Enhancing process understanding through multimodal data analysis and extended reality

Aleksandar Gavric

Business Informatics, TU Wien, Favoritenstrasse 9-11/194-3, 1040 Vienna, Austria

Abstract

The significance of process mining lies in its ability to enable organizations to gain insights and improve efficiency by analyzing vast amounts of data generated in IT supported processes. Process mining often falls short when it comes to understanding manual processes, as it primarily captures the *what* but not the *how* of such activities. We see great potential in the growth of videos capturing various processes, which can lead to a rich source of data for process understanding and analysis, enabling a more comprehensive insight into manual workflows. However, extracting actionable insights from these videos poses significant challenges due to their unstructured nature. This doctoral thesis is about an approach that combines multimodal data analysis and extended reality (XR) techniques to enhance business process understanding. By integrating visual, textual, and audio information from videos, the proposed solution facilitates comprehensive analysis of processes, facilitating process mining, monitoring, and guidance. To address the complexities arising from an amalgamation of rich data sources, we propose three primary research objectives: (1) evaluating methods for mining relevant process information from these multimodal data sources, particularly video; (2) exploring the integration of XR technologies with enriched event logs to foster an immersive, interactive data visualization experience and accurate domain-specific modeling; and (3) determining the influence of integrating these technologies on the interpretability of process mining results. Ultimately, our study explores the way for a specialized, comprehensive approach to process mining, harnessing the power of XR for enriched event log analysis. To demonstrate the effectiveness of the proposed approach, the thesis elaborates applications in various domains, including manufacturing, logistics, healthcare, and beyond.

Keywords

Multimodal data analysis, Extended reality, Process understanding, Process monitoring, Process guidance

1. Introduction

In an era where data is both ubiquitous and diverse, we envision a significant promise in the proliferation of rich data sources documenting diverse processes, which can serve as a valuable data for comprehending and dissecting processes in-depth, affording a broader comprehension of manual workflows. One such type of data source is video, which is rich in content but also inherently complex and high-dimensional. Process discovery goal is to take an event log containing example behaviors and create a process model that adequately describes the underlying process [1]. Despite the omnipresence of event logs data, most organizations diagnose problems based on fiction rather than facts [2], and there are limitations in what process-

Companion Proceedings of the 16th IFIP WG 8.1 Working Conference on the Practice of Enterprise Modeling and the 13th Enterprise Design and Engineering Working Conference, November 28 – December 1, 2023, Vienna, Austria

aleksandar.gavric@tuwien.ac.at (A. Gavric)

0009-0005-1243-7722 (A. Gavric)

mining practitioners tend to use actively [3]. Existing techniques for process understanding have traditionally relied on manual observation, interpretation, and note-taking. These techniques are not only time-consuming but are also prone to human error and inconsistency. The process mining field's growth and the emergence of challenges faced by analysts, as identified through comprehensive [4, 5] interviews and surveys, underscore the need for enhanced support and research at multiple levels, including individual, technical, and organizational, to fully leverage process mining's potential in competitive business environments.

The rapid advancements in multimodal data analysis and extended reality technologies present promising opportunities for the automatic extraction of valuable process-related information. Various studies show that interdisciplinary research field of conceptual modeling and artificial intelligence gains mutual benefits [6, 7, 8]. However, despite these advancements, the integration of multimodal data, particularly videos, into process analysis and guidance remains underexplored and lacks effective approaches.

In the past, we saw how adding additional modality to process mining showed positive results. Some studies showed that in fields such as humanities, social sciences and medicine where workers follow processes and log their execution manually in textual forms instead, we can achieve process discovery results that are very satisfactory with 88% correctly discovered activities [9]. However, other studies showed there is a considerable gap of research into the semantic aspects of process model text labels and natural language descriptions [10] which can be enhanced by attaching more modalities as descriptors and those modalities need to be mineable. Even on a scale of a meta-model for representing all aspects of the digital enterprise, adding the notion of constraints and modalities can give modelers the option to add more precision to their models where needed [11]. Next to, e.g., an enterprise's operational capabilities, enterprise's modeling capabilities will become an increasingly important foundational capability of enterprises, and the challenge is to further improve these modeling capabilities by means of tools, modeling languages, and associated processes while balancing the return on modeling effort [12].

There is an imperative need to develop robust, scalable, and versatile frameworks that can seamlessly analyze multimodal data and translate them into comprehensive process understanding and guidance. This doctoral dissertation revolves around a strategy that connects the analysis of multimodal data with extended reality (XR) methods to elevate the comprehension of business processes.

The remainder of this paper is structured as follows. In Section 2, we discuss problem formulation. Following that, in Section 3, we explore the landscape of related work. Section 4 is dedicated to an examination of the planned research strategy, encompassing our proposed solution design and the pertinent techniques involved. Moving forward to Section 5, we carefully dissect the evaluation plans of our solution design and identify the key audience and beneficiaries of this thesis. In addition, we highlight and categorize the contributions of this thesis into three organized topics. Given these intended contributions, Section 6 discusses the actual research plan. Finally, in Section 7, we draw our paper to a conclusion.

2. Problem formulation

The age of digital transformation has brought forward the prominence of complex processes in various industries. Nevertheless, gaining a profound understanding of these processes, especially from multimodal data sources like videos, remains an unresolved challenges. It is noted that in many organizations, documentation of process knowledge is scattered around various process information sources which introduces considerable problems [13], but we are introducing a new concept of information fragmentation through various modalities. To the best of our knowledge, no prior research has addressed this specific topic. Effective data analysis, visualization, and interpretation can bridge the comprehension gap, ensuring processes are not just efficient but also understandable.

To achieve such, we can capture the digital footprint of activities and transactions, allowing businesses to streamline their processes, identify bottlenecks, and enhance efficiency. However, it is important to recognize that not all aspects of real-world processes are adequately represented in any representation of the business process. To underscore the limitations of current Process Mining methods, let's consider a familiar scenario – the repair process for a mobile phone. Fig. 1 represents a simplified business process for phone repair, specifically focusing on repairing a phone with a broken camera. This figure employs a flowchart format to visually map out the sequence of activities involved in the repair process, with the manual tasks highlighted in grey. This visualization aids in comprehending the steps and stages essential for conducting phone repair within a business setting, facilitating both a high-level overview and a detailed understanding of the repairment procedure. Such visual representations are valuable tools for enhancing process efficiency, quality control, and workforce training within businesses that offer phone repair services.

