Multi-Modal Large Language Model driven Augmented Reality Situated Visualization: the Case of Wine Recognition

Vincenzo Armandi
Department of the Arts
University of Bologna, Bologna, Italy

Andrea Loretti
Department of the Arts
University of Bologna, Bologna, Italy

Lorenzo Stacchio
Department of Political Sciences, Communication and International Relations
University of Macerata, Macerata, Italy

Pasquale Cascarano *
Department of the Arts
University of Bologna, Bologna, Italy

Gustavo Marfia
Department of the Arts
University of Bologna, Bologna, Italy

Abstract

Situated Visualizations (SV) and reality-based information retrieval systems, enhanced by Mixed Reality (MR) and Augmented Reality (AR), enable the overlay of digital information onto real-world objects, providing context-aware content through computer vision. Despite their potential, these systems face significant challenges in scalability and adaptability, particularly for domains like wine recognition, where diverse label designs, frequent updates, and limited historical databases complicate automated analysis. SOLLAMA (SOmmeLier LlAMA) is a novel wine recognition framework designed to address the scalability and adaptability challenges of AR systems in recognizing diverse wine labels. Leveraging Multimodal Large Language Models (MLLMs), SOLLAMA integrates visual and textual cues for accurate label interpretation, bypassing the need for extensive image datasets and traditional OCR methods. Built on the Augmented Wine Recognition (AWR) system, it replaces the OCR module with LLAMA 3.2 for advanced text recognition and contextual understanding. Key benefits include scalability across diverse designs and simplified, server-free deployment. Experimental validation on a dataset of wine labels from Italy's Emilia–Romagna region highlights the system's effectiveness and its potential to transform wine recognition in AR-based applications.

*Corresponding author: pasquale.cascarano2@unibo.it

1. Introduction

Situated visualizations (SV) and reality-based information retrieval systems are designed to overlay context-specific digital information onto real-world entities, such as food, people, buildings, and photographs [4, 29, 44]. In SV, digital data linked to physical objects are collected and collocated to serve as anchors [12]. The increasing accessibility of Mixed Reality (MR) and Augmented Reality (AR) technologies has expanded the viability of SV across various domains, providing users with pertinent information about physical objects and guiding them through specific processes, such as learning or decision-making [16, 17, 23, 33, 48].

These systems are based on computer vision techniques, which extract relevant information to be dynamically used within an AR environment thus creating responsive experiences that adapt in real-time to the identified data, providing either the exact information or semantically related contents [4, 23, 31]. This allows for the superimposition of contextually relevant digital content onto the physical environment, enhancing user interaction and understanding. Such paradigms can be applied in a variety of contexts, from industrial settings to education [43, 45]. This integration is commonly applied using visual anchor targets, such as 2D images and 3D models, and also document targets (i.e., pictures containing structured text) [8, 23]. In particular, in document recognition, some works have focused on scanning documents to display augmented information related

to their content, including metadata extraction and contextual overlays [40, 46, 59].

Food labels can be considered as a specialized form of a document, as they contain structured information such as nutritional facts, ingredients, and origin details [37]. In the food and beverage sector, numerous initiatives have focused on scanning packages to display augmented information related to their contents, including nutritional details and reviews [34, 39, 47]. An interesting field of application amounts to wine recognition, where applications like Yuka [57] and Vivino [51] exemplify this trend, offering users insights into product quality and composition. These applications typically rely on marker-based or markerless computer vision techniques for object detection, recognition, and tracking, which often demand substantial computational resources and extensive datasets of reference images [42]. However, these methods may struggle to scale effectively with long-tail products, such as wines, due to the vast diversity and frequent updates in wine labels [18] which can be also impossible to cover, considering the limited availability of images for older vintages. Considering these limits for wine recognition, and the textual analysis approach proposed by [46], authors of [4] introduced the Augmented Wine Recognition (AWR) system leveraging a deep learning-based OCR and domain-specific knowledge to identify wine types through the textual information on bottle back labels. They address the limitations of traditional marker-based or computer vision approaches, particularly in handling the long-tail distribution of wine products, where labels may vary or lack comprehensive databases. Leveraging European regulations on wine label design, the AWR system employs a custom tree data structure and hierarchical search algorithm to match textual features efficiently. The tests on more than 2000 wine labels, from Italy's Emilia–Romagna region, demonstrated high recognition accuracy and a fast inference time.

Despite these interesting approaches, such methods may struggle to scale effectively with diverse and evolving document formats due to the wide variety and frequent document layouts and design updates [3]. Indeed, the complex visual features of wine labels—such as intricate backgrounds, diverse fonts, and distortions from the bottles' curved surfaces—can inhibit OCR accuracy. Even with accurate text recognition, automatically discerning which words are pertinent for identifying a specific wine can be time-consuming and may not align with the real-time requirements of SV [4]. Consequently, implementing an effective system necessitates the integration of methods, algorithms, and technologies to ensure both detection efficacy and efficiency.

To address these challenges, we propose leveraging Multimodal Large Language Models (MLLMs) for wine recognition through the automatic interpretation of the text, reported in the image of the label itself. MLLMs are designed to process and integrate multiple data modalities, such as text and images, enabling them to comprehend and generate content that encompasses both visual and textual information [56]. By employing MLLMs, we can develop a system that interprets the visual and textual elements of wine labels holistically, facilitating accurate wine identification without relying solely on extensive image databases or traditional OCR methods. In practice, we built upon the AWR system, changing the OCR module with an open-source MLLM (in our case LLAMA 3.2 [13]) towards completely automatizing the textual recognition from pictures through prompted instructions. The system is referred to as SOmmeLier LlAMA (SOLLAMA).

