Towards Wine Tasting Activity Recognition for a Digital Sommelier

Mario O. Parra
marioparra@cicese.edu.mx
CICESE
Ensenada, Baja California, Mexico

Jesus Favela
favela@cicese.mx
CICESE
Ensenada, Baja California, Mexico

Luis A. Castro
luis.castro@acm.org
Sonora Institute of Technology (ITSON)
Ciudad Obregon, Mexico

Daniel Gatica-Perez
gatica@idiap.ch
Idiap Research Institute & EPFL
Switzerland

Abstract

In this study, we evaluated the feasibility of using zero-shot classification models for activity recognition in a Digital Sommelier. Our experiment involved preprocessing video data by extracting frames and categorizing user activities related to a wine-tasting scenario. Image classification models demonstrated high accuracy, nearing 90%, in distinguishing between "engaged" and "disengaged" states. However, video classification models presented a lower performance in classifying user activities such as "observing wine," "smelling wine," and "sipping wine," with an average accuracy of around 50% due to the interdependent nature of the activities. Despite these challenges, our findings highlight the potential of zero-shot classification models in enhancing virtual assistants' ability to recognize and respond to user activities.

CCS Concepts

• **Human-centered computing** → **Empirical studies in ubiquitous and mobile computing**; **Empirical studies in HCI**;

Keywords

Digital Commensality, Conversational Agents, Human-Food Interaction

ACM Reference Format:
Mario O. Parra, Jesus Favela, Luis A. Castro, and Daniel Gatica-Perez. 2024. Towards Wine Tasting Activity Recognition for a Digital Sommelier. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI Companion '24), November 4–8, 2024, San Jose, Costa Rica.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3686215.3686217

1 Introduction

Digital Commensality, the practice of sharing meals through digital means, is gaining interest as technology advances [3, 21]. Integrating conversational agents into this experience can enhance user interaction and create a more engaging environment [13, 18]. One promising application is in wine tasting, where a Digital Sommelier can guide users through the process, providing information and recommendations in real time [17]. This paper explores the use of pre-trained zero-shot classification models [24] to recognize and classify user activities during a wine-tasting session.

Activity recognition is a widely explored field that seeks to develop systems capable of identifying people's actions based on data obtained from sensors [22]. While there is existing work that uses wearables [9] and cameras [8] with activity recognition models to infer user actions, recognizing eating activity, especially to assist a virtual assistant in decision-making, remains an underexplored area. Regarding alcohol consumption, activity recognition has demonstrated significant potential in identifying and analyzing drinking behaviors through mobile phone sensors [2]. Research like the DrinkSense project has leveraged accelerometer data, location services, and other smartphone sensors to effectively classify drinking events [20]. However, the application of computer vision models, particularly in the context of wine tasting with a digital assistant, remains largely unexplored.

In this study, we conducted experiments with pre-trained zero-shot classification models to evaluate their effectiveness in recognizing specific user activities related to wine tasting. This paper contains the following contributions: First, we propose the use of pre-trained zero-shot models for recognizing wine-tasting activities, offering a scalable alternative to custom-built models that may not be adaptable to different scenarios [16]. Second, we introduce a novel dataset derived from recorded sessions with participants, capturing various activities such as observing, smelling, and sipping wine. Finally, we analyze the performance of image and video classification models from platforms like HuggingFace [23] to determine their suitability for this application

The findings from this experiment provide insights into the strengths and limitations of using zero-shot classification models for activity recognition in a conversational agent context. This research not only demonstrates the potential of these models in enhancing user interaction but also highlights areas for future improvement and development.

