big data and cognitive computing

MDPI

Article

A Reference Model for the Analysis and Indexing of Metaverse Recordings for Information Retrieval

Patrick Steinert [1,*], Stefan Wagenpfeil [2], Ingo Frommholz [3] and Matthias L. Hemmje [1]

[1] Faculty of Mathematics and Computer Science, University of Hagen, Universitätsstrasse 1, 58097 Hagen, Germany
[2] Faculty of Business Computing and Software Engineering, PFH University of Applied Science, 37073 Göttingen, Germany; s.wagenpfeil@pfh.de
[3] School of Applied Data Science, Modul University Vienna, 1190 Vienna, Austria; ifrommholz@acm.org
* Correspondence: psteinert@acm.org

Abstract

After the peak of the recent hype wave of interest surrounding the metaverse, virtual world applications remained in areas such as gaming, VR training, simulations, and collaboration. In this context, recordings are created which subsequently evolve into extensive collections that users may wish to access, search through, and retrieve items from. In order to facilitate searchability of metaverse recordings, it is necessary to adapt content analysis and indexing techniques to the specific characteristics of these recordings. This paper presents a reference model, the Processing Framework for Metaverse Recordings (PFMR), which details the phases of structural analysis, feature extraction, data mining, and feature fusion. The objective is to facilitate efficient retrieval of metaverse content. Our evaluation, based on a prototypical implementation, demonstrates the applicability and effectiveness of *PFMR*. This lays the groundwork for further integration of metaverse-specific content into Multimedia Information Retrieval systems. The evaluation of the 256 Metaverse Recording dataset shows that *PFMRs*' domain-specific adaptability and integratability allows effective metaverse recording information retrieval for metaverse-specific features such as avatar detection, dialog mining, and toxicity classification.

Keywords: multimedia; content analysis; metaverse recordings; multimedia information retrieval; time-series content analysis; multimedia systems

1. Introduction

The creation and relevance of multimedia play a crucial role in today's mainstream media landscape. High quality ubiquitous cameras enable the creation of a vast amount of multimedia, while the rise of the internet and social media networks drives its creation and distribution [1]. This development leads to extensive multimedia repositories, including personal smartphone libraries [2], research-oriented collections [3], and large-scale archives on social media platforms [4]. Leveraging such collections requires efficient and effective search, a task addressed by Multimedia Information Retrieval (MMIR) [5], the research field concerned with indexing and retrieving multimedia content objects [6], i.e., videos, images, text, audio, or their combinations.

In recent years, a notable trend in multimedia has emerged: the growth of virtual reality applications [7,8]. The metaverse is a network of connected perpetual and immersive virtual environments [9] and can be regarded as a novel user interface paradigm for current multimedia platforms, including social media [10] and video conferencing [11]. After a

Check for updates

Academic Editor: Yin Zhang

Received: 23 November 2025
Revised: 24 February 2026
Accepted: 3 March 2026
Published: 9 March 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

brief appearance of the metaverse on the Gartner Hype Cycle in 2022 [12], it was estimated to take more than 10 years to reach the productivity plateau; it has already disappeared from the 2023 edition [13]. However, actual usage continued to grow [14], for example, in gaming. Users also participate in other activities in the metaverse, e.g., socializing, watching videos, attending live events, virtual travel, or attending virtual classes, meetings, or work-related events [15]. According to Statista [16], two out of five U.S. citizens are at least somewhat familiar with the metaverse. In addition to consumer use cases, collaboration [17], VR training [18], and simulation [19] have also shown increasing relevance. All such applications can produce recordings of user sessions, referred to as Metaverse Recordings (MVRs) [20].

Present-day metaverse virtual worlds are technologically based on Massive-Multiplayer Online Games [21]. The audio-video stream produced by the game software can be recorded, e.g., as a screen recording. Furthermore, virtual worlds can contain and be represented by multimedia content objects, such as 3D models and textures, which can also be recorded. In whatever format the MVR is recorded, the content and structure differ from other types of media [6]. Thus, MVRs can be considered a new type of multimedia and a new type for MMIR.

The central challenge for MMIR to support MVR lies in determining how MVRs can be effectively analyzed and indexed by integrating existing analysis methods in a manner that accounts for their composite, temporally aligned, and heterogeneous nature [6]. While numerous techniques for multimedia analysis exist, it remains unclear to what extent current MMIR analysis frameworks can be applied to *MVRs* or how they must be adapted to accommodate these characteristics. The contribution of this work therefore lies not in introducing new feature extraction methods, but in describing and validating a structured processing framework that enables flexible composition and orchestration of heterogeneous analysis components, allowing domain-specific adaptation and iterative refinement of analysis pipelines for *MVR* information retrieval.

In earlier publications [6,20], we introduced the concept of a *Processing Framework for Metaverse Recordings (PFMR)*. This paper updates and details the phases of the *PFMR*: *MVR Structural Analysis*, *MVR Feature Extraction*, *MVR Data Mining*, and *MVR Feature Fusion*. It provides a detailed description of the integration into MMIR and presents a systematic evaluation, thereby addressing an aspect that has so far been insufficiently explored in existing work. The objective of *PFMR* is to facilitate efficient information retrieval of MVRs. Our evaluation, which is based on a prototypical implementation, demonstrates the applicability and effectiveness of *PFMR*. This lays the groundwork for further integration of MVRs into MMIR systems.

We expect that the contributions are applicable to similar domains of temporal multimodal multimedia analysis, such as social media analytics, autonomous driving, or robotics.

The methodological foundation of this work follows a research approach proposed by Nunamaker et al. [22]. This approach structures the research process into four interrelated stages: problem observation, conceptual design and modeling, system realization, and evaluation. In this paper, the problem observations are established through the analysis of existing work in Section 2. The conceptual designs and models are introduced in Section 3. The realization of the proposed approach in a prototypical system is described in Section 4, while the experimental evaluation and discussion of results are presented in Section 5.

2. State of the Art and Related Work

The following section presents a literature study on the state of the art in science and technology with regard to existing methodologies and established frameworks relevant to virtual worlds, MMIR, and associated studies. Initially, groundwork is established

through an exploration of virtual environment concepts and their technological components. Subsequently, *MVR* and MMIR are introduced and contextualized. Furthermore, pertinent investigations into multimedia data examination are discussed.

2.1. Metaverse

Stephenson's 1992 novel introduced the concept of the metaverse [23], a term that has since evolved into a broad marketing descriptor [24] applied to various technologies, ranging from three-dimensional web-based platforms to what some envision as the Internet's successor [25]. Among the various interpretations, Mystakidis offers a rigorous conceptualization, characterizing the metaverse as an ongoing, multi-participant space that transcends conventional reality by integrating the physical and digital domains [26]. Users navigate these environments through avatars, typically three-dimensional representations that may resemble the user or embody imagined personas.

According to the Metaverse Roadmap [27], a cross-industry foresight project report, two fundamental variables shape metaverse development: the degree of intimacy between users and the balance between purely simulated versus augmented experiences. These variables generate different potential futures, with mirror worlds [27] and virtual worlds [27] emerging as particularly viable scenarios [28] within fully simulated environments. Simulated worlds have gained broader adoption compared to augmented reality applications [28]. Throughout this paper, virtual world encompasses both mirror worlds and virtual worlds.

While metaverse applications remain predominantly centered on gaming and entertainment contexts, which continue to drive the majority of user engagement and platform growth [14,29–31], a growing body of research demonstrates the increasing relevance of metaverse technologies in professional domains. Beyond gaming, metaverse technologies now serve professional training purposes [18] and athletic development [32,33]. The industrial metaverse [34] leverages comparable technological frameworks to create digital representations of architectural structures, manufacturing facilities, and extractive operations [35]. A comprehensive framework for Industry 4.0 technology adoption [36] identifies virtual and augmented reality alongside simulation as key enabling technologies for digital transformation in enterprise settings, demonstrating how immersive technologies support training, digital twins, and operational optimization across business domains. These developments suggest that professional applications such as VR-based training, collaborative workspaces, and industrial simulations are gaining relevance alongside entertainment-oriented use cases. Consequently, recordings from such diverse contexts may exhibit varying characteristics, which motivates the exploration of adaptable content analysis approaches.

2.2. Metaverse Recordings

Grounded in the identified application domains, recordings of the user sessions are made for use cases of replay [37], analysis [18], or machine learning training [38]. In addition to videos, there is growing interest in the metadata of user sessions [39]. This includes details such as the name and location of the virtual world, the digital identities of avatars present, and sensor data like hand controller and head movements. In a field study [40], 21 existing and potential application scenarios were identified. One example is market research, where VR technology is used to simulate supermarket environments and analyze recorded video and gaze trajectories to study product placement. Moreover, the ability to replay scenes with freedom to move around within the recording, as was previously available in AltSpace VR [41] or as is currently available in EngageXR [42], is also of interest.

Introduced in [6,20], MVRs may encapsulate multiple forms of information. Real-world scenes are typically captured through photographic, audio, or video recordings that

depict physical environments. In contrast, recordings generated within virtual worlds differ essentially [6] and therefore require specific terminology. Figure 1 illustrates the definitions. A scene in a metaverse-based virtual world can be captured using several approaches. The most straightforward method is a capture of the rendered video stream, producing a single image or a video, optionally accompanied by audio. This type of data is referred to as a Multimedia Content Object (MMCO) [43]. Beyond rendered output, a virtual world constitutes a technical space composed of elements such as geometric shapes, textures, and vectors. Consequently, it can also be recorded in a raw representation that includes the scene graph, vectors, and textures used for rendering. This representation is defined as Scene Raw Data (SRD) [6]. In addition, Peripheral Data (PD) [6] may be captured from sensors such as motion controllers or eye- and face-tracking systems. Such information, including affective cues like emotions, represents a valuable resource for MMIR [44,45].

The relevance of SRD and PD has been demonstrated in several application scenarios identified in the field study [40], including marketing research and personal use cases such as golf training, as well as VR-based police training [18,46]. Furthermore, Ref. [45] demonstrates how SRD and PD can be recorded and utilized for *MVR* information retrieval. Despite their demonstrated value, the recording of such data is still uncommon in current metaverse implementations, particularly in entertainment-oriented applications.