This approach offers several advantages. (i) **Scalability**. MLLMs can generalize across a wide array of wine labels, including those not present in existing databases, by understanding visual and textual cues. (ii) **Reduced Computational Requirements**. By processing images and text concurrently, MLLMs can operate efficiently without the need for high computational time. (iii) **Server-Free Implementation**. MLLMs can be deployed on-device, enabling server-free operations that enhance user privacy and reduce latency. (iv) **Simplified Logic**. The integrated understanding of visual and textual information by MLLMs minimizes the need for complex processing pipelines, streamlining the system architecture.

The rest of this paper is organized as follows. In Section 2, we review the related work on food and beverage product identification, particularly focusing on the methods used in AR applications for label recognition and product identification, including both commercial and academic solutions. Section 3 introduces the domain-specific wine background, detailing the regulatory structure and information that can be leveraged for wine identification. Section 4 describes the proposed SOLLAMA system, including its AR interface, the MLLM module, and the hierarchical textual database and search algorithm built upon its textual recognition. In Section 5, we evaluate the proposed system, providing an in-depth analysis of its efficacy and efficiency in recognizing wine labels. Finally, in Section 6, we discuss the limitations, propose potential improvements, and outline directions for future work.

2. Related work

We here review those works that focused on providing systems to recognize information for SV in food industry. Then, we review the most recent works adopting LLMs and MLLMs in document analysis and text extraction.

2.1. Product Recognition for Situated Visualization

Food and beverage product identification for SV is often performed with bar codes or QR codes. Many AR applications exploit image detection and recognition paradigms,

which may also be based on codes or on the recognition of a product as it appears [35,41,42,47]. Identifying and exploiting visual cues in food product pictures is an approach that appeared in various research contributions [15, 19, 26, 60]. In this scenario, some works, from both industrial and academic contexts, focused on wine label recognition and its application in the AR realm [4,5,20,24,28,38,49]. Regarding commercial solutions, *WineEngine* is an online wine label recognition service [49] which exploits a combination of OCRs and image-retrieval-based approaches using the wine bottle front label. This approach requires adding reference label images to the considered database and does not provide an AR interface. Another interesting system to recognize wine bottles is *Living Wine Labels* [28]. Finally, Vivino is the most downloaded app with a community comprising 20 million users around the globe [38, 51] and provides features such as wine exploration, evaluations, and a wine bottle front label recognizing service. Vivino does not provide an AR interface and implements an image retrieval approach based on the Vuforia Cloud Recognition service.

Considering now academic contributions, mostly have followed image retrieval-based approaches [2,5,24]. In [24] the authors proposed a CNN-SIFT framework for wine label retrieval, where a trained CNN model recognizes the wine producer to narrow the search range, while a SIFT descriptor empowered with RANSAC and TF-IDF mechanisms matches the final sub-brand. In [2], the authors presented an AR system running on a Microsoft HoloLens, making use of the Vuforia SDK to recognize markers attached to wine bottles and to display information concerning those bottles [52]. It is also possible to find other approaches in literature that concentrate on recognition sub-problems. All of the aforementioned academic contributions rely on image-retrieval-based approaches, and so present the main limit of requiring an extensive image database, which may be very difficult if not impossible considering old, out-of-production, or new wine types (i.e., long-tail samples). Differently, [4] implemented an OCR-based solution to read serial numbers from wine labels to provide counterfeit prevention and brand protection. However, this would be required to have access to all the correspondences between serial numbers and related bottle wine types. The latter, paved the way for a textual analysis of wine labels (and in general product labels) exploiting domain knowledge cues to recognize the product, without relying on any image retrieval approach. This also put the basis for the exploitation of MLLMs to automatize this process, possibly removing any kind of implementation logic.

2.2. Multi-Modal Large Language Model Document Analysis

LLMs have demonstrated remarkable proficiency in understanding and generating text, leading to their application in various Information Extraction (IE) tasks. [58] provided a comprehensive survey on the integration of LLMs into IE tasks, emphasizing their effectiveness in extracting structured information from unstructured text. Considering document-level relation extraction, [55] introduced a framework that utilizes LLMs to identify relationships between entities across entire documents. However, the development of MLLMs has enabled the processing of diverse data types, enhancing document understanding by integrating textual and visual information. On this line, [25] analyzed what MLLMs could nowadays achieve in vision-language tasks, analyzing their architectures and training techniques. In this domain, authors of [53], introduced a layout-aware generative language model, that exemplifies the application of MLLMs in document understanding. DocLLM effectively handles tasks such as information extraction and document classification by incorporating text and visual layout information. On a similar line, [27] developed an OCR-free large multimodal approach designed for document understanding. It integrates text and visual features only, achieving notable improvements across various benchmarks related to scene text-centric and document-oriented tasks. Finally, in the retail and food domain, authors of [21] replace a complex multi-step pipeline involving image preprocessing, object detection, OCR, and supervised object classification with diverse MLLM, evaluating their performance in production-level visual-question answering and OCR tasks using the Retail-786k dataset [22], which comprises approximately 786,000 high-resolution product images from European retailers. Findings revealed that performance varied significantly depending on the task: most models accurately answered questions regarding product brand and price but struggled with fine-grained classification tasks, such as correctly identifying specific product names or detecting discounts. These results suggest that further investigation on the visual-text factors influencing those must be performed. In any case, all the text extracted and analyzed by these models can be reframed to enable their automatic localization within documents, providing cues for a structured search in classical algorithms [36].

References	IR	AR	AReT	OCR	TDO	TLM
[49]	✓	✗	✓	✓	✗	✗
[28]	✓	✓	✗	✗	✗	✗
[51]	✓	✗	✗	✗	✗	✗
[14, 20, 24, 32, 54]	✓	✗	✗	✗	✗	✗
[2]	✓	✓	✓	✗	✗	✗
[5]	✗	✗	✓	✓	✓	✗
[4]	✗	✓	✓	✓	✓	✗
SOLLAMA	✗	✓	✓	✗	✓	✓

Table 1. Comparison between the characteristics of the different wine recognition systems and SOLLAMA.