2 Related Work

In the context of alcohol consumption, activity recognition has shown promising results in detecting and characterizing drinking

ICMI Companion '24, November 4–8, 2024, San Jose, Costa Rica

ACM ISBN 979-8-4007-0463-5/24/11
https://doi.org/10.1145/3686215.3686217

behaviors using mobile phone sensors [2]. Studies such as those on DrinkSense have used data from accelerometers, location services, and other smartphone sensors to accurately classify drinking events [20]. Another study integrated crowdsensing data with Instagram social media posts to classify alcohol-related activities, achieving an accuracy of 82.3% using image features and contextual cues [19]. Similarly, other research has focused on detecting fluid intake using wearable devices. Efforts have been made to identify hand gestures using smartwatches to determine when and how often a person drinks liquids throughout the day [11, 12]. These findings suggest that mobile phones and wearable sensors can effectively monitor and identify drinking episodes. However, individuals need to wear devices equipped with models specifically trained and customized for these activities.

A recent effort in activity recognition has focused on commensality [16], which aims to identify actions such as eating or talking. However, there is still a long way to go to achieve accurate and reliable recognition of eating activity in different contexts and considering the diversity of actions that can be performed during a meal [1]. One potential application of this work lies in the activity of wine tasting. Alcohol consumption, in general, is a socially relevant activity that is influenced by the social context in which it takes place [14]. Some studies focus on understanding user activity in eating contexts using multimodal data from smartphones through mobile sensors [6]. Other studies focus on identifying and tracking when people eat, using video and audio analysis [5]. However, the analysis of user behavior in the specific context of wine tasting with a virtual agent has not been explored to our knowledge.

Currently, databases such as Kinetics 400 are available [7], containing 400 human action classes specifically used for training activity recognition models. In this database, classes like "drinking beer," "tasting beer," and "opening bottle" are included. Similarly, the Kinetics 600 database includes other classes [4], such as "tasting wine," "sipping cup," and "opening wine bottle." These classes are similar to those used in this experiment. However, they do not exactly reflect what we aim to measure. For instance, the class "tasting wine" is very general and can encompass any of the three activities: observing, smelling, and sipping wine. Therefore, using a dataset with examples of these three specific classes to differentiate activities within a wine-tasting interaction represents a new contribution to the field.

3 Methods

3.1 A Digital Sommelier

The Digital Sommelier (DS) is a conversational agent that guides users through wine-tasting sessions [17]. The Digital Sommelier employs OpenAI's "gpt-3.5-turbo-0125" model for conversational interaction, configured to emulate a sommelier with specialized prompts and wine-tasting transcript examples.

Avatar embodiment and animation included in the sommelier, using tools such as Dall-E and the d-id.com API, create lifelike interactions by combining text and images into videos. Voice data is converted to text using a speech-to-text web interface, and text is transformed into speech using Amazon Polly Voices. The prototype's database includes comprehensive wine information, including grape variety, colors, aromas, pairing suggestions, winery details, and wine preparation methods.

For a future prototype, an activity recognition module will be implemented to infer user activities such as "pick up glass," "smell wine," "observe wine," and more. In general, the Digital Sommelier integrates language processing, image and video generation, and speech recognition to create an interactive, user-friendly wine-tasting experience.

3.2 Dataset

During a previous experiment with the Digital Sommelier, tests and interactions were conducted with 30 participants. A total of 30 subjects participated in this experiment, of whom 15 (50%) were women. The average age of the participants was 28.67 years (SD = 7.11). Most participants had little or no experience with wine tasting. The sessions were conducted in Spanish and recorded in a laboratory at the institute, and all participants signed a consent form that assured them of their anonymity. A setup was arranged with a monitor and a microphone for participants to interact with the virtual agent (see Figure 1). Participants sat in a chair facing a table with wine utensils and the Digital Sommelier.