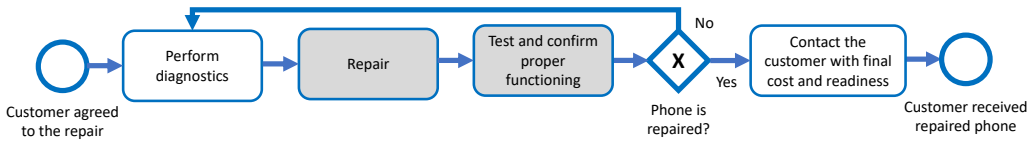

Figure 1: Simplified repair process of a phone with a broken camera. The grey activities are manual activities.

Building valuable representations of a process can be based on event logs like presented in Table 1. In this example, the event log serves as a detailed record of activities, their timestamps, sources of information, actors involved, objects in use, duration, and the respective environments in which these activities took place.

One aspect highlighted within these event logs is the source of information for each event. It is noted that most information sources are manual, indicating that a human actor manually entered data into the system to record these events. Another noteworthy element in this table is the presence of unknown timestamps, denoted by question marks, which occur when there is no interaction with a digital system to record a specific event. Consequently, this lack of digital tracking makes it challenging to determine the exact duration of certain activities, resulting in

Table 1
Sample **Pre-Enhanced** Event Logs for Phone Repair Business Process

Timestamps	**[00:18-00:28]**	**[00:28-00:33]**	**[00:33-?]**	**[?-00:59]**	**[00:59-01:04]**
Event	Start Repair	Camera Diagnosis	Replacement Part Order	Camera Replacement	Quality Check
Source of Info.	Manual	Machine log	Manual	Manual	Manual
Actor	Technician A	Technician A	Administrator	Technician A	Quality Inspector
Objects in Use	Phone, Screwdriver	Phone, Diagnostic Tool	None	Phone, Camera Module, Screwdriver	Phone
Duration	10 minutes	5 minutes	? minutes	? minutes	5 minutes
Environment	Workshop	Workshop	Office	Workshop	Workshop

estimates provided as a range of minimal to maximal values. Additionally, the table reveals that tracking the objects in use during each event is a somewhat vague aspect of this process. The list of objects is not exhaustive, and it is unclear whether some objects can be used in parallel. This ambiguity can complicate resource allocation and potentially affect the efficiency of the repair process.

Integrating multimodal data sources into event logs can offer innovative solutions to address the challenges of enriching event logs. By supplementing manual data entry with diverse data streams such as video recordings from multiple perspectives, sensor data like depth sensing maps, and audio recordings capturing specific sound events, businesses can significantly enhance the accuracy, completeness, and richness of their event logs, as illustrated in Fig. 2.

Multiple camera angles can capture different aspects of the repair procedure, allowing for precise and real-time documentation of technician actions and the condition of the device. Video data also offers the advantage of providing a visual timeline, eliminating the need for uncertain timestamps. With this visual evidence, the duration of each repair step can be accurately determined, enabling more precise process analysis. Moreover, video data can be used to verify the usage of tools and objects, making it easier to track resources in parallel and optimize resource allocation. Sensor data, particularly unstructured ones, offers a wealth of information that can complement traditional event logs. These data streams can provide insights into the intricate details of a repair, such as the depth and dimensions of components, the accuracy of alignments, and the quality of connections. Audio data recording of specific sound events can add another layer of context to event logs. For instance, the detection of particular sounds, such as clicks, whirrs, or engine sounds during the repair process, can serve as additional markers for events and their timing. These audio cues can be correlated with the corresponding visual and sensor data, offering a holistic understanding of the repair workflow. Furthermore, audio data can assist in identifying potential issues or anomalies during the repair, facilitating real-time intervention and quality assurance. Table 2 shows sample event logs by adding multimodal data sources of information to the system. The measurable exact duration of specific activities is a crucial factor in evaluating manual processes, as it often indicates efficiency, identifies areas for improvement, and guides resource allocation. Shorter durations suggest streamlined

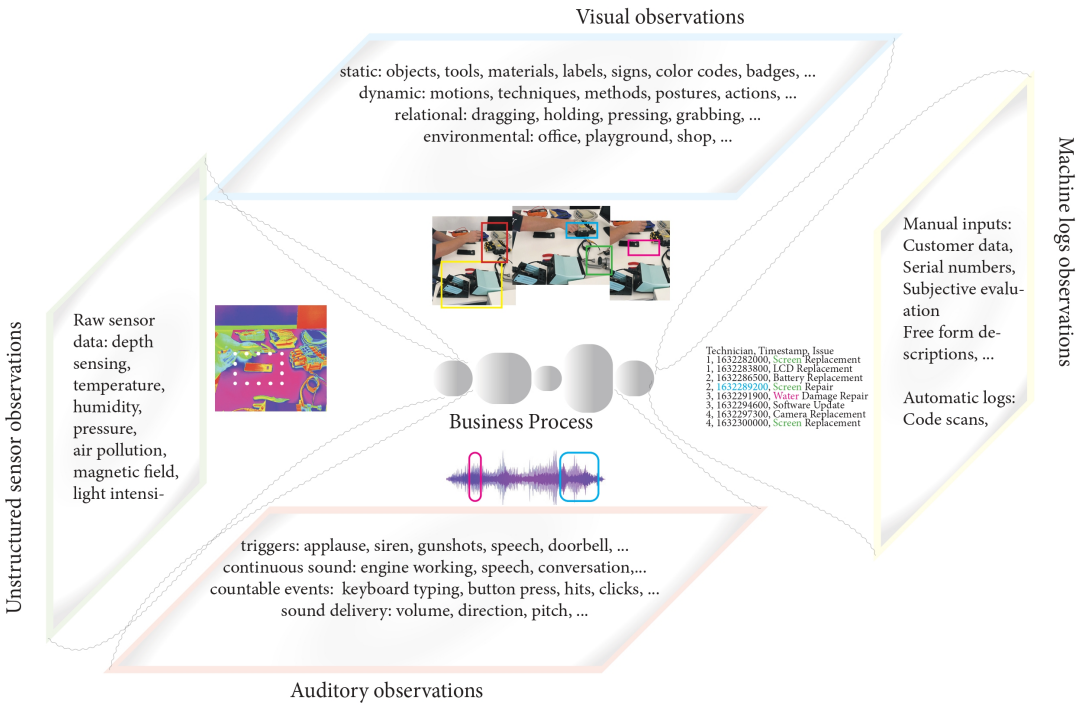

Figure 2: Multimodal enhancement of event logs for a business process.