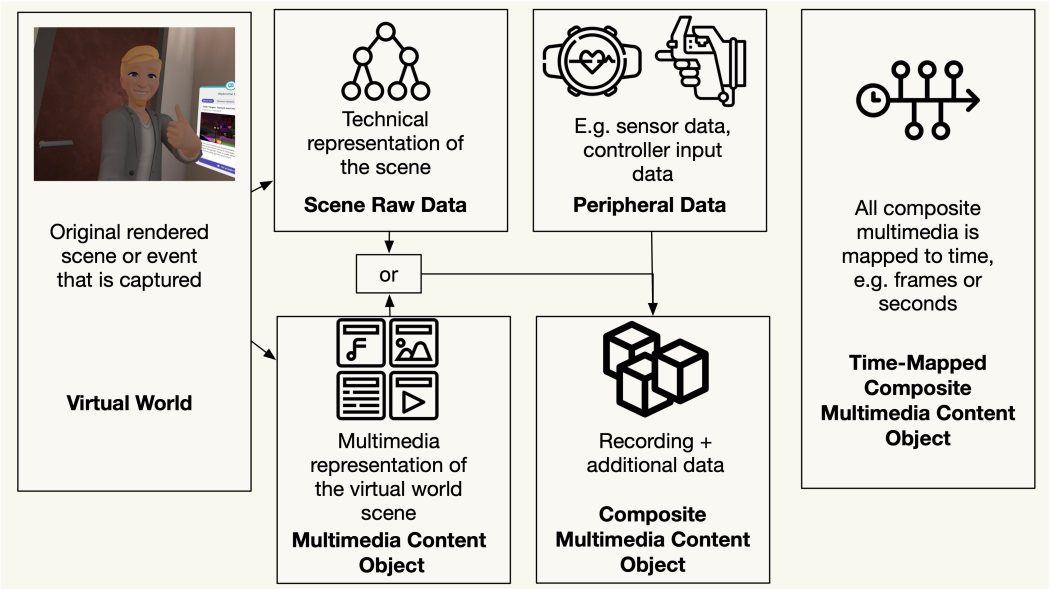


Figure 1. Metaverse multimedia terms and definitions [20].

MVRs can therefore consist of multiple heterogeneous components, including rendered multimedia content, rendering data representations, and peripheral sensor data. To define this structure formally, these components are unified under the concept of a Composite Multimedia Content Object (CMMCO) [6]. For the formal representation of a CMMCO, three sets are defined: MMCO, representing multimedia content objects; SRD, denoting the associated scene raw data; and PD, comprising peripheral sensor data. The CMMCO therefore contains various descriptive representations of MMIR-relevant features r_i, where each feature can be assigned to its originating set:

$$CMMCO = \cup_{i=1}^{n} r_i \Leftrightarrow r_i \in MMCO \cup \{SRD, PD\} \quad (1)$$

When the components of a CMMCO share a common temporal reference system, such as frame numbering or time codes, that enables data alignment, this configuration is referred to as a Time-mapped Composite Multimedia Content Object (TCMMCO) [6].

A standard for recording sessions in a way that allows the scene to be reassembled is unknown. Nvidia Capture and Replay [47] is a project that enables the recording of trajectories in a virtual VR world as SRD. The current implementation is limited in its compatibility with a wide range of VR applications. Another unclear aspect is the storage and import of the elements produced for an MVR. These can be packaged and stored in a container or archive format, which has not yet been defined.

The availability of datasets for *MVR* research is limited. Several datasets exist for virtual worlds, but specific metaverse datasets are rare. Ref. [48] presents a dataset to find virtual worlds, but it is not based on recordings. In [49], we presented the 256 Metaverse Recording dataset for research purposes. The dataset contains 256 individual MVRs based on screen recordings originating from a variety of 13 metaverses, each consisting of multiple virtual worlds. It includes recordings and annotations of activities, objects, and interactions in metaverse virtual worlds. It is the largest dataset with a high variety of virtual worlds known to the authors.

The recording of behavior in the metaverse rises privacy concerns, in particular if sensitive biometric data are recorded. In controlled environments, such as training or research, the consent of participants is required. However, this is part of future work. In addition to privacy risks, since the recording is produced in a virtual environment, trust is another concern [50]. C2PA [51] addresses the trustworthyness challenge by adding cryptographical provenance metadata to the media [51], called content credentials. Assuming the MVRs contain content credentials, these do not affect the content analysis but provide the user a feature to filter in the retrieval process.

Next, we describe MMIR and related work supporting MVRs.

2.3. Multimedia Information Retrieval

In general, MMIR [5] supports access to data in large and heterogeneous multimedia collections. An integral part of this process is the retrieval function, which matches queries against an index of the multimedia assets in the collection. This index includes metadata about each asset to make it retrievable. Semantic metadata connects features, making them machine-processable and interpretable. Metadata can be obtained through feature extraction. Semantic features are commonly organized using graph-based representations, in which information elements are modeled as nodes connected by their semantic relationships [52]. Smart MMIR [43] employs a graph model, the Multimedia Feature Graph (MMFG), which provides an expressive, scalable, semantic, and explainable feature representation.

The Generic Multimedia Analysis Framework (GMAF) [43] is an extensible and integrable MMIR framework based on the MMFG. It provides a plugin integration mechanism to add feature extraction plugins. To process the output of multiple plugins, GMAF also contains an extensible fusion mechanism. To orchestrate several plugins and fusion mechanisms, GMAF provides Flows, a workflow description mechanism [6].

Despite rich research in the field of MMIR, it lacks support for MVRs [6]. There are several open challenges. For example, potential information requirements for metaverse recordings are described in [6], but are not sufficiently researched. There are several open challenges with regard to content analysis. For example, temporal structuring of metaverse recordings for MVR information retrieval is unknown. Abdari et al. [48] researched the retrieval of metaverse worlds, i.e., finding the virtual world to enter. The approach uses natural language as a query method. The approach is limited to classifying the appearances of virtual worlds and is limited to artificial virtual worlds, not real metaverse virtual worlds.

In [6], we proposed that specialized content analysis is needed for effective MVR information retrieval. For example, avatars are a basic concept in the metaverse, hence,

avatar detection [53] is explored and provides a specialized content analysis method. The complexity of MVRs in content, structure, and data types [6] requires effective content analysis and indexing for effective retrieval.

2.4. Analysis and Indexing in Information Retrieval

Accessing media collections efficiently through search requires an index [54]. The index is created based on extracted features that describe the collection's elements. Content Analysis [54], e.g., by feature extraction, is used to gain a semantic understanding of the content of a multimedia element. Several task-specific feature extraction methods exist, such as metadata extraction, object detection [55], spatial image segmentation [56], or temporal video segmentation [57].

To optimize the result of the content analysis and consequently the indexing of the multimedia elements for retrieval, a sequence of steps is often executed. Such sequences are referred to as a process or process framework when no strict sequence of steps is specified. For example, the organization chart of the histopathological WSI analysis [58] describes a process of image analysis, including steps of pre- and post-processing, multiple feature extractions, and image classification.

Subsequent to feature extraction, data mining is a method used to make sense of the extracted data. Data mining is the analytical process of exploring and analyzing large datasets to uncover hidden information, such as hidden patterns, or unknown correlations. For data mining projects, the Cross Industry Standard Process for Data Mining (CRISP-DM) [59] is defined. However, the process is modeled for human activities in the process.

In the field of video analysis and retrieval, Hu describes [60] a comprehensive range of strategies for video data mining. These include Object Mining, which focuses on identifying and grouping the same object across different segments of a video despite variations in appearance. Special Pattern Detection is employed to recognize specific actions or events, such as human movements or traffic patterns, using predefined models. Pattern Discovery involves the autonomous detection of unknown patterns using unsupervised or semi-supervised learning, crucial for discovering novel insights or initializing additional models. Video Association Mining aims to uncover relationships between different events or objects in the video, discerning the connections between various semantic concepts. Tendency Mining tracks and analyzes trends of certain events, understanding their occurrences and spatial-temporal relationships. Lastly, Preference Mining analyzes user preferences in specific types of videos, such as news or movies, to provide personalized content. Each of these strategies applies different techniques tailored to the unique demands of video data mining based on the information of previous feature extraction steps.

The complexity of the analysis increases with the complexity of the media itself. Simple images of cats on a white background require only a single classification, while the analysis of a movie video requires multiple in an organized way. For video retrieval, a process has been established over time [60]. 'Segmenting audio-visual data into homogeneous segments in a manner suited for further processing (e.g., visual summaries) is an important step. The segmentation of video data into manageable chunks is complicated...' [61]. Continuous research is carried out to address such problems.

Overall, the literature on video analysis and processing workflows demonstrates a broad range of solutions, most of which are highly specialized for particular domains [32,62,63] and tasks [64–67]. A substantial body of work targets narrow application scenarios, for example sports analytics, where processing stages are closely tied to domain-specific assumptions and feature extractors [68–71]. Complementary research proposes more high level pipelines, typically organized into segmentation, feature extraction, and classification stages, without further implementation details. In addition, several

studies focus on scalability and execution by embedding these pipelines within workflow management systems or distributed processing environments.

