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

Rahel Arnold*, Heiko Schuldt†

Department of Mathematics and Computer Science
University of Basel, Basel, Switzerland
*† {firstname.lastname}@unibas.ch
*0000-0002-5881-4432 †0000-0001-9865-6371

***Abstract*—Recent advancements in Mixed Reality (MR) technology and the exponential growth of multimedia data production have led to the emergence of innovative approaches for efficient content retrieval. This paper introduces Mixed Reality Multimedia Retrieval (*(MR)*^{2}), a groundbreaking concept at the convergence of MR and multimedia retrieval. At its core, *(MR)*^{2} leverages MR's transformative capabilities with an innovative live query option, allowing users to initiate queries intuitively through real-world object interactions. By autonomously generating queries based on object recognition in the user's field of view, *(MR)*^{2} facilitates the retrieval of similar multimedia content from a connected database. The technical backbone of the *(MR)*^{2} framework includes object detection (YOLOv8), semantic similarity search (CLIP), and data management (Cottontail DB). Our research redefines user interactions with multimedia databases, seamlessly bridging the physical and digital domains. A successful iOS prototype application demonstrates promising results, paving the way for immersive and context-aware multimedia retrieval in the MR era.**

***Index Terms*—Mixed Reality, Multimedia Retrieval, Object Detection, Similarity Search, Visual-Text Co-Embedding, User Experience.**

I. Introduction

The rapid evolution of technology has opened up new and innovative opportunities for interacting with digital data, resulting in an exponential growth of multimedia content. Managing this surge poses a significant challenge for retrieval techniques. This paper explores the convergence of Artificial Intelligence (AI), Mixed Reality (MR), and multimedia retrieval, introducing *(MR)*^{2}, a transformative concept bridging the physical and digital worlds.

Motivation: The increasing demand for seamless user interactions with multimedia content motivates our research. Traditional retrieval systems, reliant on text-based queries, often limit immersive experiences. In MR environments, our goal is to enable effortless engagement with multimedia content, leveraging AI-powered object detection.

Use Case: MR in an Art Gallery: Consider a user in an art gallery wearing an MR headset. Instead of navigating menus, our system, *(MR)*^{2}, powered by AI, allows users to interact directly with artworks. As the user explores, the system's live query option scans their field of view, identifying potential exhibits. When the user focuses on a striking painting, *(MR)*^{2} recognises the object and offers dynamic information about it. The user can then seamlessly query for similar artworks, creating a fluid, information-rich experience within the MR environment. This transforms art exploration into an immersive journey where users can intuitively access multimedia content.

Research Objective and Contribution: We aim to redefine multimedia retrieval in MR environments through a robust framework, integrating AI-driven object detection, MR technologies, and multimedia retrieval. This paper introduces *(MR)*^{2} and demonstrates the revolutionary impact of AI-powered live queries on user interactions in MR environments.

Outline: Section II introduces tools, and Section III delves into *(MR)*^{2}'s conceptual architecture. The AI-powered prototype for iOS is detailed in Section IV, and Section V summarises the experiment results. Section VI surveys related research, and conclusions are discussed in Section VII.

II. Foundations

This section introduces fundamental concepts essential for understanding our approach to multimedia retrieval in MR, emphasising the role of Artificial Intelligence (AI).

A. Multimedia Retrieval

Multimedia retrieval involves searching and retrieving content based on user queries from diverse datasets, including images, videos, audio, and text. This process is crucial in content-based image retrieval and video recommendation applications. The challenge lies in developing efficient retrieval systems for multimedia data in various modalities and formats [1]. AI is pivotal in designing effective retrieval engines that produce a ranked list of documents based on user query relevance.

B. Mixed Reality

Mixed Reality (MR) exists on the spectrum between Virtual Reality (VR) and Augmented Reality (AR). VR immerses users in a wholly digital environment, while AR overlays digital content onto the physical world. MR, blurring the

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.

boundaries between physical and virtual domains, offers exciting possibilities for interactive and immersive experiences. Applications of MR affect various industries, including entertainment, education, healthcare, and industry. MR hardware, including headsets like Microsoft HoloLens[1] and Meta Quest Pro[2], provides users with a seamless blending of digital and physical worlds.

C. Object Detection

Effective recognition and interaction with physical objects in real-time within MR environments require a specialised approach to object detection. AI, particularly advanced deep learning techniques like YOLOv8[3] (You Only Look Once) [2] and Faster R-CNN [3], has made significant strides in detecting objects within images and video streams with remarkable accuracy and real-time performance. These AI-driven techniques serve as the backbone for object detection in MR scenarios, enabling rapid and accurate identification of objects in the user's environment. This process involves breaking down the environment into images and analysing each to identify objects and their properties.