Differently from the cited related works in wine recog-

nition, SOLLAMA aims to directly integrate multimodal LLMs to simplify the implementation logic for the recognition phase in situated visualization, to provide a novel and more efficient approach to recognize the discriminative information to distinguish a wine bottle type from another.

Table 1 compares the characteristics of our solution against existing ones, where IR stands for Image Retrieval, AR for Augmented Reality, AReT for Almost Real-Time, OCR indicates the usage of an OCR, TDO for Textual Database only, and TLM indicates that is exploiting an LLM fo textual extraction.

3. Wine Domain Knowledge

A wine bottle typically features two labels: a front label and a back label. The front label is primarily used for brand communication, while the back label provides detailed information about the wine, formatted in compliance with the regulations of its country of origin [7]. In some cases, bottles may have a single label that consolidates all required details in a compact format. For clarity, we will use the term "label" to refer specifically to those containing information necessary for distinguishing wine types, as mandated by Italian regulations.

Recent work has provided a valuable description of the historical development of wine policies in Europe inside the Common Market Organization (CMO) [1]. According to Italian regulations, specific information (e.g., wine appellation, winery) must appear on a wine bottle in the same field of view (i.e., a consumer should not have to turn a bottle to read them all). Italian labels report different information, some mandatory and some not [6, 11, 30, 50].

The wine label contains various pieces of information essential for identification and classification, listed as follows. (i) **Name**: The wine name, typically displayed at the label's top-center. (ii) **Type**: Specifies whether the wine is varietal, generic, or appellation-based. Appellation wines refer to production within specific geographical areas. (iii) **Appellation**: Appellation wines are categorized as either Protected Geographical Indication (PGI/IGP) or Protected Designation of Origin (PDO/DOP). PDO wines (e.g., DOC, DOCG) require all production phases to occur in the designated region, while PGI wines (e.g., IGT) allow at least one phase outside the region. These classifications were formalized after the 2008 CMO reform. (iv) **Appellation Value**: This is the unique "proper name" of the wine type within its class (e.g., Pignoletto for DOC wines) and is displayed alongside the appellation. (v) **Winemaker/Winery**: The label must include the name of the winery where the wine is bottled, which may produce multiple labels or brands. (vi) **Region of Origin**: While not mandatory, it can often be inferred from the appellation value for appellation wines or identified through the winery. (vii) **Effervescence**: Specifies whether the wine is still, sparkling, or spumante. If not indicated, the wine is assumed to be still. (viii) **Sweetness**: Terms vary depending on effervescence, such as *Secco*, *Brut*, or *Dolce*, with sweetness mandatory only for spumante wines. (ix) **Color**: Identified as red, white, or rosé, possibly using synonyms.

The information introduced to this point is valuable to uniquely identify wine types, which are also the wine descriptors adopted in our system to discriminate among them. These amounts to those labels that must be extracted and recognized automatically by the considered MLLM.

4. SOLLAMA system

The proposed SOLLAMA system (see Fig. 2) includes two main components: (a) a client AR interface running on a mobile device, used to take pictures of the wine label and present AR content after wine type identification, and (b) an algoritmic pipeline, which employs an MLLM to retrieve the relevant text within the image sent by the mobile device, then fed to the hierarchical search algorithm implemented in [4], providing the best wine-type candidates.

4.1. Augmented Reality interface

(a) (b) (c)

Figure 1. AR interface: (a)-(b) Wrong and Correct suggestions, (c) Correct Identification confirmation.

Fig. 1 shows the main processes and features of the SOLLAMA system, depicting four different views of the AR interface. The client side of SOLLAMA has been implemented adopting an AR approach for Android-based smartphones (developed with Unity and the Vuforia SDK). Once activated, the AR interface starts to continuously scan what is framed by the device camera, collecting more and more frames. When a certain number of frames are collected, the system to verify whether the camera is pointing at a known label. During this recognition process, a spinning loading icon appears on the screen's bottom left corner. Once a label is recognized, the interface shows the wine name, appellation, region, and region image (if available) related to