In summary, despite this diversity, existing approaches exhibit three fundamental limitations with regard to MVR information retrieval. First, they are largely based on fixed analysis pipelines designed for narrowly defined tasks. Such solutions optimize individual processing stages for a single application scenario and provide little support for flexibly reconfiguring or combining heterogeneous analysis methods within a unified framework. Consequently, they lack reusable, modular orchestration mechanisms that would allow analysis components to be adapted to different domains or evolving retrieval requirements. Moreover, current research does not sufficiently address how established multimedia analysis techniques can be systematically adapted and orchestrated for domain-specific retrieval in metaverse environments. Together, these shortcomings reveal a clear research gap concerning both the structured integration of analysis methods and their application to the characteristics of *MVRs*, motivating the need for a dedicated and flexible processing framework tailored to metaverse recordings. Second, current frameworks remain primarily focused on conventional video and other traditional media types, generally assuming a single dominant modality with structure derived implicitly from the signal. This design perspective does not accommodate the characteristics of *MVRs*, which consist of multiple heterogeneous and semantically distinct data sources—such as rendered video, scene data, and peripheral sensor streams—that are temporally aligned but conceptually independent. *MVRs* therefore integrate continuous temporal structure, multiple modalities, and semantically rich user interactions within a single composite artifact. Existing MMIR frameworks neither explicitly model this composite nature nor provide mechanisms for systematically combining heterogeneous analysis methods across these modalities. Third, many existing approaches do not demonstrate an integrable approach, i.e., how the various existing specific methods can be integrated and orchestrated in a workflow that can be adapted to domain-specific needs. Consequently, no current framework offers a generic and extensible model that simultaneously supports flexible pipeline configuration and the multimodal structure of *MVRs* in a manner suitable for retrieval-oriented MMIR systems.

In conclusion, several existing approaches provide useful methodological foundations that can be transferred to MVR content analysis. However, the limitations identified above continue to present significant challenges for the systematic analysis and retrieval of metaverse recordings.

In [6,20], we introduced an approach for *MVR* content analysis, the *Processing Framework for Metaverse Recordings (PMFR)*. *PFMR* is based on the commonly described video analysis process. *PFMR* is described in more detail in the modeling section.

2.5. Summary of the Literature Research

This overview of the current state of the art and related research in the area of MVRs describes open challenges. One of the primary gaps identified is the need for tailored content analysis methods due to the specific characteristics of MVRs. Current feature extraction approaches are often isolated and lack integration, with no comprehensive model available for MVR analysis, which hinders complex MVR information retrieval. Additionally, there is a critical need for adaptable analysis methods that can be tailored to specific application scenarios, allowing for effective integration and customization to address diverse challenges. These will be addressed in the next section.

3. Conceptual Design and Modeling

The foundation of our design draws upon Norman and Draper's User-Centered System Design principles [72], employing the Unified Modeling Language (UML) as the formal

notation framework for conceptualizing the MMIR system architecture. Our methodology begins by establishing a representative application scenario within the target domain: a user generates multimedia artifacts within a virtual environment—specifically, capturing immersive recordings of virtual scenes for archival or personal documentation purposes. These MVR instances subsequently undergo indexing and become retrievable through the MMIR infrastructure. Figure 2 illustrates the primary use cases that govern this workflow. The interaction model begins when a visitor produces an MVR and ingests it into the MVR-extended MMIR ecosystem. The system then performs automated analysis and indexing operations on the MVR content. Subsequently, consumers can execute search operations, encompassing query formulation and refinement use cases. The framework further supports result examination through viewing and filtering mechanisms. The terminal use case involves the interactive playback of MVR content within a dynamic filtering and presentation interface, enabling exploratory browsing and content consumption.

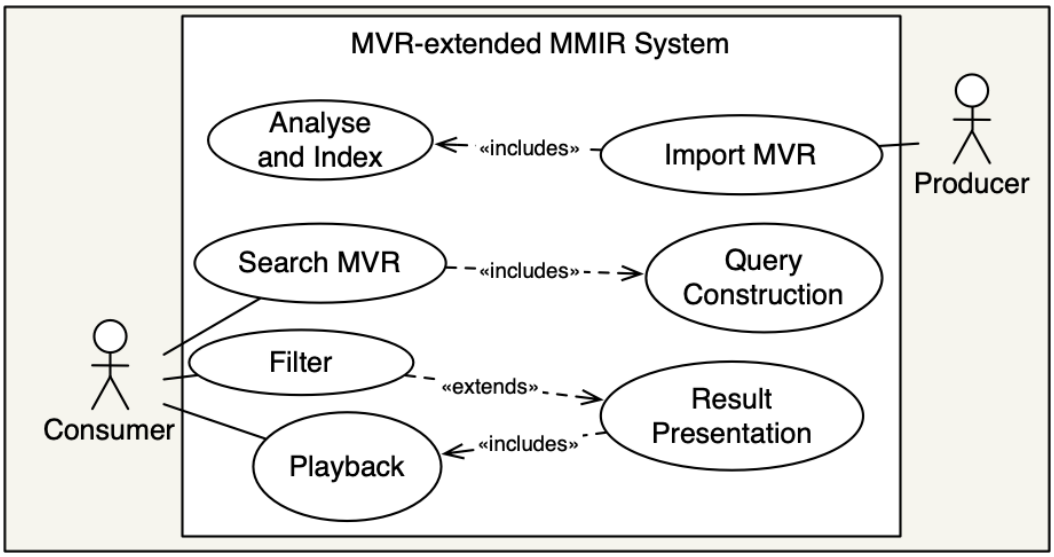


Figure 2. UML use case showing MMIR processing of metaverse multimedia [20].

The following sections describe extended modeling work for the use case of *Analyze and Index*, based on the previously proposed reference models *Processing Framework for Metaverse Recordings*.

3.1. Reference Model for the Analysis and Indexing of MVRs

The retrieval of MVRs and other media types shares a common foundation, as both aim to fulfill users' information needs through comparable feature extraction and retrieval mechanisms, relying on the same core system components [40]. Nevertheless, notable distinctions arise regarding content characteristics, data formats, and modes of interaction [6]. On the query side, only subtle differences in terminology can be observed. Whereas conventional multimedia analysis typically deals with real-world scenes, MVRs capture virtual or semi-constructed environments. Additionally, MVRs can exhibit considerable format variability, as they may consist of multiple heterogeneous data types. Regarding how users formulate their searches, their general retrieval behavior is unlikely to shift substantially; however, the language used may be specific to the virtual context. For example, when seeking to revisit a personal memory of a conversation, a user might enter a query such as 'the penguin I talked to', a phrasing that would be atypical in a real-world search scenario. Finally, how users engage with MVRs may diverge from conventional multimedia information retrieval, possibly requiring new strategies for presenting results

and facilitating user interaction. As such, the core difficulty in supporting MVRs stems from both the novelty of the multimedia type and the nature of its underlying content.

As an example use scenario, the information need of a user in the domain of cybersecurity could search for segments where toxic language occurs, e.g., insults.

To enable retrieval, the recorded data must undergo processing by the MMIR system. As discussed in Section 2.4, appropriate methods for MVR analysis and indexing have not been established, nor has the adequacy of existing frameworks been determined. Various content analysis methods are available to extract features (see Section 2.4), but they need to be orchestrated to be effective. The goal is to extract the set of relevant features R. For MVR content analysis, the introduced concept of a generic process framework, *PFMR*, supports CMMCO and extends the previously described common model for video retrieval. The process framework provides four steps: *MVR Structural Analysis*, *MVR Feature Extraction*, *MVR Data Mining*, *MVR Feature Fusion*, and conversion of the output information into a data structure through *MVR Indexing*. The flow of the elements is visualized in Figure 3. As a framework, *PFMR*can be adapted to domain-specific analysis needs. It clearly separates the phases of observation (structure and feature extraction), consolidation (feature fusion), and inference (data mining). The underlying hypothesis is that separating the analysis into three clearly defined phases leads to a more structured and transparent analysis process. The second hypothesis is that iterative refinement, realized through repeated interaction between feature fusion and data mining, improves the quality and relevance of the resulting feature set. Based on these assumptions, *PFMR*is expected to offer integratability and domain-specific adaptability, leading to effective MVR information retrieval.

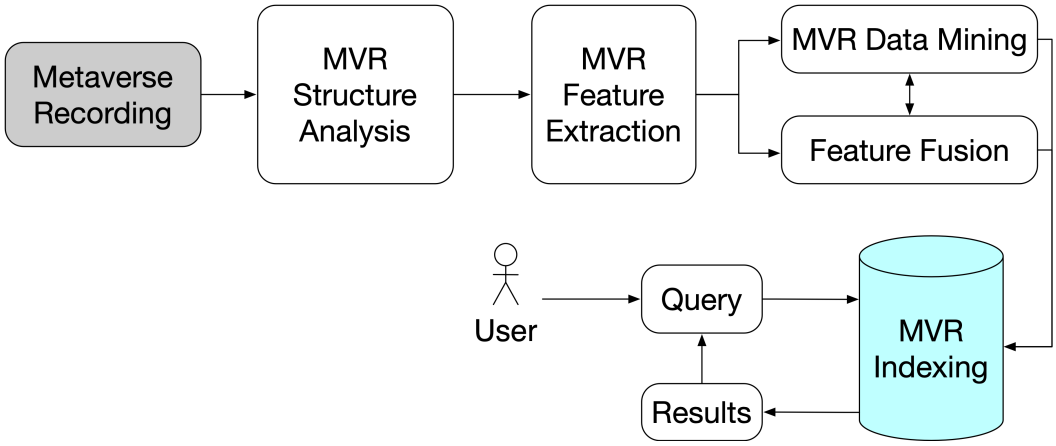


Figure 3. PFMR overview [6].

PFMR comes with a set of rules. Each process step passes the set of features to the next process step as input. In all cases, the process starts with *MVR Structure Analysis* followed by *MVR Feature Extraction*. The subsequent steps, *MVR Feature Fusion* and *MVR Data Mining*, can be modeled in various ways. The steps can occur multiple times and in any order before the final set of features is transferred to the index. A detailed explanation of the steps is given next.

The rationale to first process all the features is to extract all features from MMCO, SRD, and PD based on methods of their modality. Based on the abstract features, a feature fusion or data mining can be performed. This supports the reusability of existing methods with a low computational overhead for passing the data from one step to the next. Adding or replacing new feature extractors is easy to achieve compared to a highly specific processing method.

3.1.1. MVR Structural Analysis

Structural analysis can be spatial or temporal. For example, the analysis can recognize and isolate spatial areas of the video. Similar to the example of the described ultrasonic videos, MVR_{MMCO} contains some overlays, such as menus or indicators. Such elements could be detected and maybe removed by a plugin before passing over to an object detection to improve the results.