and optimized processes, while longer durations may signal inefficiencies. Analyzing activity durations is essential for benchmarking, compliance, quality control, and ensuring a positive customer experience. However, it's vital to consider other factors such as accuracy, quality, and process effectiveness in conjunction with timing to comprehensively evaluate manual processes, as not all processes prioritize speed above all else.

To fully harness the potential of incorporating business logic on top of enriched event logs, there is a pressing need for a specialized approach that can deal with unique demands of working with video and volumetric data in conjunction with temporal and spatial information. Such an approach may involve applications of extended reality (XR) for immersive data manipulation and precise domain-specific modeling. Incorporating XR into the analysis of enriched event logs may not only enhances the understanding of complex processes but also offers a more intuitive and interactive approach to data analysis. To this end, three research questions emerge:

[RQ1] How can we effectively mine and extract process mining-relevant information from multimodal data sources, particularly from video data?

[RQ2] How can extended reality technologies be effectively integrated with enriched event logs to facilitate immersive data manipulation and visualization for precise modeling and domain-knowledge annotation?

[RQ3] How does the integration of these technologies influence the interpretability of process mining results?

Table 2
Sample **Enhanced** Event Logs for Phone Repair Business Process

Timestamps	**[00:11-00:23]**	**[00:23-00:32]**	**[00:33-00:40]**	**[00:40-00:52]**	**[00:52-1:11]**
Event	Removing phone screen	Heating up the cleaner	Camera Cleaning	Module Replacement	Quality Check
Source of Info.	Video	Audio, Thermal camera	Video	Video	Video, Manual
Actor	Technician A	Technician B	Technician A	Technician A	Quality Inspector
Objects in Use	Phone, Screwdriver, Screen tool	Cleaner	Phone, Cleaner, Textile	Phone, Camera Module, Screwdriver	Phone
Duration	12.8 minutes	8.7 minutes	7.1 minutes	12.3 minutes	19.2 minutes
Environment	Work-desk1	Cleaner's corner	Work-desk1	Work-desk2	Workshop

3. Related work

In this section, we examine related work. A broader study into related work is still ongoing, but two areas that provide promising inputs are: *Immersive Process Manipulations* and *Video and Sensory Analysis*.

3.1. Immersive process manipulations

There are studies on how modeling tools visualize the models, and how modelers interact with the models [14]. Immersive visualizations have gained significant attention as a powerful tool for exploring and understanding complex data. Researchers have explored the application of immersive technologies, such as virtual reality (VR) and augmented reality (AR), to visualize and interact with process logs.

Several scholarly papers have investigated the utility of immersive visualizations as effective tools for immersive analytic [15]. Additionally, specific research has critically reviewed immersive environments, particularly in the context of process planning [16], where authors proposed a set of design guidelines for the development of VR-based Computer-Aided Process Planning. Their findings indicate that immersive VR technologies have the potential to enhance various process planning scenarios, including decision-making, real-time response support, verification, training, and the automatic generation of process plans. However, the identified challenges these technologies still need to address are data interoperability, incorporation of organizational aspects, and technological operational accuracy. Our approach aims to address these aforementioned challenges by mining processes from richer sources of information and semantics, specifically from unstructured and highly dimensional data.

While conducting a general review of immersive design approaches, as outlined in [17], it becomes evident that there are limitations in design reviews and untapped potential in utilizing VR during the design process. The identified potential state-of-the-art practices involve the

creation of design sketches and the activation of functions within VR using a personal data assistant. Our approach is implementing a solution that facilitates model manipulation within an immersive world through the assistance of an AI-driven personal assistant.

Oberhauser et-al previously published VR-based tools for visualizing, navigating, and interacting with various modeling notations such as ArchiMate [18], Business Process Modeling Notation (BPMN) [19], Process Mining results [20], and program code structures fly-through [21]. Zenner et-al contributed to the field with a tool for Process Model Exploration [22, 23], where they introduces a concept that spatializes event-driven process chains (EPCs) by mapping traditional 2D graphs to a 3D virtual environment. While the tools presented by these authors represent the first public implementations of room-scale floating platforms, allowing users to explore various models through natural walking, neither of them explores the potential of 1) semi-automated process mining or 2) the utilization of Audio-Visual multimodal sensory data as sources of process logs and interactive entities. Additionally, these tools are not commercially or open-source available.

A conducted survey [22] gives important contribution to our proposal, as it demonstrates that using a VR interface as an alternative to traditional paper or desktop-like monitor representations does not result in a significant decrease in model understanding performance.

As for preservation of the process knowledge we encode models in various ways to make the encoded models suitable for applying Machine Learning algorithms [24]. Studies shows that event knowledge graphs are a very versatile tool that opens the door to process mining analyses in multiple behavioral dimensions at once [25], and the production of platforms for transforming conceptual models into knowledge graphs is emerging [26].

3.2. Video and sensory analysis

Understanding the semantics and meaning conveyed in instructional videos is a fundamental challenge in computer vision and natural language processing. This research stream focuses on developing approaches to automatically analyze and interpret instructional videos to extract high-level semantic information. Researchers have explored different methods to tackle this problem, ranging from video segmentation and action recognition to language understanding and multimodal fusion.