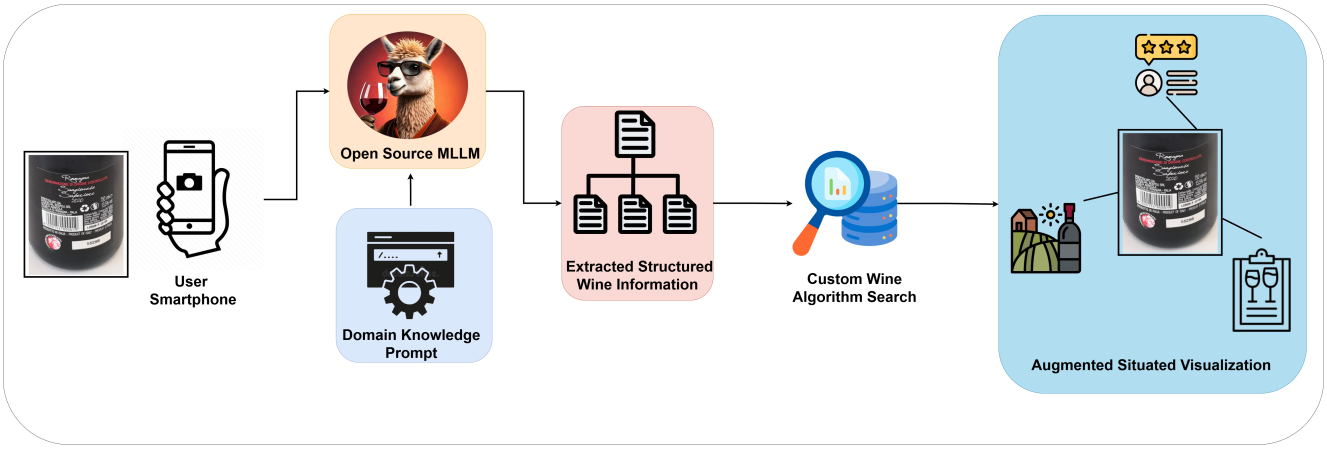


Figure 2. The SOLLAMA is an MLLM-driven AR system designed for wine label recognition and information retrieval. The process begins with a user capturing an image of a wine label using a smartphone. This image is processed through an open-source MLLM, guided by a domain knowledge prompt to extract structured wine information. The extracted data is then processed by a custom wine algorithm search, which interfaces with a hierarchical database to identify relevant wine attributes efficiently. Finally, the system provides augmented situated visualizations, offering detailed product insights and recommendations.

the first query result. The interface also lists other possible candidates on its right panel. If the right answer is present in this list, a user can select it. In this case, the interface opens a dialogue box asking the user to save the selected result. In case the back-end recognition service is not able to match the targeted wine because the related entry is not included in the textual DB, an alert is displayed.

4.2. MLMM algorithmic pipeline

The system back-end components include an algorithmic pipeline employing (i) text extracted by a considered MLLM on a wine label image and (ii) a custom hierarchical search algorithm that skims a hierarchical textual DB based on the previously extracted words, providing the best candidate wine types. Those are respectively detailed in Section 4.2.1 and Section 4.2.2.

4.2.1 Textual Extraction module

The proposed system integrates a Textual Extraction Module which leverages a MLLM to extract structured information from wine labels by integrating visual and textual processing with domain-specific knowledge. As depicted in Fig. 2, the workflow begins with the user capturing an image of the wine label, denoted as I, using their smartphone. This image is processed by the MLLM, which combines visual features extracted via a vision encoder E_V and textual features extracted via an OCR-based text encoder E_T. The extracted embeddings are concatenated to form a multimodal representation, $z = \text{Concat}(E_V(I), E_T(I))$, which serves as input to the MLLM.

To contextualize the extraction process, a domain-specific prompt, P, is provided, detailing key attributes of interest such as wine region, vintage, and alcohol content. In our case $P =$ "*Report only the label text in the image. Also read the parts that are blurry, and barely visible. Reply only with the text you find without special characters and on a single line without punctuation.*" We here adopted such a very general prompt, to preliminary explore the capabilities of our MLLM to act as an OCR on images of wine labels. However, we constrain it to retrieve only classical text without special characters on a single line, to match the structure of the adopted OCR in [4].

Then, the MLLM exploits the cross-attention mechanism to refine the multimodal embeddings, yielding a prompt-conditioned representation. This aims at maximizing the MLLM parameters, θ_{MLLM}, to maximize the accuracy of attribute recognition.

This representation is decoded to generate structured data, S, containing a list of extracted attributes.

The extracted information is then passed to a custom search algorithm, which matches the data against a curated wine database to provide additional insights, such as reviews and geographic origins. Finally, the results are visualized through an augmented interface, enabling the user to explore the wine's contextual and sensory profile interactively.

4.2.2 Wine database and Search Algorithm

We considered a subset of the wine database produced in [4], which contains more than 2000 wines from the Emilia-Romagna region. Given the hierarchical nature of the attributes for wine discrimination (Section3), we adapted that classification to fit a hierarchical tree structure as in [4]. This transformation involves the definition of the level of

Figure 3. Frames from two wine bottles. Left three images are from ID 4, and the right three images are from ID 6.

trees according to our considered features: appellation, appellation value, effervescence, sweetness, color, and wine name. Each layer of the tree corresponds to different features (nodes) representing possible matching values (e.g., DOC, DOCG for appellation). The hierarchy is constructed by optimizing a cost function that balances pruning efficiency and search complexity, ensuring features with higher discriminatory power appear at the upper levels of the tree, as detailed in [4].

The hierarchical database supports efficient wine identification by employing a search algorithm. The Textual Extraction Module furnishes the text from wine labels, preprocessed the retrieved words to crop relevant areas, and eliminates errors or noise. The corrected text is then matched against the tree's feature values using a linear search algorithm with the Levenshtein distance. The algorithm then traverses the hierarchical tree level by level, selecting the most probable match at each node and pruning irrelevant branches. This traversal continues until the wine type is identified, with the remaining nodes representing potential matches. For features with default values, the algorithm assumes these defaults if no relevant words are detected. The hierarchical structure ensures scalability and adaptability to new data, while the search algorithm effectively compensates for the Textual Extraction Module inaccuracies.

5. Experiments and results

5.1. Dataset

For this study, we adopted a high-variance subset of 10 wine bottles, as introduced [4], selected specifically to evaluate the system's robustness under challenging conditions (e.g., occlusions, reflections, blur). This dataset includes wines with diverse label designs, font styles, and structural variations. The wine bottle labels considered amount to (Wine ID 1) Colli di Faenza Rosso Riserva Secco, (Wine ID 2) Colli di Rimini Biancame Secco, (Wine ID 3) Colli di Rimini Rebola Secco, (Wine ID 4) Colli di Imola Rosso Riserva Secco, (Wine ID 5) Emilia Lambrusco Frizzante Dolce, (Wine ID 6) Emilia Lambrusco Frizzante Secco, (Wine ID 7) Emilia Malvasia Frizzante Secco, (Wine ID 8) Forlì Sangiovese Secco, (Wine ID 9) Lambrusco Di Sorbara Rosso Frizzante Secco 1, (Wine ID 10) Lambrusco Di Sorbara Rosso Frizzante Secco 2.

