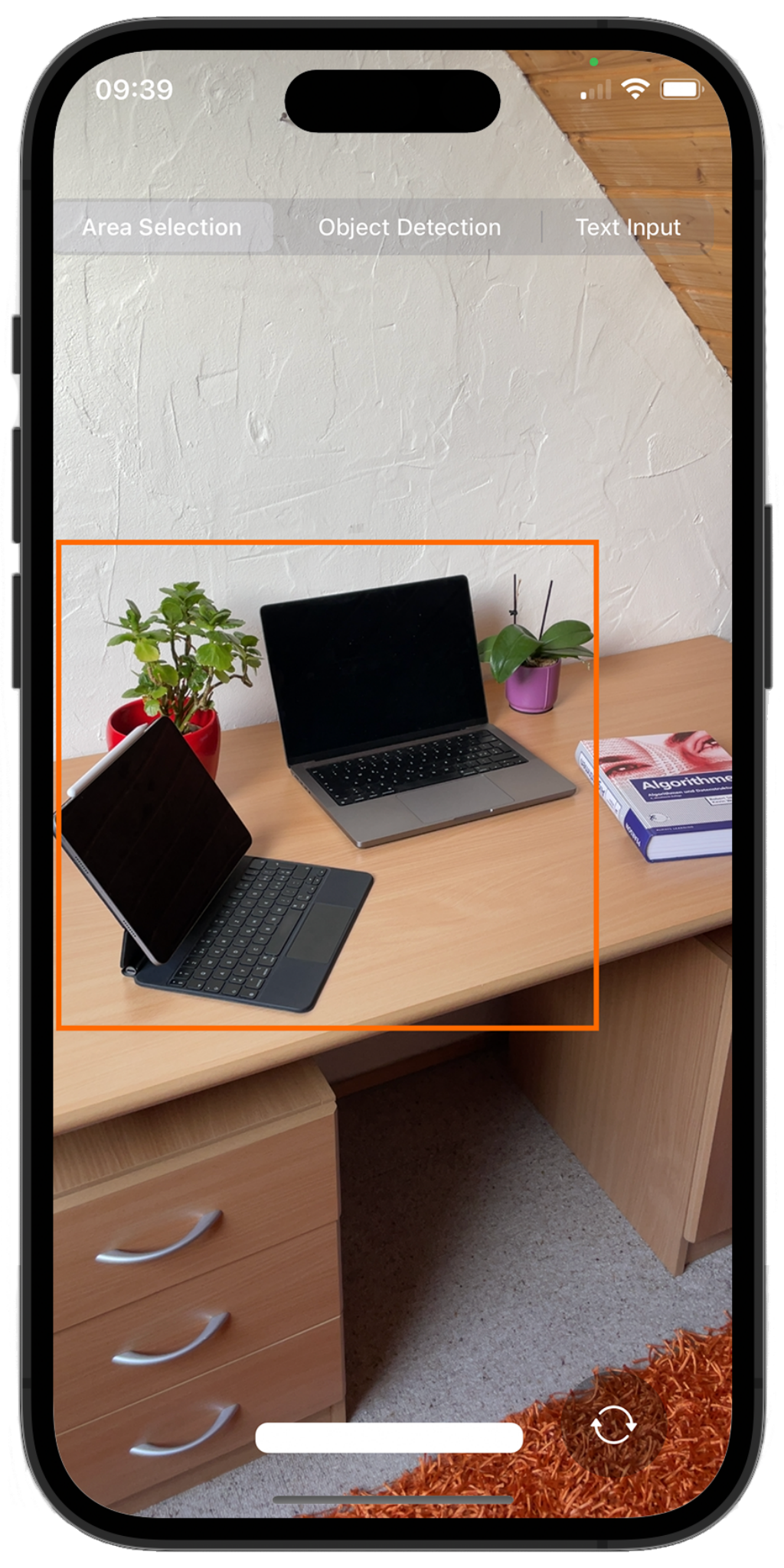
Contrastive Language-Image Pretraining (CLIP) [2], a breakthrough in visual-text co-embedding and a form of artificial intelligence, enhances multimedia retrieval by creating a joint space for direct text and image comparison. CLIP integrates textual metadata with visual content using an image and text encoder, enabling robust and context-aware retrieval.