Some presented studies like [27] proposes approach for generating graphical representations of instructional videos that doesn't require any annotations. They use cross-modal attention to utilizing agreement between multiple modalities to learn a unified graphical structure representing videos as joint embedding between visual, audio and textual signals obtained from automatic speech recognition. To learn complex activities in videos and reduced computational complexity of global/data-set level representation of sub-actions, researchers transform the pipeline down to local/video level. They performed rigour evaluation of the generated graphs with a user study, and graphical and qualitative analysis.In response to the question *Tell me what happened*, the authors present a framework [28] that enables video prediction, video rewind, and video infilling, all during inference time. They evaluated their approach in various video scenarios like animation and gaming. When it comes to action recognition, novel approaches like [29] are not using object-level graphs or scene graphs to represent the dynamics of objects and relationships between them, but rather relationship transitions directly. Their solution, OR^2G, recognises

attribute transitions of objects which leads to leveraging potential in reasoning methods that are aware of relationships between objects.

In difference to our approach, none of the found approaches focuses on aligning recognised semantics with business logic, quantifying processes into business-relevant valuable metrics, and most importantly, efficiency of incorporating human in the loop to represent domain-expert knowledge. We identify as a critical step in the pipeline from raw data to valuable representations of semantics, a modeling method that follows the complexity of data that is analysed. In our approach we aim at transforming modeling to an immersive world where complex raw data such as 360° video data can be properly visualized. Approach from [28] relies on text guidance which, in our approach, we want to provide by an engaged modeler that is fully immersed into mining process. With novel available data-set for pan-optic scene graph generation such as [30], we want to make a step toward incorporating business logic into relations between entities on the scene and for such a challenging task we want to involve modelers, giving them tools for high-quality abstract-level highly-specific process mining tool that is semi-automated through AI trained guidance. Furthermore, we propose an entirely new methodology for multi-conceptualizations of the spatio-temporal event-logs that is opening various valuable applications, such as re-mining of logged data for post-mining relevance insights changes. We plan to explore multimodal event log employment of automatic techniques to automatically identify activity correspondences that represent similar behaviors [31] or annotate process models with concepts such as taxonomy [32].

3.3. Process mining based on audio-visual sources

The topic of process mining from videos as sources is yielding very few results. In this regard, [33] is providing a reference architecture for process mining from video data. Their solution, ViProMiRA, is a supervised learning-based, case-driven, context-specific tool for extracting event-logs from unstructured video data. With a prototypical implementation, they showed that ViProMiRA was capable of automatically extracting more than 70% of the process-relevant events from a real-world synthetic video data-set. They also explicitly stated as their limitation that ViProMiRA itself is not directly transferable to practical use, as it is a prototypical instantiation serving as a blueprint. Authors also pointed that their limitation is that evaluation of ViProMiRA is done on a video data which does not represent a real-world process and is limited in duration. Notably, ViProMiRA exclusively focuses on video data and does not consider other modalities such as audio, sensor data, or text information. Furthermore, [34] discusses an approach that applies process mining on video surveillance data of pigpens. The authors highlight that the process analytic pipeline from raw video data to a discovered process model has not yet been fully implemented and they see further use cases of their approach in medicine and material science. They recorded process-specific videos of 4 pigpens with a camera installation from different angle. For knowledge mining, they tailored techniques to their use case of creating a heat-map of common pig positions that is used for pig activity recognition and tracking. Their discovery of process model is enhanced with domain-specific knowledge. Finally, in the context of process mining from videos, there is a found study [35] that is analysing the consistency between the process model and the predefined Petri-net model to do a conformance checking on process models extracted from videos. They perform video data pre-processing that removes

the background information irrelevant to the moving target in the video picture and only keeps interest point area. They are performing classifications for action placement and recognition.

Our approach is filling the gap of labeling of continuous video and other multimodal sensory data using immersive technologies and leveraging larger process-related, more detailed, spatio-temporal incorporation of semantics. With our approach, a modeler can interact in a novel and more optimal way with significantly more valuable data sources. In contrast to found solutions, our approach will consider multiple conceptualizations of detected entities utilizing novel techniques such as [27, 29] for entity and action segmentation on video and other spatio and/or temporal related data. Modern techniques have potential to make automation of understanding highly-unstructured data in a semantically richer way, and we aim at applying those techniques in an immersive VR-based setting where a modeler involved in process mining can use automatic entity detections as a toolkit to be very efficiant in specifying business logic. To the best of our knowledge, no publicly available literature demonstrates a similar application, which suggests that our approach would contribute with: 1) an innovative flexible, robust, scalable, multi-person modeling tool on real world data, 2) a larger set of semantically rich activities that can be recognised, and 3) process mining from valuable sources of multimodal data including 360° video data.

4. Research strategy

4.1. Background

The underlying principle of event log creation in videos involves the extraction and analysis of meaningful semantic information from low-level event logs. While low-level logs capture detailed data such as object detection, sound recognition, and speech detection, the higher-level analysis focuses on incorporating the underlying semantics into these logs. This process involves identifying and categorizing events based on their semantic context, such as recognizing specific events composed of actions, identifying objects or people in different contexts and understanding their interactions. By incorporating semantics into the low-level event logs, the resulting higher-level event logs provide a more comprehensive and abstract representation of the video content, enabling advanced applications like activity recognition, video summarization, and semantic search.

Scene logging involves creating a knowledge graph of entities present in videos and their spatio-temporal changes within the video content. By constructing a knowledge graph, entities can be represented as nodes and their relationships and interactions can be modeled as edges connecting these nodes with additional data which may include the date, time, location, duration, and participants of a given interaction. To capture the relationships between entities, methods such as [30] can be applied to identify and recognize relations among two given objects in videos based on a collected detaset. By analyzing the changes in spatial proximity, relative positions, and trajectories of objects, meaningful relationships can be inferred.

Ontology modeling can organize objects, people, actions and events into a hierarchical structure. For example, in a scene, different types of vehicles can be categorized into subclasses such as cars, trucks, or motorcycles. Maintaining this hierarchical organization is crucial

for efficient object identification, as it allows for a more detailed understanding of the scene and facilitates higher-level reasoning. Hierarchical object recognition can be achieved through hierarchical classifiers or by employing deep learning architectures that incorporate hierarchical features. These approaches enable the system to identify objects at different levels of abstraction, from general object categories to fine-grained subclasses.

Combining data from different modalities To gain a comprehensive understanding of the processes, data from different modalities, such as visual, audio, or sensor data, can be combined. Integration techniques and data fusion algorithms can be employed to merge information from various sources, enriching the analysis and providing a more holistic view of the tracked processes.