We expanded the analysis from static images to keyframes extracted from videos to assess the system's capability to handle realistic scenarios. Each video was recorded while rotating the camera around the bottle from left to right, capturing the label from various angles and under varying lighting conditions. From each video, we extracted multiple keyframes, simulating real-world usability and helping evaluate the system's performance in mitigating OCR detection errors caused by factors such as reflections, distortions due to the bottle's curvature, or suboptimal environmental lighting. This methodology was designed to test the capability of the adopted MLLM pipeline when acting as a text recognition system, ensuring reliable recognition even in non-ideal conditions. Examples of key frames extracted by those videos for wine bottles with IDs 4 and 6 are visually reported in Figure 3.

From each of these samples, we extracted the ground truth words (wine attributes) written on the label, which amounts to the features we must match to recognize and discriminate a particular type of wine (see Section 3). A visual example is reported in Figure 4. This particular setup aims at creating a complex experimental setting: our considered MLLM model should match not only the attributes required to recognize a wine type but all the text reported in the label.

5.2. Experimental Setting

We employed, as our target MLLM, Llama 3.2, a foundational language and vision model, designed to support a large variety of AI tasks [9]. We selected this model not

Table 2. Comparison of Llama 3.2 and the legacy OCR on data for Matches, Missing, and Incorrect metrics. Values are reported as mean (standard deviation) calculated over frames and wine IDs. Bold values indicate superior performance per metric for each Wine ID.

Wine ID	Matches		Missing		Incorrect	
	Llama 3.2	OCR	Llama 3.2	OCR	Llama 3.2	OCR
1	7.57 (±2.15)	**9.86 (±1.07)**	3.43 (±2.15)	**1.14 (±1.07)**	**58.71 (±26.31)**	103.29 (±9.52)
2	**6.29 (±2.63)**	5.14 (±2.04)	**3.71 (±2.63)**	4.86 (±2.04)	**16.86 (±12.25)**	35.71 (±15.70)
3	7.86 (±2.54)	**10.00 (±2.71)**	5.14 (±2.54)	**3.00 (±2.71)**	**12.29 (±6.90)**	24.86 (±1.95)
4	8.56 (±1.88)	**10.67 (±2.35)**	5.44 (±1.88)	**3.33 (±2.35)**	**39.11 (±5.60)**	52.22 (±3.15)
5	6.43 (±1.13)	**10.00 (±1.73)**	5.57 (±1.13)	**2.00 (±1.73)**	**24.86 (±2.67)**	29.29 (±0.95)
6	**6.71 (±2.43)**	4.71 (±2.29)	**3.29 (±2.43)**	5.29 (±2.29)	**21.86 (±5.73)**	33.71 (±7.54)
7	4.86 (±1.57)	**7.14 (±2.85)**	5.14 (±1.57)	**2.86 (±2.85)**	**22.00 (±11.39)**	46.00 (±5.07)
8	**8.86 (±1.95)**	7.57 (±1.13)	**3.14 (±1.95)**	4.43 (±1.13)	**46.14 (±24.12)**	75.29 (±3.90)
9	**8.00 (±1.63)**	5.14 (±1.46)	**4.00 (±1.63)**	6.86 (±1.46)	**24.57 (±6.55)**	57.71 (±4.31)
10	**7.71 (±2.56)**	6.00 (±1.91)	**4.29 (±2.56)**	6.00 (±1.91)	**36.00 (±8.39)**	52.29 (±7.61)

Figure 4. Ground truth example labeled for Wine Bottle with ID 9.

only for its multimodal understanding but also because in its training pipeline, data optimized for document understanding (OCR-oriented) were included [9]. We adopted the 3B parameters version provided through the Ollama interface [1]. Then, to compare its performance concerning the legacy OCR module already included in the system introduced in [4] which amounts to a state-of-the-art OCR [10]. For each method, we considered all the possible extracted words.

Then, to have a fair comparison, we applied a trivial regular expression to post-process all the text extracted by both the MLLM and the classical OCR to filter out any form of punctuation. The extracted words were matched against a predefined list of ground truth words within our dataset to assess system performance. Quantitatively, we then calculate the **exact match accuracy** (number of correct words on the retrieved ones), **precision** (number of correct words in the retrieved ones concerning the total), and **recall** (number of missing words in the retrieved ones concerning the total). These are referred as respectively *Matches*, *Missing*, and *Incorrect*.

This experimental setting ensures that the generated word list aligns with the predefined ground truth words, enabling robust evaluation and analysis of system performance. All the experiments were conducted on a workstation, equipped with an Intel Xeon Gold CPU operating at 3.80 GHz, 96 GB of RAM, and an NVIDIA Quadro RTX 5000 GPU, with 16 GB of VRAM.

5.3. Results

Table 2 report the average performance of applying respectively the adopted Llama model and legacy OCR, over frames and different wine labels.

This comparative analysis between Llama 3.2 and traditional OCR systems on data reveals distinct strengths and limitations in text extraction tasks. In terms of *Matches*, Llama 3.2 outperforms OCR in 5 out of 10 Wine IDs, indicating an equal rate of correct text extractions in these instances. For example, Wine ID 2 shows Llama 3.2 achieving 6.29 (±2.63) matches compared to OCR's 5.14 (±2.04). However, OCR surpasses Llama 3.2 in the remaining Wine IDs, such as Wine ID 1, where OCR records 9.86 (±1.07) matches against Llama 3.2's 7.57 (±2.15).