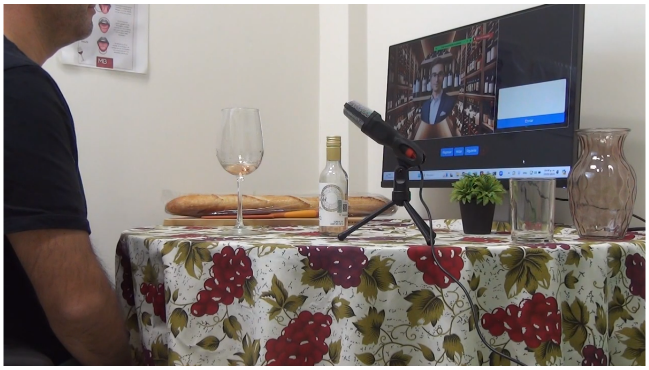

Figure 1: Experimental setup

Each session lasted approximately 15 minutes, and each participant interacted twice with the conversational agent. The sessions were recorded using two cameras: one positioned behind the participant to capture the screen displaying the avatar and a front-facing camera to record the participant's expressions and reactions.

Subsequently, during the analysis of the recordings with the participants, the front-facing videos were labeled with the following four classes: observing wine, smelling wine, sipping wine, and being disengaged. Video clips of approximately 3 seconds were extracted to represent these activities.

To explore a future implementation of activity recognition, video clips were extracted to test various image and video classification models. 220 images (samples in Figure 2) divided into 4 different classes: 55 observing wine, 55 smelling wine, 55 sipping wine, and 55 disengaged. 90 video clips divided into three classes: 30 observing wine, 30 smelling wine, and 30 sipping wine. These images and videos were extracted from 11 of the 30 participants, selected based on their better positioning within the camera frame of the frontal camera. The videos from the remaining participants will be used to test and compare with our own activity recognition models in future research.

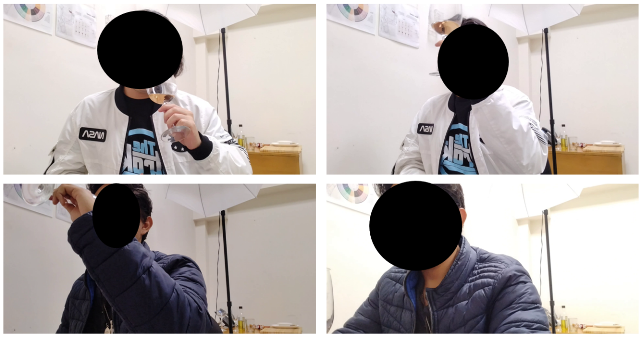

Figure 2: Wine tasting activities. (top-left) smelling wine; (top-right) observing wine; (bottom-left) sipping wine; (bottom-right) disengaged

3.3 Models

This experiment uses existing classification models from the Hugging Face platform. These models have been previously trained and are specifically utilized for zero-shot classification. We selected the most commonly used models for testing by filtering the Hugging Face site (as of June 2024) based on the highest number of downloads.

- **laion/CLIP-ViT-H-14-laion2B-s32B-b79K**: This model is a variant of the CLIP architecture, utilizing a Vision Transformer (ViT) with a large capacity (ViT-H-14).
- **laion/CLIP-ViT-B-32-laion2B-s34B-b79K**: Another CLIP model variant employs a Vision Transformer with a smaller capacity (ViT-B-32).
- **openai/clip-vit-large-patch14**: Developed by OpenAI, this CLIP model uses a Vision Transformer with a large patch size (ViT-large-patch14).
- **openai/clip-vit-base-patch32**: This model, also from OpenAI, features a Vision Transformer with a base capacity and a patch size of 32 (ViT-base-patch32).

Additionally, this experiment tested models used for zero-shot video classification. The most commonly used models on the platform were also selected:

- **microsoft/xclip-base-patch32**: This model, developed by Microsoft, is a variant of the XCLIP architecture. It utilizes a Vision Transformer with a base capacity and a patch size of 32 (ViT-base-patch32).
- **microsoft/xclip-large-patch14**: Another model from Microsoft, this one features a larger Vision Transformer architecture with a patch size of 14 (ViT-large-patch14).

To implement a future Digital Sommelier, image classification models will be primarily used due to their lower computational demands than video models, although both may be necessary to accurately classify similar activities such as observing, smelling, and sipping wine. The Digital Sommelier must first determine whether the user is engaged in wine tasting or merely interacting with the wine glass. After establishing this, video classification models will identify the specific activities of observing, smelling, and sipping wine.