4.2. Solution architecture design

The solution architecture depicted in Fig. 3 offers a comprehensive, layered approach to integrating and analyzing multimodal data sources in the context of process mining. This design embodies a holistic strategy, enabling the conversion of raw data inputs into enriched conceptual models. The goal of this architecture is its applicability in a multitude of fields. The solution should show transformative impact across domains of Process Mining, where the model aids in deciphering complex processes, or Process Monitoring/Tracking, where real-time insights are extracted, or even Process Guidance, where the model provides direction and recommendations that can be valuable for both training purposes or real-time on-field guidance.

Source data acquisition The foundation of our architecture begins with the Source Data Acquisition phase. This phase accumulates a wide array of input sources, ranging from videos captured from diverse vantage points, to audio recordings, in-depth sensor data maps, and detailed machine logs that track all interactions and database entries with the associated software. Given the heterogeneity of these data sources, it's imperative to have a unified architecture to ensure that all these disparate data streams are captured, processed, and made ready for subsequent layers of analysis.

Automated contextualization (observable entity recognition) The Automated Observable Entity Recognition phase, also termed as Automated Contextualization, is predominantly AI-driven and boasts of capabilities like object recognition (initially unnamed), identification of people (who are not yet recognized as system actors), motion tracking that detects sequences of postures of individuals and objects, classifiers for predicting the categories of various resources or activities, and background isolation to discern between diverse environments and settings. The power of AI in this phase ensures that the raw data is swiftly transformed into recognizable entities and contexts, paving the way for deeper analysis and integration.

Immersive contextualization The Immersive Contextualization phase is manual and calls upon domain-knowledge experts to embed business logic into the evolving model. This immersive experience is facilitated through cutting-edge augmented reality glasses or virtual reality headsets, enabling experts to navigate through 360° videos. At this juncture, the Entity Naming process comes to the fore, serving as a verification mechanism for the automated classifications performed earlier. This is complemented by Artifact Manipulations, which provide the flexibility to either group entities into more abstract categories or further specify them. For instance, an environment can be identified as a derivative of another setting, contingent on the availability

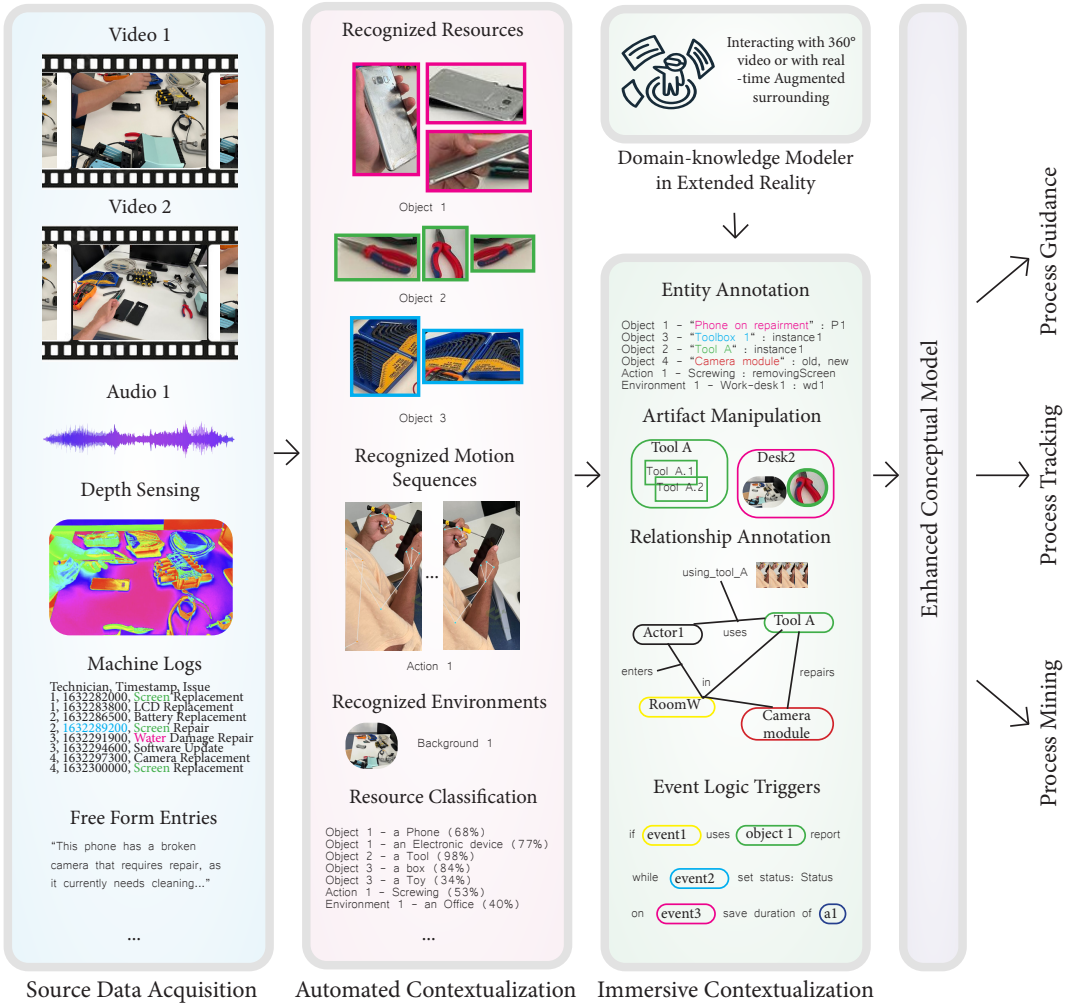

Figure 3: Solution design architecture.

of specific tools. Relationship Annotation further enriches the model, charting out a graph of pertinent links between artifacts and activities. The culmination of this phase is the Event Triggers module, where intricate business logic rules and advanced connections between the recognized entities and activities are established.

Enhanced conceptual model After traversing through the intricate layers of the architecture, what emerges is the Enhanced Conceptual Model. This model is not only aware of the enriched multimodal event log but is also primed for diverse applications. It encapsulates the depth and breadth of the information from the varied data sources and the insights garnered through automated and manual contextualization.