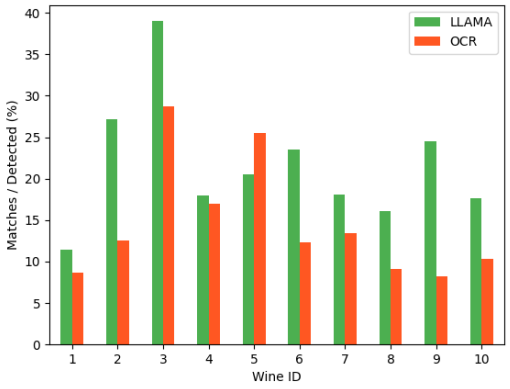
Regarding the *Missing* metric, Llama 3.2 demonstrates superior performance in 5 out of 10 Wine IDs, indicating fewer missed word extractions. For instance, Wine ID 2 shows Llama 3.2 with 3.71 (±2.63) missing extractions, while OCR has 4.86 (±2.04). Conversely, OCR exhibits better results in the remaining 5 Wine IDs, such as Wine ID 1, where it records 1.14 (±1.07) missing extractions compared to Llama 3.2's 3.43 (±2.15).

In the *Incorrect* metric, Llama 3.2 consistently outperforms OCR across all Wine IDs, indicating a lower rate of erroneous extractions. For example, Wine ID 1 shows Llama 3.2 with 58.71 (±26.31) incorrect extractions, significantly lower than OCR's 103.29 (±9.52). These findings indicate that while classical OCR has similar perfor-

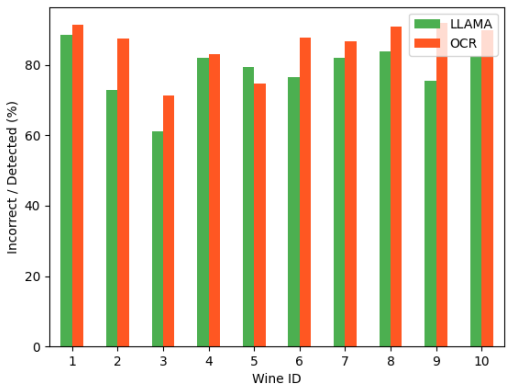
[1] https://ollama.com/library/llama3.2

mance for text extraction, Llama 3.2 may offer advantages in complex scenarios requiring a lower false positive rate. This is exactly the use case we here took into consideration, since to recognize a wine type, we have to match only the words needed to perform an algorithmic search, that otherwise would require a more complex filtering logic, as the one implemented in [4]. Despite this, Llama 3.2 exhibit a standard deviation higher than the OCR one, highlight a lack of robustness. This is likely due to the fact that the MLLM was subjected to frames where the textual information on the wine label was only partially visible—a scenario that deep learning-based OCR systems are designed to handle but can still struggle with under certain conditions.

To further analyze the performance of Llama 3.2 and traditional OCR systems, we present two bar plots in Figure 5 comparing the Matches (%) and Incorrect (%) metrics relative to detected words across various wine IDs.



(a) Matches (%)



(b) Incorrect (%)

Figure 5. Comparison of Matches, Incorrect in percentage to detected words

Figure 5a compares the percentages of Matches over all the detected words for both methods. It is evident that Llama 3.2 consistently achieves a higher ratio of correct detections relative to detected words across the majority of IDs. **Figure 5b** instead reports the percentages of Incorrect words over all the detected ones. Across all wine IDs, Llama 3.2 consistently reports a lower percentage of incorrect extractions compared to OCR. For instance, in Wine ID 4, the Incorrect (%) for Llama 3.2 is considerably lower than that of OCR, indicating better alignment with the ground truth.

These results further confirms the initial metrics: despite having similar text extraction performances, Llama 3.2 exhibit a lower false positive rate.

6. Conclusions

We here introduced the SOLLAMA system, a MLLM-driven situated augmented reality visualization system. It demonstrates the potential of MLLMs to simplify system logic deployment when implementing multimodal information retrieval systems, here contextualized for wine label recognition. By leveraging Llama 3.2, SOLLAMA effectively addresses challenges associated with traditional OCR-based methods, such as scalability, computational efficiency, and adaptability to diverse wine label designs. Experimental evaluations validate the superior performance of Llama 3.2 in handling complex scenarios providing a high recognition accuracy and reduced false positive rates.

SOLLAMA's innovative approach bridges the gap between AR interfaces and real-world product recognition, offering practical applications for the food and beverage industry and beyond.

Future research will explore a more varied mixture of dataset in the wine domain, along with the performance comparison inclusion of other OCRs and MLLMs. A the same time, we will set a robustness analysis to address the challenges posed by varying text visibility, lighting conditions, reflections, and contrast commonly encountered with wine labels on bottles, comparing OCR and LLMs. Moreover, we will extending SOLLAMA to other domains requiring complex document analysis, integrating real-time user feedback mechanisms, and improving the system's robustness to extreme environmental variations. At the same time, we will focus on improving our system prompt, to directly identify the attributes that we need in a pre-defined template that could be directly used for optimized search.

References

[1] Julian M Alston and Davide Gaeta. Reflections on the political economy of european wine appellations. *Italian Economic Journal*, 7(2):219–258, 2021.

[2] Jesús Omar Álvarez Márquez and Jürgen Ziegler. Improving the shopping experience with an augmented reality-enhanced shelf. *Mensch und Computer 2017-Workshopband*, 2017.

[3] J Amudha, Manmohan Singh Thakur, Anupriya Shrivastava, Shubham Gupta, Deepa Gupta, and Kshitij Sharma. Wild ocr: Deep learning architecture for text recognition in images. In *Proceedings of International Conference on Computing and Communication Networks: ICCCN 2021*, pages 499–506. Springer, 2022.