User sessions in the metaverse can be considered long term. VR training can take multiple minutes, classes can take 1.5 h, and games may even last multiple hours. Searching for specific moments is likely. Furthermore, video retrieval [61] describes that a breakdown in structural components often benefits further analysis.

Common Video Retrieval performs a temporal segmentation of the videos into frames, shots and scenes [73]. For MVR, segmentation in shots and scenes is not suitable, because metaverse sessions often are continuous movements. Another analysis method is the application of event-based segmentation from lifelong research. The analysis of SRD data can also be interesting if certain events are logged, e.g., new game round. A structural analysis provides valuable information for further feature extraction. Hence, the structural analysis is a first step in further processing.

The Structural Analysis fs for MVR, which produces a set of segments $S = \{s_1, \ldots, s_n\}$ can be defined as

$$fs(MVR) = S = \{s_1, \ldots, s_n\}, \forall s_i \in MVR \tag{2}$$

In addressing the aforementioned exemplary use case, temporal segmentation can be exemplified through division by fixed time intervals or via event-based segmentation, wherein an event is, for example, a teleportation to another world.

3.1.2. MVR Feature Extraction

A feature extraction based on a semantic concept [74] helps to bridge the semantic gap, i.e., the gap between a technical feature representation and its human understandable meaning. For example, object detection methods can be used to detect objects and more, such as positional, compositional, and semantic relationships, e.g., interaction. The features F are derived from feature extraction fe on the components $MMCO$, SRD, and PD of the MVR.

Let fe be a function for feature extraction. fe, applied to an input $x \in MVR$, produces a set E of extracted features e.

$$fe(x \in MVR) \rightarrow e \in MVR \Leftrightarrow e \in E, e \subset R \tag{3}$$

Many feature extraction methods are unimodal, hence, depending on the use case, the step requires $fe_{MMCO}, fe_{SRD}, fe_{PD}$.

In the case of multiple segments from the previous segmentation, in particular spatial segmentation, the feature extraction is applied to the segments only:

$$\begin{aligned} E = & fe_{MMCO}(S_{MMCO_{MVR}}) + fe_{SRD}(S_{SRD_{MVR}}) \\ & + fe_{PD}(S_{PD_{MVR}}) \end{aligned} \tag{4}$$

Let there be a segment s_i, where $s_{i,MMCO_{MVR}}$ is the segment in data type MMCO, $s_{i,SRD_{MVR}}$ is the segment in data type SRD, and $s_{i,PD_{MVR}}$ is the segment in data type PD. The extracted features of the segment are e_{s_i}

$$\begin{aligned} e_{s_i} = & fe_{MMCO}(s_{i,MMCO_{MVR}}) + fe_{SRD}(s_{i,SRD_{MVR}}) \\ & + fe_{PD}(s_{i,PD_{MVR}}) \end{aligned} \tag{5}$$

An example of feature extraction is audio transcription, which detects and transcribes spoken words that occur within the segment. Another example is avatar detection, which detects and classifies occurring avatars within the segment.

3.1.3. MVR Feature Fusion

Feature fusion joins multiple features, for example, can merge multiple classification results into a single homogeneous result. The process can be based on rules such as a confidence threshold or feature mapping. Feature fusion improves retrieval by processing, e.g., by retaining significant features.

The *PFMR*does not enforce a single fusion strategy but explicitly allows both early and late feature fusion depending on the configuration of the processing pipeline. When feature fusion is applied directly after the feature extraction stage, integrating heterogeneous features, e.g., common and domain-specific visual concepts, it corresponds to early feature-level fusion. In contrast, when fusion is applied after the data mining step, where higher-level semantic constructs such as dialogs, activities, or inferred events have already been derived, the process constitutes late, decision-level fusion. This design reflects the flexible nature of *PFMR*, in which fusion is not bound to a fixed stage but can be positioned according to the analytical requirements of the use case.

A *MVR Feature Fusion* function ff takes the output of the *MVR Feature Extraction* E, or, because of the flexible connection, the result of the *MVR Data Mining* G'. Formally, it can be described as

$$ff(E), ff(G') \rightarrow E' \subseteq MVR \tag{6}$$

Advanced feature fusion can also improve the results by comparing the extracted features of different models and their confidence. A high confidence threshold of a single model, e.g., 0.9, automatically qualifies the result. If the confidence is below a certain level but two detectors have a confidence above a certain threshold θ, e.g., 0.5, the fusion strategy also retains it.

Let there be the extracted features e_a and e_b with the computed confidences ec_a and ec_b. An applied feature fusion ff can then be defined as follows.

$$ff(e_a, e_b) = \begin{cases} 1 & \text{if } ec_a \geq \theta \text{ and } ec_b \geq \theta \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

Such a feature fusion simply merges the features detected by the feature extraction. Following the previous examples, it joins spoken words and detected avatars for each case.

3.1.4. MVR Data Mining

In contrast to feature fusion, *MVR Data Mining* computes new features based on the extracted features and optional external knowledge. For example, if several people and a ball are visible in a segment, data mining could automatically infer that the segment shows a ball game. Another example is the presence of the objects knife, fork, spoon, and dishes, which indicates a restaurant or place to eat.

A data mining function dm, processing the output of *MVR Feature Extraction* E or *MVR Feature Fusion* E', computes new features G', which are present in the MVR but not contained in R'.

$$dm(E) \text{ or } dm(E') \rightarrow G' \in MVR, G' \notin E', G' \notin E \tag{8}$$

Continuing the previous examples, in this step the transcribed words are analyzed with a toxicity classification. While one could mention that this could also be part of the feature fusion during the transcription, the presented approach is more flexible. For example, in addition to audio transcription, feature extraction could also extract written text from chat messages as features. A subsequent toxicity classification can be performed for all characteristics of any source.

In conclusion, the PFMR provides a guideline for analyzing MVRs. The hypothesis is that the processing framework improves the effectiveness of the analysis.

3.1.5. MVR Indexing

After passing an MVR through a process instance, the results are the sets of features R' and maybe G'. The indexing of the feature sets can be achieved by transforming them into a feature representation.A suitable representation is the MMFG; however, the representation is not required to be a graph-based structure and could be replaced by any other formal feature representation. Formally, the mapping function fm produces the MMFG.

$$R = E' \cup G' \tag{9}$$

$$fm(R) \rightarrow MMFG \tag{10}$$

The features can be mapped in several ways to one or more MMFGs. Ref. [6] describes a single MMFG for the MVR, with all segments at the first level, e.g., temporal segments. All features that appear in a segment are child nodes of the segments. Let MMFG be a $graph(V, E)$ with a set of vertices (nodes) V, the root v_{root}, and a set of edges E.

$$MMFG = (V, E) \tag{11}$$

The described mapping can be formalized as follows:

$$V = \{v_{root}\} \cup V_{segments} \cup V_{features} \tag{12}$$

$$E = \{(v_{root}, v_s) | v_s \in V_{segments}\} \cup \{(s_i, r_j) | rj \in R_{s_i}, s_i \in S\} \tag{13}$$

The resulting feature representation can be stored in the MVR index. A popular way of feature indexing is the use of embeddings. An alternative approach is Graph Codes, which can be generated from the MMFG.

3.2. Modeling Discussion

Several models and tools exist for extracting features from multimedia content. For example, object detection, interaction detection for human–human and human–object interaction, language detection, or transcription. *PFMR* allows the orchestration of existing methods. Continuing the given examples, a process instance following the *PFMR* could use methods as follows. For the *MVR Structural Analysis*, the a black frame detection of FFmpeg [75] can be used to create temporal segments. *MVR Feature Extraction* can be executed by YOLO [76] object detection and Whisper [77] transcription. A simple *MVR Feature Fusion* joins the extracted features. Finally, an *MVR Data Mining* can apply Detoxify [78] as a toxicity classification. This example shows that several existing methods could be

integrated while the framework orchestrates the execution and transfer of results to the index. In addition, individual methods in the process steps can be exchanged, added, or extended.

With regard to the given use scenario of the domain of cybersecurity, the described process instance produces a representation of the MVRs, divided into segments, which contain features of occurrences of insults, a class of toxicity classified by the toxicity classifier, based on extracted spoken words from the audio.

Unlike conventional video analysis pipelines that operate on isolated clips or predefined tasks, *PFMR*introduces a process-level abstraction tailored to the characteristics of metaverse recordings. It enables the coordinated analysis of heterogeneous data types, preserves temporal alignment across modalities, and supports incremental semantic enrichment through feature fusion and data mining. While existing MMIR frameworks provide extensible processing architectures, they do not model metaverse-specific structures such as avatar interactions, continuous session dynamics, or time-aligned composite content. *PFMR*closes this gap by defining a reference model that operationalizes these requirements and enables their integration into retrieval-oriented systems.

The provided framework has its benefits in domain-specific adoption. However, this adoption needs to be done in advance of the processing itself. It does not change dynamically and a change requires a reprocessing. In terms of scalability, the processing is highly scalable, since each asset is processed individually, and, hence, can be done fully in parallel. Adding or modifying elements requires only a processing of the element.

The processing of MVRs potentially includes personal and personal relatable data. With regards to privacy, security, and trustworthyness, these are not particularly addressed in the *PFMR*.

3.3. Summary of Conceptual Design and Modelling

The models presented are *PFMR*, providing an integratable framework. It allows reuse of existing methods for MVR content analyses consisting of any type, MMCO, SRD, and PD. *PFMR* is technology agnostic and can be integrated independently. It allows a domain-related adoption by selecting domain specific structure analysis, feature extraction, feature fusion, and data-mining processes.

A prototypical implementation of the conceptual models is described in the next section.

4. Implementation and Testing

The conceptual models is designed as a reference model which is integratable into existing MMIR models and systems. As described in Section 2.3, GMAF is an MMIR system with a flexible and extensible architecture and is used to demonstrate the integration of *PFMR*, however, it should be able to work with any extensible MMIR system.