Ultzralytics has developed the eighth version of YOLO, which divides an input image into a grid of cells responsible for predicting objects within their boundaries. YOLOv8 indicates a set number of bounding boxes for each grid cell defined by their coordinates, confidence scores, and class probabilities for different object categories. The model employs a deep neural network with convolutional layers to extract features from the input image at multiple scales. Feature fusion techniques are then used to combine information from different levels of the network to capture context and details simultaneously, making it robust in detecting objects of varying sizes and contexts. After prediction, YOLOv8 applies non-maximum suppression to filter out redundant bounding boxes and keep only the most confident ones for each object. With its efficient architecture and feature fusion techniques, YOLOv8 is widely employed in applications such as autonomous driving, surveillance, and object recognition for its speed and reliability.

D. Visual-Text Co-Embedding

Visual-text co-embedding is a highly effective method that combines visual and textual features to enhance multimedia retrieval systems. The Contrastive Language-Image Pretraining (CLIP) architecture [4], [5], a breakthrough in this field, relies on AI to create a joint embedding space where text and images can be directly compared. This allows for seamless integration of textual metadata with visual content, resulting in more robust and context-aware retrieval. The AI-driven CLIP architecture employs a vision and text encoder jointly trained to encode images and their textual descriptions into the same space. Using a contrastive loss function, CLIP brings similar image-text pairs closer in the shared space while pushing dissimilar pairs apart. CLIP's capabilities extend to tasks such as zero-shot image classification, showcasing the transformative impact of AI on multimedia retrieval.

III. Mixed Reality Multimedia Retrieval Framework

(MR)^2 revolutionises multimedia retrieval in Mixed Reality environments, creating a first iteration in the vision of a seamless blend of natural and virtual worlds. Leveraging advanced machine learning (ML) technologies, *(MR)*^{2} enables real-time interaction and a user-centric design.

A. Key Principles

(MR)^2 follows fundamental principles that profoundly influence its core philosophy:

Immersion: *(MR)*^{2} is designed to immerse users in an MR environment, seamlessly blending the physical and digital realms. The goal is to transport users into an augmented space, allowing them to interact with their surroundings while seamlessly accessing and engaging with digital content. This immersive experience aims to be compelling, making users feel fully present within the MR environment.

Real-time Interaction: A strong focus on real-time interaction is central to *(MR)*^{2}. This principle ensures that users can perform actions and receive responses without perceptible delays. Whether capturing images, selecting objects or retrieving multimedia content, *(MR)*^{2} prioritises immediate and fluid interactions, enhancing the overall user experience and making it dynamic and engaging.

User-Centric Design: *(MR)*^{2} adopts a user-centric design approach, placing the user's needs and preferences at the forefront. The system is crafted to be intuitive, user-friendly, and adaptable to individual requirements. This user-centric design encompasses the entire user journey within the Mixed Reality environment, aiming to cater to a diverse user base and ensure the concept is accessible and enjoyable for all.

Integration of Cutting-Edge ML Models: To achieve accurate object detection and relevance in content retrieval, *(MR)*^{2} integrates cutting-edge machine learning models. These models represent ML and computer vision research standards, underscoring *(MR)*^{2}'s dedication to leveraging technological advancements to provide users with precise and meaningful results.

B. Conceptual Model

The conceptual model of *(MR)*^{2} features a frontend and backend, each serving distinct purposes, as shown in Figure 1.

Frontend: The frontend is designed to capture user interactions and perform computations on the device. It offers three query modalities: object detection, area selection, and text queries. Moreover, the frontend ensures that the query results are presented engagingly, making it easier for users to understand and utilise the information.