[4] Alessia Angeli, Lorenzo Stacchio, Lorenzo Donatiello, Alessandro Giacchè, and Gustavo Marfia. Making paper labels smart for augmented wine recognition. *The Visual Computer*, 40(8):5519–5531, 2024.

[5] Stevan Čakić, Tomo Popović, Stevan Šandi, Srdjan Krčo, and Anita Gazivoda. The use of tesseract ocr number recognition for food tracking and tracing. In *2020 24th International Conference on Information Technology (IT)*, pages 1–4. IEEE, 2020.

[6] Camera di Commercio Molise. Guida etichettature vino. https://www.molise.camcom.gov.it/sites/default/files/guida_etichettatura_vino.pdf, 2016.

[7] Steve Charters, Larry Lockshin, and Tim Unwin. Consumer responses to wine bottle back labels. *Journal of Wine Research*, 10(3):183–195, 1999.

[8] Sue Ding et al. *Re-enchanting spaces: location-based media, participatory documentary, and augmented reality*. PhD thesis, Massachusetts Institute of Technology, 2017.

[9] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[10] Easy Ocr. JadedAI. https://github.com/JaidedAI/EasyOCR, 2021.

[11] FEDERDOC. I VINI ITALIANI A DENOMINAZIONE D'ORIGINE 2020. https://www.federdoc.com/new/wp-content/uploads/2020/06/vini_italiani_denominazione_origine_2020.pdf, 2021.

[12] George W Fitzmaurice. Situated information spaces and spatially aware palmtop computers. In *Communications of the ACM*, volume 36, pages 39–49. ACM New York, NY, USA, 1993.

[13] Yuying Ge, Sijie Zhao, Ziyun Zeng, Yixiao Ge, Chen Li, Xintao Wang, and Ying Shan. Making llama see and draw with seed tokenizer. *arXiv preprint arXiv:2310.01218*, 2023.

[14] Timnit Gebru, Oren Hazi, and Vickey Yeh. Mobile wine label recognition. 2022.

[15] Venugopal Gundimeda, Ratan S Murali, Rajkumar Joseph, and NT Naresh Babu. An automated computer vision system for extraction of retail food product metadata. In *First International Conference on Artificial Intelligence and Cognitive Computing*, pages 199–216. Springer, 2019.

[16] Aditya Gunturu, Shivesh Jadon, Nandi Zhang, Jarin Thundathil, Wesley Willett, and Ryo Suzuki. Realitysummary: On-demand mixed reality document enhancement using large language models. *arXiv preprint arXiv:2405.18620*, 2024.

[17] Shirin Hajahmadi, Lorenzo Stacchio, Alessandro Giacché, Pasquale Cascarano, and Gustavo Marfia. Investigating extended reality-powered digital twins for sequential instruction learning: the case of the rubik's cube. In *2024 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 259–268. IEEE, 2024.

[18] Oliver Hinz, Jochen Eckert, and Bernd Skiera. Drivers of the long tail phenomenon: an empirical analysis. *Journal of management information systems*, 27(4):43–70, 2011.

[19] Bin Hu, Nuoya Zhou, Qiang Zhou, Xinggang Wang, and Wenyu Liu. Diffnet: a learning to compare deep network for product recognition. *IEEE Access*, 8:19336–19344, 2020.

[20] Jeong-Mun Jung, Hyung-Jeong Yang, Soo-Hyung Kim, Guee-Sang Lee, and Sun-Hee Kim. Wine label recognition system using image similarity. *The Journal of the Korea Contents Association*, 11(5):125–137, 2011.

[21] Bianca Lamm and Janis Keuper. Can visual language models replace ocr-based visual question answering pipelines in production? a case study in retail. *arXiv preprint arXiv:2408.15626*, 2024.

[22] Bianca Lamm and Janis Keuper. Retail-786k: a large-scale dataset for visual entity matching, 2024.

[23] Benjamin Lee, Michael Sedlmair, and Dieter Schmalstieg. Design patterns for situated visualization in augmented reality. *IEEE Transactions on Visualization and Computer Graphics*, 2023.

[24] Xiaoqing Li, Jiansheng Yang, and Jinwen Ma. Cnn-sift consecutive searching and matching for wine label retrieval. In *International Conference on Intelligent Computing*, pages 250–261. Springer, 2019.

[25] Chia Xin Liang, Pu Tian, Caitlyn Heqi Yin, et al. A comprehensive survey and guide to multimodal large language models in vision-language tasks. *arXiv preprint arXiv:2411.06284*, 2024.

[26] Mingyuan Lin, Longhua Ma, and Binchao Yu. An efficient and light-weight detector for wine bottle defects. In *2020 16th International Conference on Control, Automation, Robotics and Vision (ICARCV)*, pages 957–962. IEEE, 2020.

[27] Yuliang Liu, Biao Yang, Qiang Liu, et al. Textmonkey: An ocr-free large multimodal model for understanding document. *arXiv preprint arXiv:2403.04473*, 2024.

[28] livingwinelabels. livingwinelabels. https://www.livingwinelabels.com/, 2021.

[29] Nuno Cid Martins, Bernardo Marques, João Alves, Tiago Araújo, Paulo Dias, and Beatriz Sousa Santos. Augmented reality situated visualization in decision-making. *Multimedia Tools and Applications*, 81(11):14749–14772, 2022.

[30] Michele A. Fino. Questione di Etichetta. https://www.spazioprever.it/salabar/vino/pdf/Questione_di_etichetta.pdf, 2013.

[31] Khaled Salah Mohamed. Deep learning for spatial computing: augmented reality and metaverse "the digital universe". In *Deep Learning-Powered Technologies: Autonomous Driving, Artificial Intelligence of Things (AIoT), Augmented Reality, 5G Communications and Beyond*, pages 131–150. Springer, 2023.