Implementation

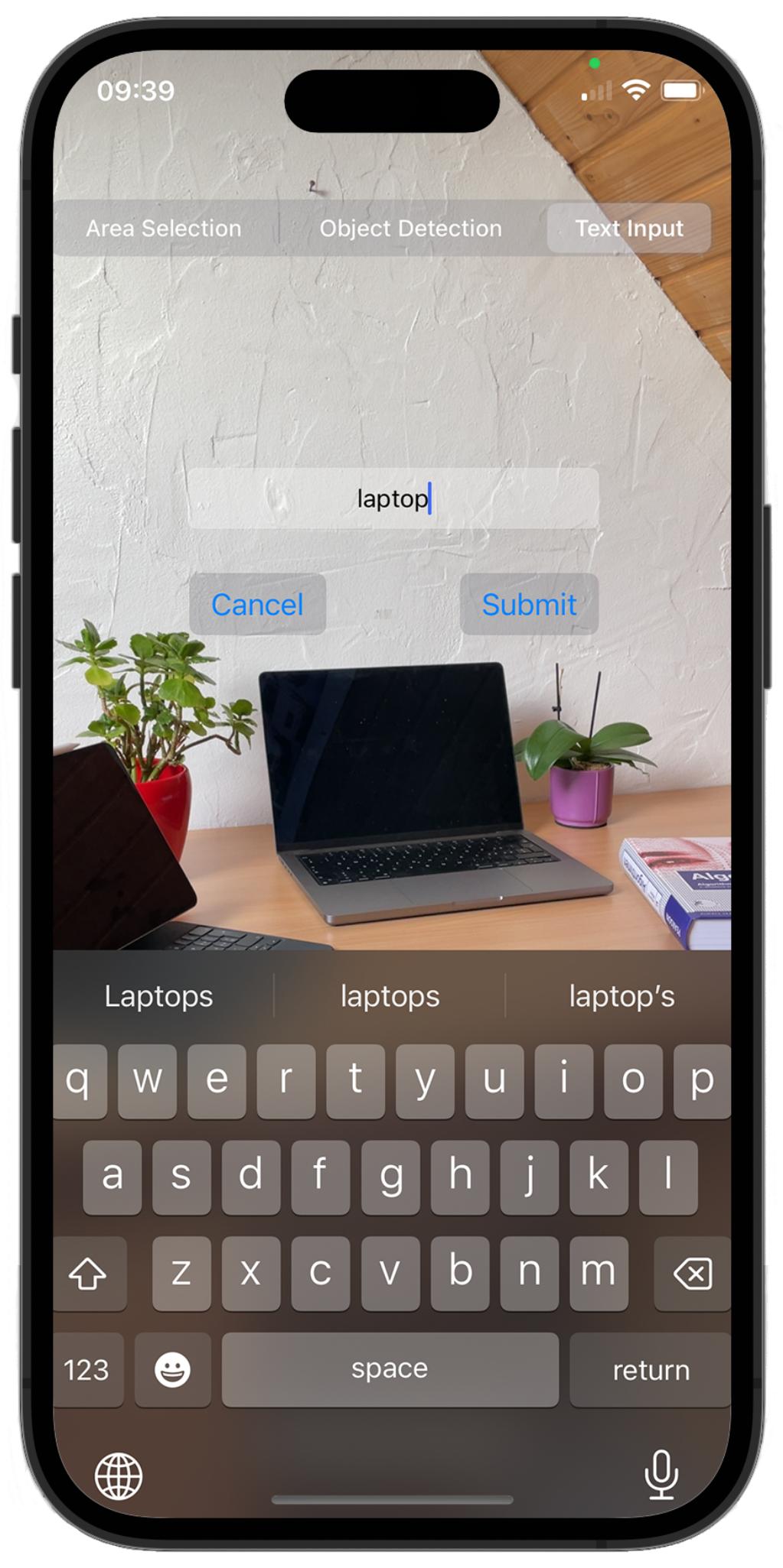
The $(MR)^2$ prototype implementation, an iOS application designed for iPhones and iPads, incorporates AI technologies. It features Swift-based MR interactions and employs YOLOv8, an AI-driven object detection model, for enhanced visual perception. The Python server manages CLIP feature extraction, leveraging AI capabilities, and utilises Cottontail DB [3] for similarity searches. The AVFoundation seamlessly integrates the camera feed, providing a responsive interface for real-world interactions. Search results are presented in a scrollable grid within a dedicated ViewController.