4 Results

4.1 Engaged vs. Disengaged Classification

First, the image classification models were tested to determine if they could distinguish whether a person is "engaged" or "disengaged" in wine tasting. For this purpose, images corresponding to the classes "observing wine," "smelling wine," and "sipping wine" were grouped into a single class called "engaged." However, to use these images in the classifiers, a more specific text corresponding to the activity of being engaged in wine tasting was needed. Therefore, the text "raising a glass of wine" vs. "other" was used, as in all three mentioned activities, the person always raises the wine glass.

According to Table 1, all models demonstrated an average efficiency above 84% and just below 90%. Interestingly, the openai/clip-vit-large-patch14 model demonstrates a varied performance, excelling in the smelling and sipping categories (each at 100%) but showing lower accuracy in the disengaged category (56.36%), resulting in an average score of 86.82%. The openai/clip-vit-base-patch32 model performs the best on average, with a high average score of 89.55%, particularly strong in the Sipping category (98.18%) and balanced performance in other categories. This table highlights the strengths and weaknesses of each model in detecting different engagement levels, providing insights into their suitability for zero-shot classification tasks in similar contexts.

4.2 Activity Classification

For the video analysis, zero-shot classification models were utilized. Initially, the videos were preprocessed to make them suitable for analysis by extracting 8 uniformly distributed frames from each video. These frames were then fed into the models, which were tasked with classifying them into one of three possible categories: "observing wine," "smelling wine," and "sipping wine." Each class consisted of 30 videos for the models to analyze.

According to Table 2, the xclip-base-patch32 model shows a low performance with an average score of 36.67%. This model performs best in the "observing" category with a score of 60%, but its performance drops significantly in the "smelling" category with a score of only 6.67%. The xclip-large-patch14 model demonstrates a slightly higher overall performance with an average score of 58.89%. This model shows a balanced performance across all categories, improving particularly in the "sipping" category with a score of 70%.

5 Discussion

The results of this experiment helped us evaluate the feasibility of using existing models for activity classification in a scenario involving a Digital Sommelier. An important finding was that image classification models can help determine if a person is "engaged or disengaged" in wine tasting. This required an additional class to encompass when a person is engaged, and the model had to be correctly trained to identify a person engaged in wine tasting. This

Table 1: Performance of Zero-Shot Classification Models in Detecting Engagement (Engaged vs. Disengaged)

Model Name	Disengaged	Observing	Smelling	Sipping	Avg
laion/CLIP-ViT-H-14-laion2B-s32B-b79K	96.36	83.64	83.64	90.91	88.6375
laion/CLIP-ViT-B-32-laion2B-s34B-b79K	87.27	81.81	80.00	89.09	84.5425
openai/clip-vit-large-patch14	56.36	90.91	100.00	100.00	86.8175
openai/clip-vit-base-patch32	83.64	83.63	92.73	98.18	89.5450

Table 2: Performance of Zero-Shot Classification Video Models for Activity Recognition

Model Name	Observing	Smelling	Sipping	Avg
microsoft/xclip-base-patch32	60.00	6.67	43.33	36.67
microsoft/xclip-large-patch14	50.00	56.67	70.00	58.89

was addressed by providing a class to recognize when the person raises the wine glass. The results were positive, with the models achieving nearly 90% accuracy in correctly identifying engagement.

However, when identifying each wine-tasting activity, the models' efficiency was lower due to the similarity and interdependence of actions. Video classification was necessary to capture movement, but activities like sipping and observing wine looked similar in initial frames. For example, as shown in Figure 3, initial frames (f1 and f2) of sipping can be mistaken for observing, and just before sipping (f3), it appears the person is merely smelling the wine. Monitoring the entire movement of the glass or hand might be a better approach, though this introduces the challenge of determining the optimal time to start and stop recording.