[1] https://www.microsoft.com/en/hololens

[2] https://www.meta.com/en/quest/quest-pro

[3] https://docs.ultralytics.com

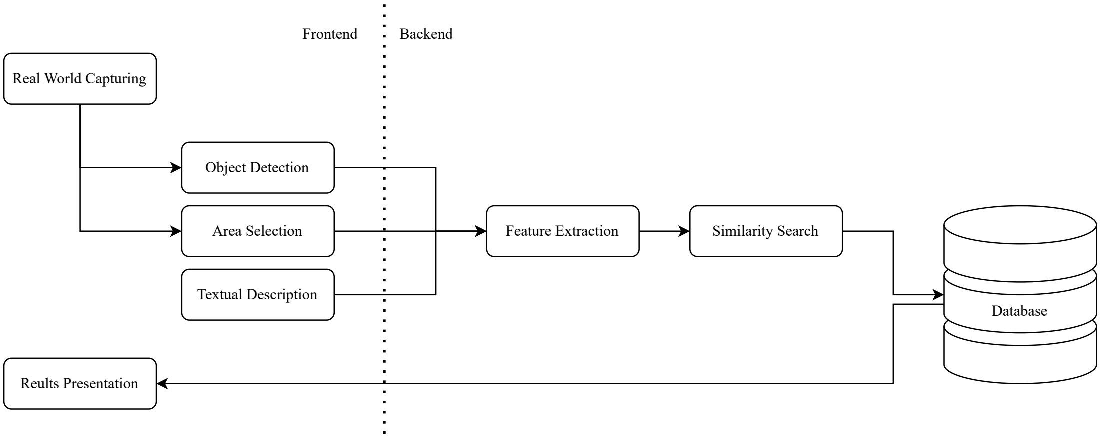


Fig. 1: Conceptual Architecture of *(MR)*^{2}

Backend: It handles data from queries, processes inputs employing a machine learning model, and performs similarity searches using feature vectors.

This model aligns with *(MR)*^{2}'s principles, seamlessly merging real and digital worlds, ensuring real-time functionality, and prioritising user-centric design. Its flexibility in ML model selection enables adaptability to future breakthroughs.

IV. Implementation

This chapter presents the prototype implementation of *(MR)*^{2}, covering its architecture and key components. *(MR)*^{2} operates as an iOS application on iPhones and iPads, communicating with a server-based backend. *(MR)*^{2}'s architecture comprises three essential components: (i) The iOS application, written in Swift, facilitates mixed-reality interactions and object detection. (ii) Communicating through a Python-based API, the server manages CLIP feature extraction and (iii) executes similarity searches in Cottontail DB [6].

Camera Feed: *(MR)*^{2} integrates the camera feed using AVFoundation, capturing real-world surroundings for an interactive experience. The app ensures a responsive interface, allowing users to switch between front and back cameras effortlessly.

Object Detection: Object detection is powered by Apple's Vision and CoreML frameworks, featuring the YOLOv8 model. This dynamic combination efficiently identifies objects in the live camera feed. Vision and CoreML collaboratively process predictions, presenting real-time bounding boxes around detected objects, as shown in Figure 2a.

Similarity Search: The backend performs CLIP feature extraction on captured images, followed by a similarity search in Cottontail DB. This process ensures that *(MR)*^{2} retrieves the top 100 similar objects in real-time, concluding the multimedia retrieval process.

Result Presentation: *(MR)*^{2} presents search results in real-time within a dedicated ViewController. A scrollable grid displays the most similar images, with the highest similarity in the top-left corner (Figure 2d). Users can navigate through results and enlarge individual objects for closer inspection.

Further Query Types: Apart from the live query, *(MR)*^{2} supports additional options: (i) Users initiate a region-based query through a touch input, positioning a rectangle on the screen in the camera feed, as seen in Figure 2b. (ii) For a text query, presented in Figure 2c, users enter a description, and CLIP enables cross-modal similarity comparisons with image content.

V. Evaluation

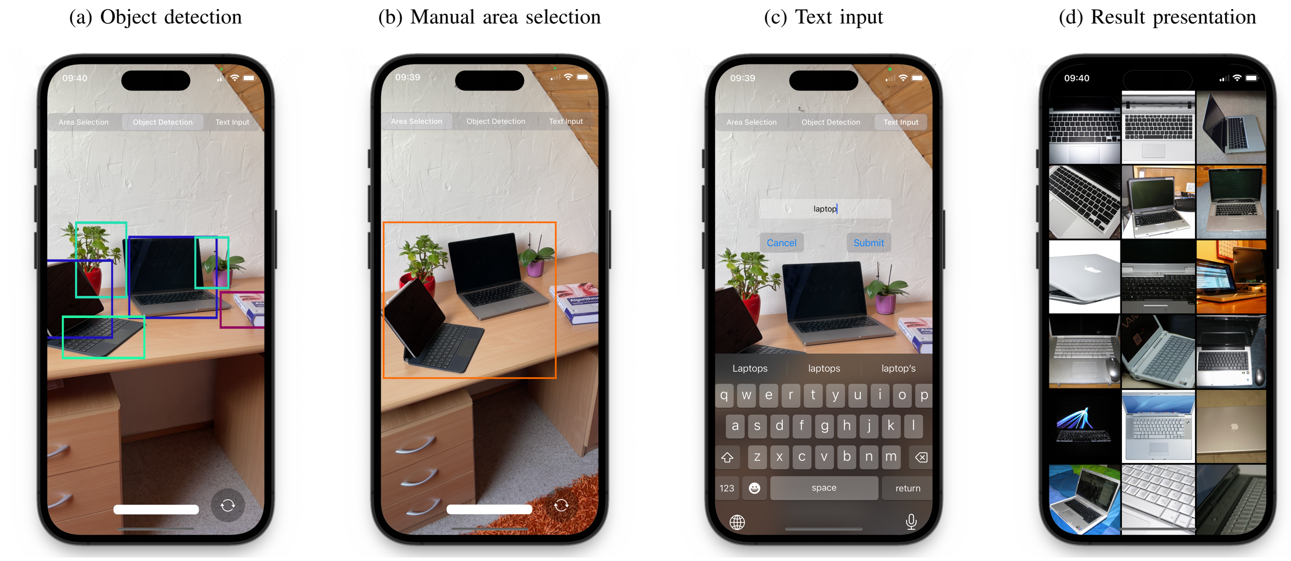
This section provides a comprehensive evaluation of *(MR)*^{2}, encompassing both analytical and user-centric assessments. We begin by analysing the functional performance through performance and user evaluations, measuring object detection inference time, query response time, and real-world usability. The section concludes with a dedicated discussion on the overall performance and potential future advancements.