The incorporation of *PFMR* into a conventional MMIR architecture necessitates a highly adaptable workflow module for the processing stage, as well as specialized feature extraction procedures tailored to metaverse-specific content. In this context, the presented implementation illustrates how *PFMR* is integrated with GMAF. GMAF offers a configurable workflow environment [43] together with a plugin-based interface for feature extraction components, which can be exploited to realize dedicated feature extraction algorithms. Furthermore, GMAF employs an extensible set of feature fusion mechanisms that combine the heterogeneous features produced by the individual plugins. Through this combination of feature extraction and fusion, a single MMFG representation is ultimately generated for every CMMCO, thereby providing a unified multimodal feature graph for subsequent retrieval and analysis tasks.

As the data structure to store the analysis results, an extensible semantic feature representation can be used, in this case the MMFG.

4.1. Integration of PFMR in GMAF

Storing temporal structure information in the MMFG requires an extension. The data are stored as a time range. Hence, the `Node` object is extended with an attribute of class `TimeRange`, shown in Listing 1. To make `TimeRange` flexible for both full date and time as well as relative time information from a reference point, e.g., the start of a recording, a flexible temporal representation is used.

Listing 1. TimeRange in GMAF.

```
public class TimeRange {
...
}

public class Node
...
private TimeRange timeRange;
public TimeRange getTimeRange() {...}
public void setTimeRange(TimeRange timeRange) {...}
...
}
```

The prototypical implementation of the conceptual *PFMR* model is integrated in GMAF by the GMAF Flows. Visualized in Figure 4, *MVR Structure Analysis* and *MVR Feature Extraction* are implemented as a `GMAF_Plugin`, which uses the media as input and returns a populated MMFG. *MVR Data Mining* and *MVR Feature Fusion* are implemented by a `FeatureFusionStrategy`, which uses MMFGs as input and returns a single MMFG.

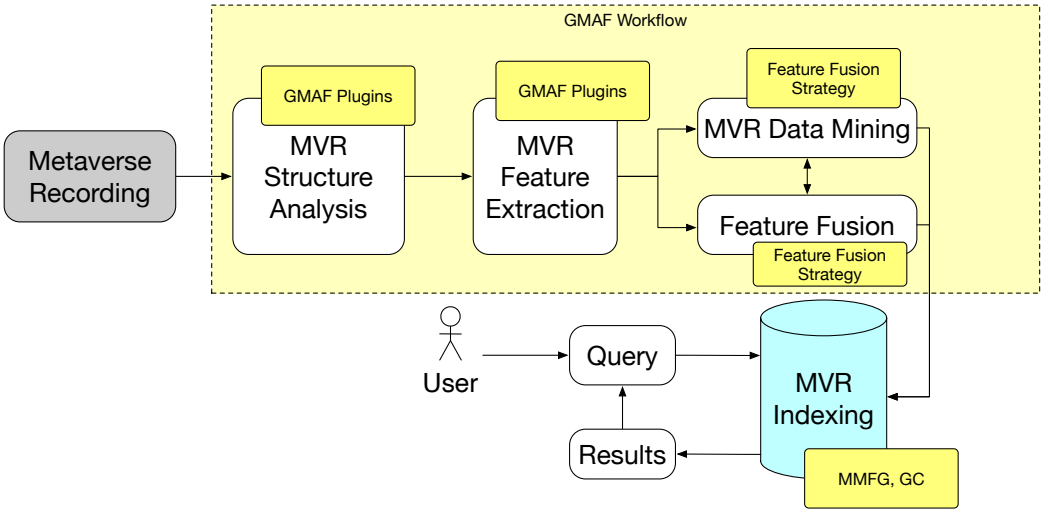


Figure 4. Use of GMAF components for *PFMR* [6].

For the evaluation of *PFMR*, a set of methods was implemented, as described next.

4.2. GMAF Plugins and Fusion Strategies

For evaluation purposes, the prototypical implementation of the *PFMR* models covers the following set of plugins and strategies. In each step, an existing method is used to demonstrate the integratability of the implementation.

4.2.1. Prototypical Integration of MVR Structure Analysis Methods

For *MVR Structure Analysis*, a temporal segmentation was implemented as a `GMAF_Plugin` by leveraging FFmpeg [75]. A scene segmentation with a parameterized threshold of 0.3 creates a `Node` for each segment with the relative time range.

4.2.2. Prototypical Integration of MVR Feature Extraction Methods

For *MVR Feature Extraction*, the established models Faster-RCNN [79] and YOLO were used in a plugin, creating a Node for each detected class with the corresponding TimeRange for the frame of occurrence. The Faster-RCNN implementation is based on Detectron, and the plugin is referred to as objectdetectron. The Yolo implementation is based on the OpenCV Java implementation, referred to as objectdetection.

An MVR-specific avatar detection ADET [53] was used in the same manner. The analysis of speech in the extracted audio stream was implemented using OpenAI Whisper. The extraction produces a list of transcribed text with a time frame. For each given time frame, a `Node` named speech is created for detected audio, `Node` language with the detected language, and nodes with the text of each detected time segment as provided by the transcript. The nodes receive a `TimeRange` corresponding to the time frame reported by Whisper output. The executed tests revealed pitfalls, such as the fact that recordings with no spoken words and silence produce hallucinated text. In rare cases, the language detection fails; in such cases, English is used as a fallback.

4.2.3. Prototypical Integration of MVR Feature Fusion Methods

A late feature fusion strategy termed Scene Clustering takes the temporal segments and groups all other features within the segments, if the time range of the Nodes fall within the segment ranges.

4.2.4. Prototypical Integration of MVR Data Mining Methods

Two data mining classes based on `FusionStrategies` are implemented, the `DialogMiner` to detect dialogs in the transcribed audio and a `ToxicityDetection` to classify the dialogs. The `DialogMiner` depends on the preceeding scene clustering and the accuracy of the temporal segmentation. It takes all nodes with transcripts in a segment, concatenates the sentences, and analyzes them. The analysis is performed by a Spacy-based dialog detection, which checks whether the current sentence starts with a response indicator like “Yes”, “No”, “I’m”, etc. If a dialog is detected, a `Node` named dialog is added to the segment.

Similarly, the `ToxicityDetection` concatenates and analyzes all transcribed text within a segment. It does not extract new features from the source but analyzes the extracted features, i.e., text. The analysis is performed by the text classification model of Detoxify’s model unbiased-toxic-roberta [78]. The classification result contains multiple classes with a probability value, for example, sexual_explicit, insult, or obscene. For all classes above a probability threshold of 0.5, a `Node` is added to the MMFG. Since the model is trained using binary cross-entropy with logits, a confidence threshold of 0.5 corresponds to the canonical decision boundary, at which the predicted class becomes more likely than not, making it a principled and commonly used baseline for filtering low-confidence predictions in a proof-of-concept setting [80].

4.3. Summary of Implementation and Testing

The prototype demonstrates that the implementation of *PFMR* is feasable. Furthermore, it shows that GMAF provides the necessary extension points to implement and

integrate *PFMR*, which confirms its capabilities in terms of integration, extendability, and flexibility.

Following the use scenario, several existing analysis methods could be implemented based on the *PFMR* model. In the next section, we present the results of the evaluation of the prototypical implementation.

5. Evaluation

The purpose of the evaluation is therefore not to demonstrate that *PFMR* outperforms individual feature extractors, but to show that *PFMR* enables the systematic integration of domain-specific analysis steps and makes their impact on retrieval performance explicit, comparable, and reproducible. The evaluation of our conceptual models and prototypical implementations is performed on the 256 Metaverse Records dataset. As presented, the dataset consists of 256 video recordings from 13 different metaverse platforms, covering a wide range of graphical styles and scene characteristics. The videos range from two to five minutes an provide a variety of visible objects, avatars, and interactions in the virtual worlds. In addition, the recordings represent diverse interaction contexts and usage scenarios within the virtual environments, resulting in a highly heterogeneous and representative dataset suitable for the *PFMR* evaluation. Different flows are constructed and evaluated to demonstrate that the identified challenges are addressed. Common metrics Precision@k (P@k) [81] and Recall@k (R@k) [82] are used.

5.1. MVR Structure Analysis

The first experiment addresses the validation of segment-based search for the user with temporal segmentation as the *MVR Structure Analysis*. An example search for different objects produces either a set of full videos or, based on the temporal structure analysis, a set of segments. Without segmentation, the result set includes videos that, at any given point, contain the queried keywords, but not necessarily in the same segment.

Flow: Simple object detection through concatenation of Fast-RCNN-MSCOCO and YOLOv3-MSCOCO, followed by scene clustering. The flow configuration is shown in Listing 2.

Listing 2. Flow configuration for scene clustering.

```
<mmfg processor="ffsegment,objectdetectron,objectdetection"/>
<fusion processor="scene-cluster"/>
```

Query: The example queries for query by keyword 'boat, person' and 'person, cup, fork, knife' are used for the evaluation. Based on the index produced by the described flow, the query produces the results shown in Table 1 where *rel* denotes the number of relevant items in the collection and *tp* denotes the true positives. The number of false positives in the result set is affected by the object detection precision, as demonstrated in Figures 5 and 6.

Table 1. Comparison of different keyword queries for objects.

Query	*rel*	Result@20	*tp*	P@k	R@k
boat, person	13	20	5	0.25	0.38
person, cup, fork, knife	6	20	5	0.25	0.3



Figure 5. False classification of structural object as knife with 97% confidence.



Figure 6. False detection of a hand as knife with 57% confidence.

5.2. Enhancing Retrieval for a Domain-Specific Usage Scenario

We conducted an additional experiment to determine the efficacy of specializing the workflow for the analysis process. The original workflow employs basic *MVR Feature Extraction*, i.e., object detection steps. By systematically enhancing the flows, we aimed

to create more focused and efficient features in the index. The exemplary use case is cybersecurity, where an investigator searches the collection.

Since GMAF provides Query by Keyword (QbK) and Query by Similarity (QbS), these queries are used for the example scenario.