4.3. Source data acquisition

The foundation of our architecture begins with the Source Data Acquisition phase. This section accumulates a wide array of input sources, ranging from videos captured from diverse vantage points, to audio recordings, in-depth sensor data maps, and detailed machine logs that track all interactions and database entries with the associated software. Given the heterogeneity of these data sources, it's imperative to have a unified architecture to ensure that all these disparate data streams are captured, processed, and made ready for subsequent layers of analysis.

4.4. Automated contextualization (observable entity recognition)

Following the data collection phase, the architecture delves into the Automated Observable Entity Recognition section, also termed as Automated Contextualization. This segment is predominantly AI-driven and boasts of capabilities like object recognition (initially unnamed), identification of people (who are not yet recognized as system actors), motion tracking that detects sequences of postures of individuals and objects, classifiers for predicting the categories of various resources or activities, and background isolation to discern between diverse environments and settings. The power of artificial intelligence in this phase ensures that the raw data is swiftly transformed into recognizable entities and contexts, paving the way for deeper analysis and integration.

4.5. Immersive contextualization

The subsequent tier, Immersive Contextualization, is manual and calls upon domain-knowledge experts to embed business logic into the evolving model. This immersive experience is facilitated through cutting-edge augmented reality glasses or virtual reality headsets, enabling experts to navigate through 360-degree videos. At this juncture, the Entity Naming process comes to the fore, serving as a verification mechanism for the automated classifications performed earlier. This is complemented by Artifact Manipulations, which provide the flexibility to either group entities into more abstract categories or further specify them. For instance, an environment can be identified as a derivative of another setting, contingent on the availability of specific tools. Relationship Annotation further enriches the model, charting out a graph of pertinent links between artifacts and activities. The culmination of this phase is the Event Triggers module, where intricate business logic rules and advanced connections between the recognized entities and activities are established.

4.6. Enhanced conceptual model

After traversing through the intricate layers of the architecture, what emerges is the Enhanced Conceptual Model. This model is not only aware of the enriched multimodal event log but is also primed for diverse applications. It encapsulates the depth and breadth of the information from the varied data sources and the insights garnered through automated and manual contextualization.

5. Evaluation discussion

5.1. Identification of the target audience and beneficiaries of the thesis

This thesis primarily targets industries that experience gaps or "blind spots" in their event logs due to manual or physically-based processes. Industries like manufacturing, logistics, warehousing, agriculture, and construction, which heavily rely on manual labor and physical operations, often lack comprehensive electronic tracking of every nuanced event. The proposed approach for extracting information from multimodal data sources, especially video data, offer a groundbreaking way for these sectors to digitize and analyze manual processes that were previously opaque. Extended reality technologies can provide these industries with immersive training and process understanding tools, making the shift from manual oversight to digital monitoring smoother and more intuitive. Additionally, with the inclusion of BPMN in multimodal data analysis, businesses can gain contextually richer insights, aiding in process optimization. Operational managers in these sectors can not only interpret the newly available process mining results but also predict future process bottlenecks and constraints. Furthermore, organizations aiming to create a standardized blueprint of their manual processes can leverage the business process libraries proposed by this research. All in all, this thesis promises a transformational shift for industries traditionally limited by manual process constraints.

5.2. Validation techniques

To ensure the robustness and applicability of the approach proposed for the planned thesis, it is imperative to employ rigorous validation techniques. One such technique is prototyping. Given the novel amalgamation of video, audio, and sensor data with XR technologies, creating a prototype offers a tangible representation of our concepts and allows stakeholders to interact with and refine the proposed system. By offering an early visualization of the system's functionalities and interfaces, prototyping facilitates immediate feedback, highlights potential pitfalls, and offers avenues for iterative refinement. Moreover, a prototype can simulate the data integration process, giving researchers a glimpse into the challenges and solutions of working with disparate data types in a unified XR environment.

Additionally, field experiments stand out as a powerful validation approach for our research. By applying our methods in real-world settings, we can assess the practicality, efficiency, and accuracy of our proposed techniques. Observing domain-experts as they engage with our XR-powered process mining tool can provide valuable insights into its usability, effectiveness, and areas for improvement. Such experiments also help in determining the tool's impact on business process workflows, stakeholders satisfaction, and the overall quality of monitored processes. Meanwhile, argumentation, although more theoretical, offers a structured platform to articulate and defend the reasoning behind our approach, ensuring that the choices made in data integration, XR application, and process analysis are not only innovative but also logically sound and consistent with existing domain knowledge.

5.3. Contribution of the thesis

The contribution of the thesis follows three topics:

Topic 1 – Mining: Focused on extracting meaningful information from data. It handles the visualization of data, manages multi-view video data, detects process loops, visualizes dependencies, identifies critical paths, and more.
Topic 2 – Tracking and monitoring: Uses videos, other modalities and extended reality for real-time process tracking, ensuring continuity of processes. It can detect anomalies, handle variations in processes over time, and create reports like summaries.
Topic 3 – Guiding: Utilizes extended reality for process guidance, transforming data into interactive guides. The topic also delves into data visualization techniques, user-friendly interface design, and training requirements.

6. Research execution plan

In our journey thus far, we have explored and tested an array of technologies, harnessing their potential for process mining in an enriched multimodal environment. Our current explorations have centered around fine-tuning pre-trained models tailored for both visual and audio recognition. Recognizing the power of visual language models, we conducted experiments to mine business logic from video frames. By furnishing these models with tailored prompts, we've made significant strides in extracting meaningful insights from the visual data. Additionally, our tests of human body pose estimation, image segmentation, and image classification have further equipped us with tools to understand and interpret the vast and varied visual inputs at our disposal.

Looking ahead, our roadmap is illustrated in Fig. 4. Our immediate endeavor is the development of an XR tool designed to empower modelers in navigating 360° videos, allowing them to pinpoint and select relevant objects with precision. This selection process is envisioned to be intuitive and immersive, leveraging a cube selector to volumetrically extract the mesh of the scene. This hands-on approach ensures that the extracted data is both relevant and contextual. Once this foundational tool is in place, our energies will shift toward the creation of a comprehensive immersive modeling tool. This tool will be pivotal in preparing us for the subsequent phase of data labeling, ensuring that the raw data is categorized and marked for further analysis. As we progress, the collection of source data will be paramount, and we anticipate diving deeper into state-of-the-art pre-trained models that can assist in AI-driven contextualizations.