A. Performance and User Evaluation

The analytical evaluation focused on pure performance, measuring object detection inference time and query response time. The median inference time of 10,000 measured detections stood at 24.8 ms, ensuring real-time applicability. The average query time (100 measurements), at 4191 ms, demonstrated efficiency, considering the extensive ImageNet Large Scale Visual Recognition Challenge (ILSVRC) dataset [7] stored in Cottontail DB.

Concurrently, the user evaluation, involving 14 participants, delved into real-world scenarios, emphasising practical usability. Participants displaying a moderate to high technology affinity (average ATI score of 4.43 [8]) found *(MR)*^{2} user-friendly and efficient. The System Usability Scale (SUS) score of 87 reinforced this, reflecting positive perceptions of usability [9].

Open feedback from users underlined the positive SUS result, highlighting *(MR)*^{2}'s intuitiveness and practicality. Minor concerns included the selection of overlapping bounding boxes when multiple objects are detected. Despite this, participants expressed a willingness to continue using *(MR)*^{2}, showcasing its user-friendliness and potential for future adoption.


(a) Object detection

(b) Manual area selection

(c) Text input

(d) Result presentation

Fig. 2: The four screenshots present the different views using *(MR)*^{2}.

B. Discussion

While the feedback is positive, *(MR)*^{2}acknowledges areas for improvement, such as expanding supported object types and laying the foundation for future advancements. Incorporating user feedback is essential in refining and extending *(MR)*^{2}. Envisioning an immersive future in multimedia retrieval, we list five additional exciting avenues: (i) Advances in query modes, including OCR, ASR, and tag search modalities, are worthwhile integrations supporting more classical methods. (ii) We can create temporal queries to introduce complexity and context awareness, enriching the search landscape. (iii) Developing immersive result presentations in MR contexts, seamlessly overlaying search results, is a promising direction. (iv) Beyond enhancing iOS accessibility, compatibility with MR glasses like Meta Quest Pro or future Apple Vision Pro devices opens avenues for more immersive applications. (v) Advancements in object detection models and self-training approaches are crucial in shaping the future of MR multimedia retrieval. Therefore, testing and comparing different models could be interesting.

VI. Related Work

Within the dynamic landscape of Extended Reality (XR), various multimedia retrieval systems have paved the way for innovations preceding *(MR)*^{2}. These systems offer diverse perspectives on integrating XR with multimedia, enriching user interactions in immersive environments.

Google Lens[4] bridges the physical and digital realms by employing image recognition and AI through smartphone cameras. It provides real-time information, translations, and interactive actions, leveraging Google's search capabilities.

Innovating VR interfaces for multimedia retrieval, such as *vitrivr-VR* [10], [11], connected to Cineast [12], prioritises text input in VR, diverging from *(MR)*^{2}'s live queries. Meanwhile, *ViRMA* [13] projects multimedia objects for visual analytics using the multi-dimensional multimedia model (M^3) [14] in VR.

Focusing on cultural heritage exploration, *GoFind!* [15] combines content-based multimedia retrieval with Augmented Reality (AR). Unlike *(MR)*^{2}, *GoFind!* highlights historical querying with varied query modalities.