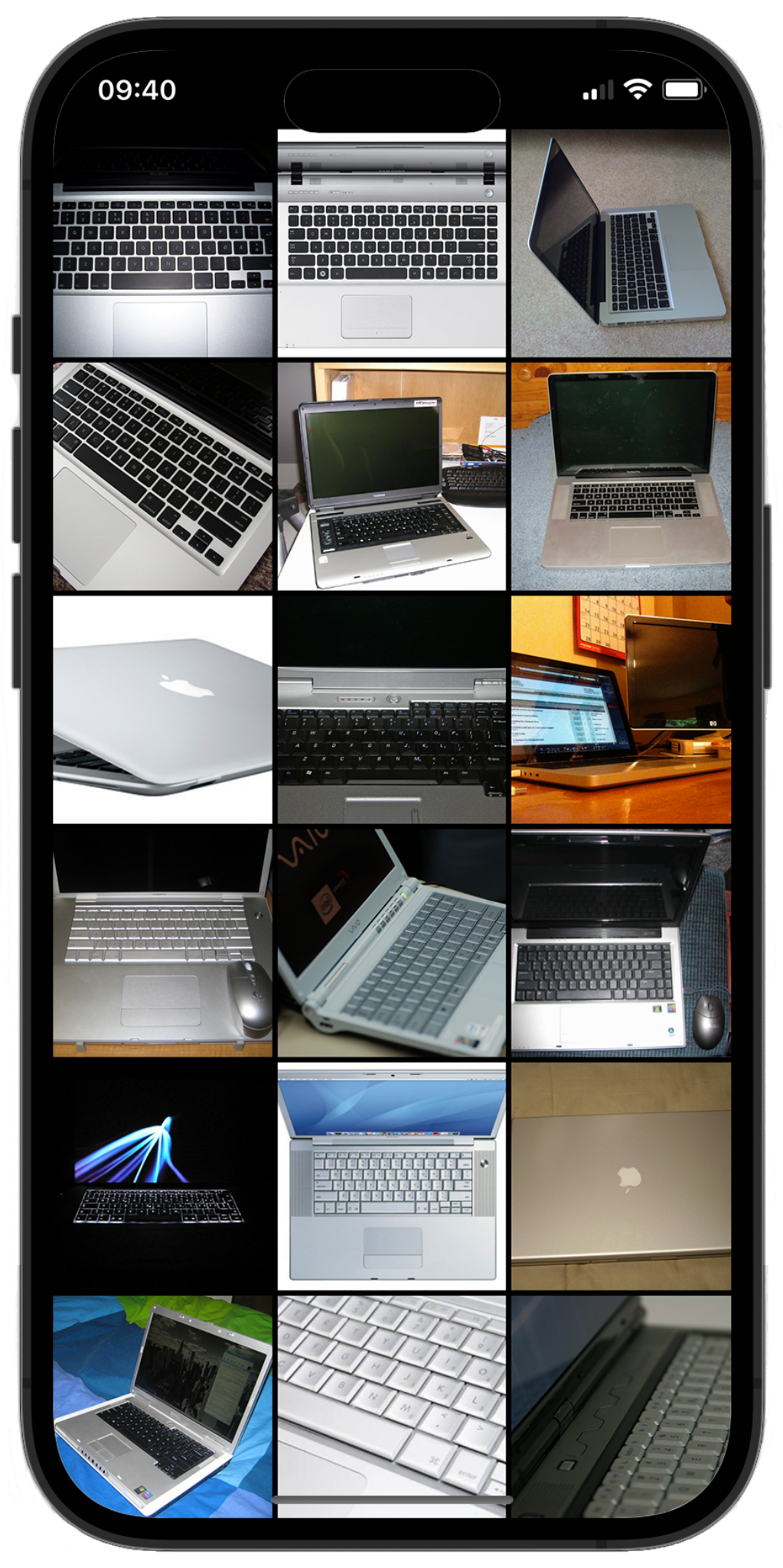
Object Detection



Manual Area Selection



Text Input



Result Presentation

Evaluation

Performance analysis focused on object detection and query response times. With a median inference time of 24.8 ms for 10,000 detections, $(MR)^2$ ensures real-time applicability. The average query time (4191 ms for 100 measurements) demonstrates efficiency, especially with the extensive ImageNet dataset in Cottontail DB.

In the user evaluation with 14 participants, emphasising practical usability, those with moderate to high technology affinity found $(MR)^2$ user-friendly and efficient (average ATI score of 4.43). The System Usability Scale (SUS) score of 87 reinforced positive perceptions.

Future Research

Proposed directions for advancing $(MR)^2$ involve integrating Optical Character Recognition (OCR) and Automatic Speech Recognition (ASR). Additionally, introducing temporal queries, developing immersive result presentations in MR, exploring compatibility with MR glasses, and emphasising advancements in object detection models and self-training approaches are crucial aspects of this research.

References

[1] J. Redmon, S. K. Divvala, R. B. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” *CoRR*, vol. abs/1506.02640, 2015. [Online]. Available: http://arxiv.org/abs/1506.02640

[2] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, “Learning transferable visual models from natural language supervision,” *CoRR*, vol. abs/2103.00020, 2021. [Online]. Available: https://arxiv.org/abs/2103.00020

[3] R. Gasser, L. Rossetto, S. Heller, and H. Schuldt, “Cottontail DB: An open source database system for multimedia retrieval and analysis,” in *Proceedings of the 28th ACM International Conference on Multimedia*, Oct. 2020.

Acknowledgement

This work was supported by the Horizon Europe project XR mEdia eCOsystem (XReco), based on a grant from the Swiss State Secretariat for Education, Research and Innovation (SERI), contract no. 22.00268.