[32] In Seop Na, Yan Juan Chen, and Soo Hyung Kim. Automatic segmentation of product bottle label based on grabcut algorithm. *International Journal of Contents*, 10(4):1–10, 2014.

[33] Michael Nebeling and Katy Madier. 360proto: Making interactive virtual reality & augmented reality prototypes from paper. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2019.

[34] Stefano Orsini and Daniela Magro. Augmented reality applications in the food industry. *Journal of Food Engineering*, 214:1–8, 2017.

[35] Lara Penco, Francesca Serravalle, Giorgia Profumo, and Milena Viassone. Mobile augmented reality as an internationalization tool in the "made in italy" food and beverage industry. *Journal of Management and Governance*, 25(4):1179–1209, 2021.

[36] Vincent Perot, Kai Kang, Florian Luisier, Guolong Su, Xiaoyu Sun, Ramya Sree Boppana, et al. Lmdx: Language model-based document information extraction and localization. *arXiv preprint arXiv:2309.10952*, 2023.

[37] Valentina Pini, Valeria Orso, Patrik Pluchino, and Luciano Gamberini. Augmented grocery shopping: fostering healthier food purchases through ar. *Virtual Reality*, 27(3):2117–2128, 2023.

[38] PTC. Vivino and Vuforia's Image Recognition Solution Make a Great Pairing. `https://www.ptc.com/en/case-studies/vivino`, 2022.

[39] Abderahman Rejeb, Karim Rejeb, and John G Keogh. Enablers of augmented reality in the food supply chain: a systematic literature review. *Journal of Foodservice Business Research*, 24(4):415–444, 2021.

[40] Ho-Sub Ryu and Hanhoon Park. A system for supporting paper-based augmented reality. *Multimedia Tools and Applications*, 75:3375–3390, 2016.

[41] Nareen OM Salim, Subhi RM Zeebaree, Mohammed AM Sadeeq, AH Radie, Hanan M Shukur, and Zryan Najat Rashid. Study for food recognition system using deep learning. In *Journal of Physics: Conference Series*, volume 1963, page 012014. IOP Publishing, 2021.

[42] Andreas Sonderegger, Delphine Ribes, Nicolas Henchoz, and Emily Groves. Food talks: visual and interaction principles for representing environmental and nutritional food information in augmented reality. In *2019 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, pages 98–103. IEEE, 2019.

[43] Lorenzo Stacchio, Vincenzo Armandi, Lorenzo Donatiello, and Gustavo Marfia. Annholotator: A mixed reality collaborative platform for manufacturing work instruction interaction. In *2023 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pages 418–424. IEEE, 2023.

[44] Lorenzo Stacchio, Shirin Hajahmadi, and Gustavo Marfia. Preserving family album photos with the hololens 2. In *2021 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pages 643–644. IEEE, 2021.

[45] Lorenzo Stacchio, Giacomo Vallasciani, Giulio Augello, Silvano Carrador, Pasquale Cascarano, and Gustavo Marfia. Wixard: Towards a holistic distributed platform for multiparty and cross-reality webxr experiences. In *2024 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pages 264–272. IEEE, 2024.

[46] Jannis Strecker, Kimberly García, Kenan Bektaş, Simon Mayer, and Ganesh Ramanathan. Socrar: Semantic ocr through augmented reality. In *Proceedings of the 12th International Conference on the Internet of Things*, pages 25–32, 2022.

[47] Georgios D Styliaras. Augmented reality in food promotion and analysis: Review and potentials. *Digital*, 1(4):216–240, 2021.

[48] Markus Tatzgern. Situated visualization in augmented reality. *Graz University of Technology*, 2015.

[49] TinEye. Wineengine is image recognition for the beverage industry. `https://services.tineye.com/WineEngine`, 2021.

[50] Vittorio Portinari. Elementi di Legislazione Vitivinicola: le norme per l'etichettatura e la tracciabilità dei vini. `http://www.sardegnaagricoltura.it/documenti/14_43_20160531144229.pdf`, 2016.

[51] Vivino. Vivino. `https://www.vivino.com/`, 2021.

[52] Vuforia. Vuforia SDK. `https://developer.vuforia.com/downloads/SDK`, 2022.

[53] Dongsheng Wang, Natraj Raman, Mathieu Sibue, et al. Docllm: A layout-aware generative language model for multimodal document understanding. *arXiv preprint arXiv:2401.00908*, 2023.

[54] Mei-Yi Wu, Jia-Hong Lee, and Shu-Wei Kuo. A hierarchical feature search method for wine label image recognition. In *2015 38th International Conference on Telecommunications and Signal Processing (TSP)*, pages 568–572. IEEE, 2015.

[55] Lilong Xue, Dan Zhang, Yuxiao Dong, and Jie Tang. Autore: Document-level relation extraction with large language models. *arXiv preprint arXiv:2403.14888*, 2024.

[56] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *National Science Review*, page nwae403, 2024.

[57] Yuka. Yuka. `https://yuka.io/it/`, 2021.

[58] et al. Zhang. Large language models for generative information extraction: a survey. *Journal of Computer Science and Technology*, 39(5):987–1003, 2024.

[59] Qintong Zhang, Victor Shea-Jay Huang, Bin Wang, Junyuan Zhang, Zhengren Wang, Hao Liang, Shawn Wang, Matthieu Lin, Wentao Zhang, and Conghui He. Document parsing unveiled: Techniques, challenges, and prospects for structured information extraction. *arXiv preprint arXiv:2410.21169*, 2024.

[60] Lili Zhu, Petros Spachos, Erica Pensini, and Konstantinos N Plataniotis. Deep learning and machine vision for food processing: A survey. *Current Research in Food Science*, 4:233–249, 2021.