VII. Conclusions

In this paper, we introduced *(MR)*^{2}, a new concept in MR Multimedia Retrieval that harnesses the capabilities of MR technology to transform user interactions with multimedia content. *(MR)*^{2} presents a distinctive approach to query formulation and pioneers the innovative live query option, seamlessly connecting the digital and physical worlds. Implementing a prototype on iOS devices demonstrated *(MR)*^{2}'s potential to enhance the user's multimedia retrieval experience.

Utilising the YOLOv8 model for object detection in *(MR)*^{2} marks a critical component. While acknowledging opportunities for expanding supported object types, *(MR)*^{2} presents the potential for object recognition in MR environments.

Leveraging the CLIP machine learning model for similarity searches enhances retrieval accuracy. *(MR)*^{2} delivers refined and personalised experiences by efficiently identifying and presenting relevant multimedia files.

Positive user evaluations underscored *(MR)*^{2}'s potential for engaging multimedia retrieval experiences. The live query option and seamless MR integration received particular acclaim, validating *(MR)*^{2}'s user-centered design.

As we look towards the future, *(MR)*^{2} lays the foundation for further advancements, building upon the insights gained from analytical and user evaluations. This paper sets the stage for ongoing research, envisioning a future where multimedia retrieval is efficient, deeply immersive, and intuitive.

[4] https://lens.google

References

[1] S. Rüger and G. Marchionini, *Multimedia Information Retrieval*. Morgan & Claypool, 2009.

[2] 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

[3] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015*, Montreal, Quebec, Canada, Dec. 2015, pp. 91–99. [Online]. Available: https://proceedings.neurips.cc/paper/2015/hash/14bfa6bb14875e45bba028a21ed38046-Abstract.html

[4] R. Mokady, A. Hertz, and A. H. Bermano, "ClipCap: CLIP Prefix for Image Captioning," *CoRR*, vol. abs/2111.09734, 2021.

[5] 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

[6] 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.

[7] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[8] T. Franke, C. Attig, and D. Wessel, "A Personal Resource for Technology Interaction: Development and Validation of the Affinity for Technology Interaction (ATI) Scale," vol. 35, no. 6, pp. 456–467. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/10447318.2018.1456150

[9] J. Brooke, "SUS: A 'Quick and Dirty' Usability Scale," in *Usability Evaluation In Industry*. CRC Press, 1996, num Pages: 6.

[10] F. Spiess, R. Gasser, S. Heller, L. Rossetto, L. Sauter, and H. Schuldt, "Competitive Interactive Video Retrieval in Virtual Reality with vitrivr-VR," in *Proceedings of the 27^{th} International Conference on MultiMedia Modeling (MMM 2021) – Part II*, ser. Lecture Notes in Computer Science, vol. 12573. Prague, Czech Republic: Springer, Jun. 2021, pp. 441–447. [Online]. Available: https://doi.org/10.1007/978-3-030-67835-7_42

[11] F. Spiess, P. Weber, and H. Schuldt, "Direct Interaction Word-Gesture Text Input in Virtual Reality," in *IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR 2022), Virtual Conference*. IEEE, Dec. 2022, pp. 140–143. [Online]. Available: https://doi.org/10.1109/AIVR56993.2022.00028

[12] L. Rossetto, I. Giangreco, C. Tanase, and H. Schuldt, "vitrivr: A flexible retrieval stack supporting multiple query modes for searching in multimedia collections," in *Proceedings of the 24th ACM international conference on Multimedia*. ACM, Oct. 2016.

[13] A. Duane and B. Þ. Jónsson, "ViRMA: Virtual Reality Multimedia Analytics," in *ICMR '22: International Conference on Multimedia Retrieval*. Newark, NJ, USA: ACM, Jun. 2022, pp. 211–214. [Online]. Available: https://doi.org/10.1145/3512527.3531352

[14] S. Gíslason, B. Þ. Jónsson, and L. Amsaleg, "Integration of Exploration and Search: A Case Study of the M^3 Model," in *Proceedings of the 25^{th} International Conference on MultiMedia Modeling (MMM 2019) – Part I*, ser. Lecture Notes in Computer Science, vol. 11295. Thessaloniki, Greece: Springer, Jan. 2019, pp. 156–168. [Online]. Available: https://doi.org/10.1007/978-3-030-05710-7_13

[15] L. Sauter, T. Bachmann, L. Rossetto, and H. Schuldt, "Spatially Localised Immersive Contemporary and Historic Photo Presentation on Mobile Devices in Augmented Reality," in *Proceedings of the 5^{th} International Workshop on the Analysis, Understanding and Promotion Heritage Contents (SUMAC'23) – Advances in Machine Learning, Signal Processing, Multimodal Technical and Human-Computer Interaction*. Ottawa, Canada: ACM, Nov. 2023.