5.2.1. Flow 1

The first flow is based on the previous experiments. The object detection based on the MS COCO [83] classes extracts detected object features, e.g., person. The investigator might search for situations where people use toxic language and select from the available keywords, e.g., person. The keyword knife could also be an option, but appears frequently in games, e.g., Minecraft, or as cutlery. Additionally, the investigator uses an example, which is the dataset the video `13_Video_16.mp4`, which contains profane chitchat [84].

The flow configuration is listed in Listing 3 and the result is shown in Table 2.

Listing 3. Configuration Flow 1.

```
<!-- PFMR 2 -->
<mmfg processor="ffsegment,objectdetectron,objectdetection"/>
<fusion processor="scene-cluster"/>
```

Table 2. Comparison of different domain-specific *PFMR*configurations.

Flow	Search Keyword	rel	k	tp	P@k	R@k
1	person	1285 *	20	5	0.25	0.00
2	avatar	1615 *	20	18	0.9	0.01
3	speech	318 *	20	5	0.25	0.02
	speech,avatar	309 *	20	5	0.25	0.02
4	dialog	106 *	20	10	0.5	0.09
5	toxicity	33 *	20	17	0.85	0.52
	avatar,dialog,toxicity	33 †	20	20	1	1
	avatar,dialog,toxicity	18	20	18	0.9	1

* not all relevant items have been validated. † count of toxicity occurrences.

5.2.2. Flow 2

A more metaverse-specific object detection method is the avatar detector ADET [53]. ADET replaces object detection in the second flow configuration. The flow configuration is listed in Listing 4 and the result is shown in Table 2.

Listing 4. Configuration Flow 2.

```
<!-- PFMR 3 -->
<mmfg processor="ffsegment,avatardetector"/>
<fusion processor="scene-cluster"/>
```

5.2.3. Flow 3

Since hate speech is a problem in online communication, the investigator might be interested in segments where avatars are talking. This distinguishes segments of interest from others, e.g., those where avatars are just moving around. The transcription process allows detection of the language, but in some cases this is not possible if no or few words are spoken. In such cases, the default is English. In cases with no words, Whisper produces hallucinated output, which is a known problem [85]. The flow configuration is listed in Listing 5 and the result is shown in Table 2.

Listing 5. Configuration Flow 3.

```
<!-- PFMR 4 -->
<mmfg processor="ffsegment,avatardetector,whisperplugin"/>
<fusion processor="scene-cluster"/>
```

5.2.4. Flow 4

Regarding speech, Whisper translates all spoken words, whether it is from a song in the background or commands played in the scene. The `DialogMiner` is added to mine dialogs from the transcript `Nodes`. Hence, the investigator can search for dialogs only, or in combination with other keywords. The flow configuration is listed in Listing 6 and the result is shown in Table 2.

Listing 6. Configuration Flow 4.

```
<!-- PFMR 5 -->
<mmfg processor="ffsegment,avatardetector,whisperplugin"/>
<fusion processor="scene-cluster"/>
<fusion processor="dialogminer"/>
```

5.2.5. Flow 5

Since the number of dialogs in a huge collection is significant, the analysis can add a toxicity classification of dialogs. This enables the toxicity class-specific keywords.

The flow configuration is listed in Listing 7 and the result is shown in Table 2. Comparing the results for the query *toxicity*, P@20 = 0.85 and *avatar, dialog, toxicity*, P@20 = 1.0, the fusion of domain-specific features yields higher precision, indicating that progressively adapting the analysis pipeline to the target domain improves retrieval effectiveness.

Listing 7. Configuration Flow 5.

```
<!-- PFMR 6 -->
<mmfg processor="ffsegment,objectdetection,avatardetector,
    ↪ whisperplugin"/>
<fusion processor="scene-cluster"/>
<fusion processor="dialogminer"/>
<fusion processor="toxicityminer"/>
```

5.2.6. Similarity Search Study

Based on the toxicity classification, all eight videos (not-segmented) with the occurrence of *obscene* speech, can be identified. Table 3 shows the mean Average Precision (mAP) [86] for the similarity search using each video as a query. As noticeable, the mAP improved with the specialized pipeline. A Wilcoxon Signed-Rank test [87] was conducted to compare the average precisions between the queries of Flow 2 and Flow 5. Results indicated that the APs were significantly different, signed-rank sum $T = 5$, significance level $\alpha = 0.05$, effective sample size $n = 7$ (8 subtracted by 1, because one query shows a difference of 0), indicating a relevant effect.

Table 3. Comparison of mAP@20.

	Flow 2	**Flow 3**
mAP@20	0.118	0.313

5.2.7. Interpretation of Evaluation Results

The experiments revealed that the specialized workflow not only preserved the accuracy and integrity of the analysis but also significantly reduced the time and resources required for completion. This empirical evidence strongly suggests that the streamlined approach could serve as a more effective standard for future applications of the procedure. Furthermore, the comparative analysis of different workflows demonstrated that multiple methods could be effectively integrated into the analytical framework.

In the context of domain-specific analysis, our results indicate that the *PFMR* methodology delivers superior outcomes in meeting user needs by improving the P@K for similarity search, while facilitating the extraction of more detailed features, i.e., keywords. As Table 2 demonstrates, the P@K values exhibit a positive correlation with increasing levels of domain-specific analysis.

The observed increase in retrieval performance with progressively specialized workflows is, to some extent, expected, as domain-specific feature extractors naturally provide more discriminative information. However, the contribution of *PFMR* does not lie in the individual feature extraction methods themselves, but in the structured orchestration of heterogeneous analysis steps. *PFMR* provides a formalized process model that enables the systematic integration, comparison, and extension of analysis components. The evaluation therefore demonstrates not merely improved retrieval performance, but the practical value of *PFMR* as an architectural framework that supports controlled experimentation and domain-specific adaptation. The improvements observed across flows illustrate how *PFMR* enables such adaptations in a reproducible and extensible manner.

Compared to an application specific method, it can be argued that the generic framework introduces an computational overhead. Given that the processing of the 256 MVRs for Flow 2 and Flow 6 both required around 38 h using an i5-8500 CPU and an RTX 3090Ti GPU, it can be stated that the majority of the computation is required for the object detection methods and the transition between process steps is negligible. Not part of the measurement was the reduction of computational effort through spatial segmentation of certain recording areas, such as chat windows.

5.2.8. Limitations

The 256 MVR dataset, while the largest known metaverse-specific dataset, is relatively small compared to standard MMIR benchmarks, yet it provides a variety of 20 metaverse virtual worlds and over 12 h of video. The evaluation focuses on a single domain. While this demonstrates the frameworks adaptability, generalizability to other domains, it remains to be validated. Results depend on the accuracy of the underlying tools, e.g., Whisper hallucinations, object detection false positives as shown in Figures 5 and 6 *PFMR* orchestrates these tools but does not resolve their inherent limitations. The experiments use screen recordings only, not the full CMMCO data types, i.e., SRD and PD. This reflects current data availability but leaves multimodal fusion partially unexplored.

5.3. Summary of Evaluation

The evaluation demonstrated the flexibility and usefulness of the *PFMR* for MVR content analysis for MVR information retrieval. The demonstrated domain-specific flows improve the retrieval results by adding relevant feature extraction and fusion, which supports the thesis that integrative and flexible analysis is beneficial for MVR content analysis and indexing. With regard to the challenges, *PFMR* provides a foundation for analyzing and indexing of MVRs for MVR information retrieval.

6. Conclusions and Outlook

This paper has shown that the *PFMR* is a solid foundation for content analysis of MVRs. The *PFMR* can integrate solitary methods for content analysis and orchestrate them to achieve an effective indexing and querying. The flexibility of *PFMR* enables domain-specific analysis and indexing. A prototypical implementation demonstrated the integration of *PFMR* into existing MMIR architectures using GMAF, and evaluation results confirmed its ability to be adapted to domain-specific scenarios such as cybersecurity. The findings highlight the potential of *PFMR* to support structured and adaptable content analysis in various metaverse use cases.

Further research on temporal analysis is required to improve temporal segmentation. The demonstrated approach of integrating time in the MMFG is sufficient for the demonstration of *PFMR*, but lacks support for temporal elements in queries, e.g., when I drove the car *after* I played the marble game.

The analysis and indexing of metaverse recordings raise important ethical and privacy considerations that extend beyond the technical scope of *PFMR*. Since MVRs may contain sensitive personal data such as speech, behavior, or interaction patterns, their automated processing and indexing introduce risks related to privacy, consent, and unintended secondary use. This is particularly relevant when multimodal features are fused and semantically enriched, as higher-level inferences may be derived that were not explicitly captured in the original recordings.

Author Contributions: Conceptualization and methodology: P.S., S.W. and M.L.H. Software, validation, formal analysis, investigation, resources, data curation, writing: P.S. Review, editing, and supervision: S.W., I.F. and M.L.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Informed Consent Statement: Not applicable.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors have no conflicts of interest to declare that are relevant to the content of this article.

References

1. Austin, D. 2023 Internet Minute Infographic. eDiscovery Today by Doug Austin. April 2023. Available online: https://ediscoverytoday.com/2023/04/20/2023-internet-minute-infographic-by-ediscovery-today-and-ltmg-ediscovery-trends/ (accessed on 10 February 2024).
2. Richter, F. Smartphones Cause Photography Boom. Statista. August 2017. Available online: https://www.statista.com/chart/10913/number-of-photos-taken-worldwide/ (accessed on 18 October 2021).
3. National Library of Medicine. What Is Open-i? Available online: https://openi.nlm.nih.gov/faq#collection (accessed on 8 September 2022).
4. Jenik, C. A Minute on the Internet in 2021. Statista. July 2022. Available online: https://www.statista.com/chart/25443/estimated-amount-of-data-created-on-the-internet-in-one-minute/ (accessed on 17 October 2022).
5. Raieli, R. *Multimedia Information Retrieval: Theory and Techniques*; Chandos Publishing: Oxford, UK, 2013.
6. Steinert, P.; Wagenpfeil, S.; Frommholz, I.; Hemmje, M.L. Integration of Metaverse Recordings in Multimedia Information Retrieval. In Proceedings of the ICSCA 2024, Bali Island, Indonesia, 1–3 February 2024; pp. 137–145. [CrossRef]
7. Lee, J.Y. A Study on Metaverse Hype for Sustainable Growth. *Int. J. Adv. Smart Converg.* **2021**, *10*, 72–80. [CrossRef]
8. Nast, C. The Metaverse Is Coming. Wired UK. February 2021. Available online: https://www.wired.co.uk/article/metaverse (accessed on 30 October 2022).
9. Alsop, T. Virtual Reality (VR)-Statistics & Facts. Statista. August 2023. Available online: https://www.statista.com/topics/2532/virtual-reality-vr/ (accessed on 22 August 2023).