The proposed solution is designed to address a multitude of challenges and considerations across various domains. One critical aspect is ensuring interoperability with different data formats. This entails the integration, standardization, and compatibility of diverse data inputs, allowing the solution to handle data from various sources effectively. This feature is pivotal in enabling the solution to be adaptable and versatile in processing information. Another essential feature focuses on accommodating diverse user needs. By offering customization options, user interface personalization, and adaptive features, the solution aims to enhance user-friendliness and satisfaction. This approach ensures that users from different backgrounds and with varying requirements can effectively utilize the solution to meet their specific needs.

Handling biases in multimodal data is another crucial aspect addressed in the solution. It explores methods for detecting, mitigating, and assessing biases, which is essential for

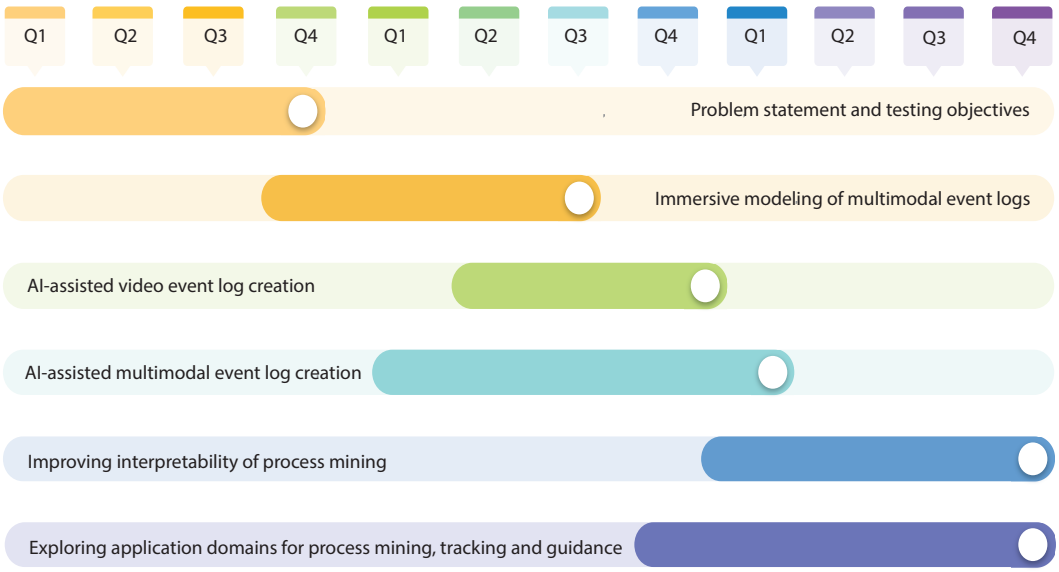

Figure 4: Roadmap of future plans time-framed in three sets of four time portions.

maintaining the integrity and reliability of analysis results. This feature underscores the commitment to fairness and objectivity in data analysis. Dealing with missing or incomplete data is a common challenge in data analysis, and the solution provides strategies for addressing this issue. Techniques such as imputation, data augmentation, and sensitivity analysis are explored to mitigate the impact of missing information on analysis results, ensuring more robust outcomes.

Concluding our roadmap, our focus will pivot towards diverse applications for our integrated system. Moreover, as we venture deeper into process mining, ensuring the interpretability of these mined processes will be crucial. This will ensure that the insights gleaned from our system are not only accurate but also actionable, lending them real-world relevance and applicability.

The future work can dive into advanced topics like real-time monitoring, faster data analysis, and accelerated process improvement, showcasing the potential for innovation in data analysis methods. Additionally, future work considers efficient storage strategies, concurrent process mining, uncovering hidden process patterns, predicting process performance improvement, and structuring business process libraries, highlighting a comprehensive approach to data analysis and process optimization. The solution's ability to be applied successfully in various industries is a testament to its adaptability and potential for widespread utility.

7. Conclusion

Our investigation into the integration of multimodal data sources and extended reality technologies in process mining has illuminated the transformative potential of this synergistic approach.

As demonstrated in our solution design architecture, the journey from Source Data Acquisition to an Enhanced Conceptual Model underscores the power of combining raw data inputs with both automated and immersive manual contextualization techniques. The goal of this integrated approach is not only in its capacity to transform heterogeneous data into actionable insights but also in its versatility, offering applications across process mining, monitoring, and guidance domains.

The implications of this research are profound for industries reliant on process analyses, especially those in repair and maintenance sectors. Our architectural model offers organizations a blueprint to harness their diverse data streams, contextualize them through AI and human expertise, and subsequently derive enhanced, actionable conceptual models.

In summation, while the horizon of process mining and analysis is expansive, our study introduces a pioneering approach that marries the best of technology and human expertise. Future endeavors in this field should focus on scalability, ensuring that our architecture can cater to even more complex and voluminous data environments, and further bridging the gap between raw data and actionable insights.

Acknowledgments

Authors are deeply grateful to Prof. Dr. Henrik Leopold at Kühne Logistics University (KLU) for his invaluable guidance in shaping this work.

References

[1] W. M. van der Aalst, Foundations of process discovery, in: Process Mining Handbook, Springer, 2022, pp. 37–75.

[2] W. van der Aalst, Data science in action, in: Process mining, Springer, 2016, pp. 3–23.

[3] W. M. Van Der Aalst, A practitioner's guide to process mining: Limitations of the directly-follows graph, 2019.

[4] L. Zimmermann, F. Zerbato, B. Weber, What makes life for process mining analysts difficult? a reflection of challenges, Software and Systems Modeling (2023). doi:10.1007/s10270-023-01134-0.

[5] J. Michael, D. Bork, M. Wimmer, H. C. Mayr, Quo vadis modeling?, Software and Systems Modeling (2023). doi:10.1007/s10270-023-01128-y.

[6] D. Bork, S. J. Ali, B. Roelens, Conceptual modeling and artificial intelligence: A systematic mapping study, 2023. arXiv:2303.06758.