10. Hackl, C. Social VR, Facebook Horizon and the Future of Social Media Marketing. Forbes. August 2020. Available online: https://www.forbes.com/sites/cathyhackl/2020/08/30/social-vr-facebook-horizon--the-future-of-social-media-marketing/ (accessed on 15 May 2023).
11. Meta Platforms, Inc. Horizon Worlds | Virtual Reality Worlds and Communities. 2023. Available online: https://www.meta.com/horizon-worlds/ (accessed on 28 October 2023).
12. Perri, L. What's New in the 2022 Gartner Hype Cycle for Emerging Technologies. Gartner. August 2022. Available online: https://www.gartner.com/en/articles/what-s-new-in-the-2022-gartner-hype-cycle-for-emerging-technologies (accessed on 22 October 2024).
13. Perri, L. 4 Exciting New Trends in the Gartner Emerging Technologies Hype Cycle. Gartner. August 2023. Available online: https://www.gartner.com/en/articles/what-s-new-in-the-2023-gartner-hype-cycle-for-emerging-technologies (accessed on 22 October 2024).
14. KZero Worldwide. Exploring the Q1 24' Metaverse Radar Chart: Key Findings Unveiled-KZero Worldswide. February 2024. Available online: https://web.archive.org/web/20240206134622/https://kzero.io/2024/02/06/2633/ (accessed on 2 March 2026).
15. Statista. U.S. Top Non-Gaming Activities Within Video Games by Engagement 2023. Statista. October 2023. Available online: https://www.statista.com/statistics/1275467/us-gamers-participating-select-non-gaming-activities-within-video-games/ (accessed on 22 October 2024).
16. Statista. U.S. Metaverse Familiarity 2024. Statista. May 2024. Available online: https://www.statista.com/statistics/1290378/unites-states-adults-familiarity-with-the-metaverse/ (accessed on 22 October 2024).
17. RAUM Virtual Collaboration GmbH. RAUM | a Metaverse Platform for Organisations & Companies. August 2023. Available online: https://www.raum.app/ (accessed on 22 October 2024).
18. Uhl, J. C.; Nguyen, Q.; Hill, Y.; Murtinger, M.; Tscheligi, M. xHits: An Automatic Team Performance Metric for VR Police Training. In *Proceedings of the 2023 IEEE International Conference on Metrology for eXtended Reality, Artificial Intelligence and Neural Engineering (MetroXRAINE)*; IEEE: Milano, Italy, 2023; pp. 178–183. [CrossRef]
19. Li, X.; Yalcin, B.C.; Christidi-Loumpasefski, O.-O.; Martinez Luna, C.; Hubert Delisle, M.; Rodriguez, G.; Zheng, J.; Olivares Mendez, M.A. Exploring NVIDIA Omniverse for Future Space Resources Missions. Space Resources Week 2022. May 2022. Available online: https://hdl.handle.net/10993/51217 (accessed on 3 May 2022).
20. Steinert, P.; Wagenpfeil, S.; Frommholz, I.; Hemmje, M.L. Towards the Integration of Metaverse and Multimedia Information Retrieval. In *Proceedings of the 2023 IEEE International Conference on Metrology for eXtended Reality, Artificial Intelligence and Neural Engineering (MetroXRAINE)*; IEEE: Milano, Italy, 2023; pp. 581–586. [CrossRef]
21. Wikipedia. Massively Multiplayer Online Game. Wikipedia. Page Version ID: 1138213454. February 2023. Available online: https://en.wikipedia.org/w/index.php?title=Massively_multiplayer_online_game&oldid=1138213454 (accessed on 3 May 2024).
22. Nunamaker, J.F., Jr.; Chen, M.; Purdin, T.D. Systems development in information systems research. *J. Manag. Inf. Syst.* **1990**, *7*, 89–106. [CrossRef]
23. Stephenson, N. *Snow Crash*; Bantam Books: New York, NY, USA, 1992.
24. Zyda, M. Let's Rename Everything "the Metaverse!". *Computer* **2022**, *55*, 124–129. [CrossRef]
25. Mason, R. Is the Next Generation of the Internet Ready to Be Tested? Forbes. Section: Innovation. April 2022. Available online: https://www.forbes.com/sites/forbestechcouncil/2022/04/29/is-the-next-generation-of-the-internet-ready-to-be-tested/ (accessed on 7 March 2023).
26. Mystakidis, S. Metaverse. *Encyclopedia* **2022**, *2*, 486–497. [CrossRef]
27. Smart, J.; Cascio, J.; Paffendorf, J. Metaverse Roadmap Overview. 2007. Available online: https://web.archive.org/web/20230202145251/https://www.metaverseroadmap.org/overview/ (accessed on 20 February 2023).
28. Ball, M. *The Metaverse: And How It Will Revolutionize Everything*, 1st ed.; Liveright Publishing Corporation, a Division of W.W. Norton & Company: New York, NY, USA, 2022.
29. Metaversed Consulting. The Metaverse Reaches 600m Monthly Active Users. October 2023. Available online: https://metaversed.webflow.io/blog/the-metaverse-reaches-600m-monthly-active-users (accessed on 8 October 2023).
30. Gartner Inc. Gartner Predicts 25% of People Will Spend at Least One Hour per Day in the Metaverse by 2026. Gartner. July 2022. Available online: https://www.gartner.com/en/newsroom/press-releases/2022-02-07-gartner-predicts-25-percent-of-people-will-spend-at-least-one-hour-per-day-in-the-metaverse-by-2026 (accessed on 28 October 2022).
31. Statista. Metaverse Worldwide Market Forecast. Statista. March 2024. Available online: https://www.statista.com/outlook/amo/metaverse/worldwide (accessed on 28 April 2025).
32. Liu, H.; Wang, Z.; Mousas, C.; Kao, D. Virtual Reality Racket Sports: Virtual Drills for Exercise and Training. In *Proceedings of the 2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*; IEEE: Porto de Galinhas, Brazil, 2020; pp. 566–576. [CrossRef]

33. Zhang, Y.; Tsai, S.-B. Application of Adaptive Virtual Reality with AI-Enabled Techniques in Modern Sports Training. *Mob. Inf. Syst.* **2021**, *2021*, 6067678. [CrossRef]
34. Zheng, Z.; Li, T.; Li, B.; Chai, X.; Song, W.; Chen, N.; Zhou, Y.; Lin, Y.; Li, R. Industrial Metaverse: Connotation, Features, Technologies, Applications and Challenges. In *Methods and Applications for Modeling and Simulation of Complex Systems*; Fan, W., Zhang, L., Li, N., Song, X., Eds.; Springer Nature: Singapore, 2022; pp. 239–263. [CrossRef]
35. Stothard, P.; Ryan, P.; Kurata, T.; Stapleton, D. Towards a mining metaverse. *Min. Technol. Trans. Inst. Min. Metall.* **2024**, *133*, 142–161. [CrossRef]
36. Fantozzi, I.C.; Olhager, J.; Johnsson, C.; Schiraldi, M.M. Guiding organizations in the digital era: Tools and metrics for success. *Int. J. Eng. Bus. Manag.* **2025**, *17*, 1–16. [CrossRef]
37. Bestie Let's Play. Wir Verbringen Einen Herbsttag Mit der Großfamilie!!/Roblox Bloxburg Family Roleplay Deutsch. October 2022. Available online: https://www.youtube.com/watch?v=sslXNBKeqf0 (accessed on 28 February 2023).
38. NVIDIA Corp. NVIDIA DRIVE Sim. NVIDIA Developer. 2024. Available online: https://developer.nvidia.com/drive/simulation (accessed on 7 June 2024).
39. INTERPOL. Metaverse—A Law Enforcement Perspective. Whitepaper. January 2024. Available online: https://www.interpol.int/News-and-Events/News/2024/Grooming-radicalization-and-cyber-attacks-INTERPOL-warns-of-Metacrime (accessed on 10 February 2024).
40. Steinert, P.; Mischkies, J.; Wagenpfeil, S.; Frommholz, I.; Hemmje, M. Information Need in Metaverse Recordings—A Field Study. *Inf. Res. Int. Electron. J.* **2026**, *31*, 226–246. [CrossRef]
41. Wikipedia. AltspaceVR. Wikipedia. Page Version ID: 1248210194. September 2024. Available online: https://en.wikipedia.org/w/index.php?title=AltspaceVR&oldid=1248210194 (accessed on 28 November 2024).
42. Engage XR Holdings Plc. Recordings | ENGAGE. February 2023. Available online: https://docs.engagevr.io/engage/advanced-features/recordings (accessed on 23 December 2024).
43. Wagenpfeil, S.; Kevitt, P.M.; Hemmje, M. Smart Multimedia Information Retrieval. *Analytics* **2023**, *2*, 198–224. [CrossRef]
44. Moshfeghi, Y. Role of Emotion in Information Retrieval. Ph.D. Thesis, University of Glasgow, Glasgow, UK, January 2012. Available online: https://theses.gla.ac.uk/3118/ (accessed on 7 April 2023).
45. Steinert, P.; Wagenpfeil, S.; Frommholz, I.; Hemmje, M.L. Uses of Metaverse Recordings in Multimedia Information Retrieval. *Multimedia* **2025**, *1*, 2. [CrossRef]
46. Uhl, J.C.; Murtinger, M.; Zechner, O.; Tscheligi, M. Threat Assessment in Police VR Training: Multi-Sensory Cues for Situation Awareness. In *Proceedings of the 2022 IEEE International Conference on Metrology for Extended Reality, Artificial Intelligence and Neural Engineering (MetroXRAINE)*; IEEE: Rome, Italy, 2022; pp. 432–437. [CrossRef]
47. NVIDIA Corp. NVIDIA/vr-Capture-Replay. NVIDIA Corporation. Original-Date: 2024-03-11T16:04:24Z. July 2024. Available online: https://github.com/NVIDIA/vr-capture-replay (accessed on 24 August 2024).
48. Abdari, A.; Falcon, A.; Serra, G. Metaverse Retrieval: Finding the Best Metaverse Environment via Language. In *Proceedings of the 1st International Workshop on Deep Multimodal Learning for Information Retrieval*; ACM: Ottawa, ON, Canada, 2023; pp. 1–9. [CrossRef]
49. Steinert, P.; Wagenpfeil, S.; Hemmje, M.L. 256-Metaverse Records Dataset. December 2023. Available online: https://patricksteinert.de/256-metaverse-records-dataset/ (accessed on 3 January 2024).
50. Kowald, D.; Scher, S.; Pammer-Schindler, V.; Müllner, P.; Waxnegger, K.; Demelius, L.; Fessl, A.; Toller, M.; Mendoza Estrada, I.G.; Šimić, I.; et al. Establishing and evaluating trustworthy AI: Overview and research challenges. *Front. Big Data* **2024**, *7*, 1467222. [CrossRef] [PubMed]
51. Coalition for Content Provenance and Authenticity. Content Credential C2PA Technical Specification Version 2.3. January 2026. Available online: https://spec.c2pa.org/specifications/specifications/2.2/specs/C2PA_Specification.html (accessed on 7 February 2026).
52. Davies, J.; Studer, R.; Warren, P. *Semantic Web Technologies: Trends and Research in Ontology-Based Systems*; John Wiley & Sons: Chichester, UK; Hoboken, NJ, USA, 2006.
53. Becker, F.; Steinert, P.; Wagenpfeil, S.; Hemmje, M.L. Avatar Detection in Metaverse Recordings. *Virtual Worlds* **2024**, *3*, 459–479. [CrossRef]
54. Rüger, S.; Marchionini, G. *Multimedia Information Retrieval*; Springer: Berlin/Heidelberg, Germany, 2010.
55. Zou, Z.; Chen, K.; Shi, Z.; Guo, Y.; Ye, J. Object Detection in 20 Years: A Survey. *Proc. IEEE* **2023**, *111*, 257–276. [CrossRef]
56. Ebadollahi, S.; Chang, S.-F.; Wu, H.D.; Takoma, S. *Echocardiogram Video Summarization*; SPIE: Bellingham, WA, USA, 2001.
57. Sidiropoulos, P.; Mezaris, V.; Kompatsiaris, I.; Meinedo, H.; Bugalho, M.; Trancoso, I. Temporal Video Segmentation to Scenes Using High-Level Audiovisual Features. *IEEE Trans. Circuits Syst. Video Technol.* **2011**, *21*, 1163–1177. [CrossRef]
58. Li, C.; Li, X.; Rahaman, M.; Li, X.; Sun, H.; Zhang, H.; Zhang, Y.; Li, X.; Wu, J.; Yao, Y.; et al. A Comprehensive Review of Computer-aided Whole-slide Image Analysis: From Datasets to Feature Extraction, Segmentation, Classification, and Detection Approaches. *arXiv* **2021**, arXiv:2102.10553[cs]. http://arxiv.org/abs/2102.10553. [CrossRef]

59. Wirth, R.; Hipp, J. *CRISP-DM: Towards a Standard Process Model for Data Mining*; Practical Application Company: Manchester, UK, 2000.
60. Hu, W.; Xie, N.; Li, L.; Zeng, X.; Maybank, S. A Survey on Visual Content-Based Video Indexing and Retrieval. *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.* **2011**, *41*, 797–819. [CrossRef]
61. *Multimedia Information Retrieval and Management: Technological Fundamentals and Applications*; Springer: Berlin/Heidelberg, Germany, 2003.
62. Wu, H.; Chen, K.; Liu, H.; Zhuge, M.; Li, B.; Qiao, R.; Shu, X.; Gan, B.; Xu, L.; Ren, B. NewsNet: A Novel Dataset for Hierarchical Temporal Segmentation. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023.
63. Naik, V.; Rathod, G. An Algorithm for Retrieval of Significant Events Near Goal Post from Soccer Videos Using Fuzzy Systems. *Int. J. Emerg. Technol. Adv. Eng.* **2025**, *3*, 808–814.
64. Liu, M.; Nie, L.; Wang, Y.; Wang, M.; Rui, Y. A Survey on Video Moment Localization. *Acm Comput. Surv.* **2023**, *55*, 1–37. [CrossRef]
65. Yeh, M.-C.; Cheng, K.-T. Video copy detection by fast sequence matching. In *Proceedings of the ACM International Conference on Image and Video Retrieval*; ACM: Santorini, Greece, 2009; pp. 1–7. [CrossRef]
66. Cavaliere, D.; Greco, L.; Ritrovato, P.; Senatore, S. A knowledge-based approach for video event detection using spatio-temporal sliding windows. In *Proceedings of the 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*; IEEE: Lecce, Italy, 2017; pp. 1–6. [CrossRef]
67. Karbalaie, A.; Abtahi, F.; Sjöström, M. Event detection in surveillance videos: A review. *Multimed. Tools Appl.* **2022**, *81*, 35463–35501. [CrossRef]
68. Chen, C.-M.; Chen, L.-H. Novel framework for sports video analysis: A basketball case study. In *Proceedings of the 2014 IEEE International Conference on Image Processing (ICIP)*; IEEE: Paris, France, 2014; pp. 961–965. [CrossRef]
69. Lin, T.; Aouididi, A.; Zhu-Tian, C.; Beyer, J.; Pfister, H.; Wang, J.-H. VIRD: Immersive Match Video Analysis for High-Performance Badminton Coaching. *IEEE Trans. Vis. Comput. Graph.* **2024**, *30*, 458–468. [CrossRef]
70. Xu, C.; Wang, J.; Lu, H.; Zhang, Y. A Novel Framework for Semantic Annotation and Personalized Retrieval of Sports Video. *IEEE Trans. Multimed.* **2008**, *10*, 421–436. [CrossRef]
71. Liu, C.-C.; Lin, Q.-F. A Scalable Video Analysis Framework for Character Identification and Statistical Profiling. In *Proceedings of the 2025 IEEE 14th Global Conference on Consumer Electronics (GCCE)*; IEEE: Osaka, Japan, 2025; pp. 501–504. [CrossRef]
72. *User Centered System Design: New Perspectives on Human-Computer Interaction*; L. Erlbaum Associates: Hillsdale, NJ, USA, 1986.
73. Spolaôr, N.; Lee, H.D.; Takaki, W.S.R.; Ensina, L.A.; Coy, C.S.R.; Wu, F.C. A systematic review on content-based video retrieval. *Eng. Appl. Artif. Intell.* **2020**, *90*, 103557. [CrossRef]
74. Snoek, C.G.M.; Worring, M. Concept-Based Video Retrieval. *Found. Trends® Inf. Retr.* **2008**, *2*, 215–322. [CrossRef]
75. FFmpeg Project. FFmpeg. 2024. Available online: https://ffmpeg.org/ (accessed on 9 October 2023).
76. Wang, C.-Y.; Bochkovskiy, A.; Liao, H.-Y.M. YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors. In *Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Vancouver, BC, Canada, 2023; pp. 7464–7475. [CrossRef]
77. OpenAI. Whisper. OpenAI. original-date: 2022-09-16T20:02:54Z. October 2024. Available online: https://github.com/openai/whisper (accessed on 6 October 2024).
78. Hanu, L.; Unitary Team. Detoxify. tex.howpublished: Github. 2020. Available online: https://github.com/unitaryai/detoxify (accessed on 24 August 2024).
79. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef]
80. Narasimhan, H.; Agarwal, S. On the relationship between binary classification, bipartite ranking, and binary class probability estimation. In *Advances in Neural Information Processing Systems*; Burges, C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2013. Available online: https://proceedings.neurips.cc/paper_files/paper/2013/file/05311655a15b75fab86956663e1819cd-Paper.pdf (accessed on 30 January 2026).
81. Liu, T.-Y. Learning to Rank for Information Retrieval. *Found. Trends® Inf. Retr.* **2007**, *3*, 225–331. [CrossRef]
82. Cakaloglu, T.; Szegedy, C.; Xu, X. Text Embeddings for Retrieval from a Large Knowledge Base. *arXiv* **2019**, arXiv:1810.10176. http://arxiv.org/abs/1810.10176. [CrossRef]
83. Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common objects in context. In *Computer Vision—ECCV 2014*; Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T., Eds.; Springer International Publishing: Cham, Switzerland, 2014; pp. 740–755.
84. Merriam Webster. Definition of CHITCHAT. October 2024. Available online: https://www.merriam-webster.com/dictionary/chitchat (accessed on 22 October 2024).

85. Koenecke, A.; Choi, A.S.G.; Mei, K.X.; Schellmann, H.; Sloane, M. Careless Whisper: Speech-to-Text Hallucination Harms. In *Proceedings of the The 2024 ACM Conference on Fairness, Accountability, and Transparency*; ACM: Rio de Janeiro, Brazil, 2024; pp. 1672–1681. [CrossRef]
86. Rahman, M.M.; Roy, C.K. TextRank based search term identification for software change tasks. In *Proceedings of the 2015 IEEE 22nd International Conference on Software Analysis, Evolution, and Reengineering (SANER)*; IEEE: Montreal, QC, Canada, 2015; pp. 540–544. [CrossRef]
87. Smucker, M.D.; Allan, J.; Carterette, B. A comparison of statistical significance tests for information retrieval evaluation. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*; ACM: Lisbon, Portugal, 2007; pp. 623–632. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.