[7] V. Kulkarni, S. Reddy, T. Clark, H. Proper, The AI-Enabled Enterprise, Springer International Publishing, Cham, 2023, pp. 1–12. doi:10.1007/978-3-031-29053-4_1.

[8] H. Proper, B. van Gils, Coordinated Continuous Digital Transformation, Springer International Publishing, Cham, 2023, pp. 101–120. doi:10.1007/978-3-031-29053-4_6.

[9] E. V. Epure, P. Martin-Rodilla, C. Hug, R. Deneckère, C. Salinesi, Automatic process model discovery from textual methodologies, in: 2015 IEEE 9th International Conference on Research Challenges in Information Science (RCIS), IEEE, 2015, pp. 19–30.

[10] J. Mendling, H. Leopold, F. Pittke, 25 challenges of semantic process modeling, International Journal of Information Systems and Software Engineering for Big Companies 1 (2015) 78–94.
[11] B. v. Gils, H. A. Proper, Next-Generation Enterprise Modeling, Springer Nature Switzerland, Cham, 2023, pp. 279–305. doi:10.1007/978-3-031-30214-5_21.
[12] H. A. Proper, B. van Gils, K. Haki, Final Conclusions and Outlook, Springer Nature Switzerland, Cham, 2023, pp. 311–314. doi:10.1007/978-3-031-30214-5_23.
[13] H. van der Aa, H. Leopold, F. Mannhardt, H. A. Reijers, On the fragmentation of process information: Challenges, solutions, and outlook, in: Enterprise, Business-Process and Information Systems Modeling, Springer, 2015, pp. 3–18.
[14] D. Bork, G. De Carlo, An extended taxonomy of advanced information visualization and interaction in conceptual modeling, Data & Knowledge Engineering 147 (2023) 102209. doi:10.1016/j.datak.2023.102209.
[15] M. Kraus, K. Klein, J. Fuchs, D. A. Keim, F. Schreiber, M. Sedlmair, The value of immersive visualization, IEEE computer graphics and applications 41 (2021) 125–132.
[16] J. Zhou, J. D. Camba, Computer-aided process planning in immersive environments: A critical review, Computers in Industry 133 (2021) 103547.
[17] D. Weidlich, L. Cser, T. Polzin, D. Cristiano, H. Zickner, Virtual reality approaches for immersive design, CIRP annals 56 (2007) 139–142.
[18] R. Oberhauser, C. Pogolski, Vr-ea: Virtual reality visualization of enterprise architecture models with archimate and bpmn, in: Business Modeling and Software Design: 9th International Symposium, BMSD 2019, Lisbon, Portugal, July 1–3, 2019, Proceedings 9, Springer, 2019, pp. 170–187.
[19] R. Oberhauser, C. Pogolski, A. Matic, Vr-bpmn: Visualizing bpmn models in virtual reality, in: Business Modeling and Software Design: 8th International Symposium, BMSD 2018, Vienna, Austria, July 2-4, 2018, Proceedings 8, Springer, 2018, pp. 83–97.
[20] R. Oberhauser, Vr-processmine: Immersive process mining visualization and analysis in virtual reality, in: Proceedings of the Fourteenth International Conference on Information, Process, and Knowledge Management, 2022, pp. 75–80.
[21] R. Oberhauser, C. Lecon, Virtual reality flythrough of program code structures, in: Proceedings of the Virtual Reality International Conference-Laval Virtual 2017, 2017, pp. 1–4.
[22] A. Zenner, A. Makhsadov, S. Klingner, D. Liebemann, A. Krüger, Immersive process model exploration in virtual reality, IEEE transactions on visualization and computer graphics 26 (2020) 2104–2114.
[23] A. Zenner, S. Klingner, D. Liebemann, A. Makhsadov, A. Krüger, Immersive process models, in: Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems, 2019, pp. 1–6.
[24] S. J. Ali, A. Gavric, H. Proper, D. Bork, Encoding conceptual models for machine learning: A systematic review (2023) 562–570. doi:10.1109/MODELS-C59198.2023.00094.
[25] D. Fahland, Process mining over multiple behavioral dimensions with event knowledge graphs, in: Process Mining Handbook, Springer, 2022, pp. 274–319.
[26] M. Smajevic, S. J. Ali, D. Bork, Cm2kgcloud – an open web-based platform to transform

conceptual models into knowledge graphs, Science of Computer Programming 231 (2024) 103007. doi:10.1016/j.scico.2023.103007.
[27] M. C. Schiappa, Y. S. Rawat, Svgraph: Learning semantic graphs from instructional videos, in: 2022 IEEE Eighth International Conference on Multimedia Big Data (BigMM), IEEE, 2022, pp. 45–52.
[28] T.-J. Fu, L. Yu, N. Zhang, C.-Y. Fu, J.-C. Su, W. Y. Wang, S. Bell, Tell me what happened: Unifying text-guided video completion via multimodal masked video generation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 10681–10692.
[29] Y. Ou, L. Mi, Z. Chen, Object-relation reasoning graph for action recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 20133–20142.
[30] J. Yang, Y. Z. Ang, Z. Guo, K. Zhou, W. Zhang, Z. Liu, Panoptic scene graph generation, in: European Conference on Computer Vision, Springer, 2022, pp. 178–196.
[31] H. Leopold, Business process model matching, in: S. Sakr, A. Y. Zomaya (Eds.), Encyclopedia of Big Data Technologies, Springer, 2019.
[32] H. Leopold, C. Meilicke, M. Fellmann, F. Pittke, H. Stuckenschmidt, J. Mendling, Towards the automated annotation of process models, in: International Conference on Advanced Information Systems Engineering, Springer, 2015, pp. 401–416.
[33] W. Kratsch, F. König, M. Röglinger, Shedding light on blind spots–developing a reference architecture to leverage video data for process mining, Decision Support Systems 158 (2022) 113794.
[34] A. Lepsien, J. Bosselmann, A. Melfsen, A. Koschmider, Process mining on video data., ZEUS 3113 (2022) 56–62.
[35] S. Chen, M. Zou, R. Cao, Z. Zhao, Q. Zeng, Video process mining and model matching for intelligent development: Conformance checking, Sensors 23 (2023) 3812.