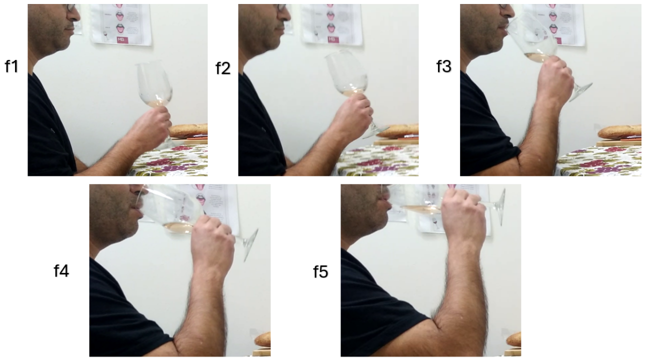

Figure 3: Sequence of Frames: Participant Sipping Wine

Despite this, it was demonstrated that using pre-trained models for zero-shot classification can be helpful in a virtual assistant aiming to recognize user activity. Applying these models in an agent could enhance the user experience. By having knowledge of the user's activity, the agent could, for instance, in the context of wine tasting, detect if the user is "disengaged" from the activity and suggest trying another wine or engage in a discussion about their opinion on what they are tasting. This would make the conversational agent a proactive tool [10], potentially improving the overall user experience.

5.1 Future Work

With regard to the implementation of a real-time model for supporting a virtual assistant such as a Digital Sommelier, one line of future work would be the implementation of multimodal models. In addition to the video with images of the user, the generated audio would be considered to verify whether the person is speaking. An example of this approach can be found in the work [15], where a multimodal approach is used that employs the person's speech state to help detect head nods in a face-to-face interaction.

On the other hand, it would be interesting to compare the generation of classification models specifically designed for this wine-tasting scenario. While this type of model may not be scalable for other scenarios such as cooking, it is important to compare them with existing models for "zero-shot" cases and make an informed decision about the type of model to use.

6 Conclusion

This paper presented the use of pre-trained classification models. Image classification models were able to identify with near 90% accuracy when a person is engaged or disengaged in wine tasting. However, using video classification models to identify specific activities such as "observing wine," "smelling wine," and "sipping wine" yielded less favorable results, with an average accuracy of around 60%. Despite this, there is potential for using zero-shot classification models to recognize user activities and enhance the user experience with a virtual assistant. Additionally, we utilized a private novel dataset derived from recorded sessions with participants, capturing a range of wine-tasting activities, which we can use for future experiments and the creation of models tailored to these scenarios.

7 Acknowledgements

This work was partly supported by the AI4Media project (Junior Fellow program), funded by the European Commission (Grant 951911) under the H2020 Programme ICT-48-2020. It was also partially funded by the CONAHCYT through a scholarship provided to the first author.

References

[1] Ferran Altarriba Bertran, Samvid Jhaveri, Rosa Lutz, Katherine Isbister, and Danielle Wilde. 2019. Making sense of human-food interaction. In *Proceedings of*

the 2019 CHI Conference on Human Factors in Computing Systems. 1–13.

[2] Sangwon Bae, Denzil Ferreira, Brian Suffoletto, Juan C Puyana, Ryan Kurtz, Tammy Chung, and Anind K Dey. 2017. Detecting drinking episodes in young adults using smartphone-based sensors. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 1, 2 (2017), 1–36.

[3] Mimi Bocanegra, Mailin Lemke, Roelof AJ De Vries, and Geke DS Ludden. 2022. Commensality or reverie in eating? Exploring the solo dining experience. In *Proceedings of the 2022 International Conference on Multimodal Interaction*. 25–35.

[4] Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. 2018. A short note about kinetics-600. *arXiv preprint arXiv:1808.01340* (2018).

[5] Simone Hantke, Maximilian Schmitt, Panagiotis Tzirakis, and Björn Schuller. 2018. EAT- The ICMI 2018 Eating Analysis and Tracking Challenge. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*. 559–563.

[6] Nathan Kammoun, Lakmal Meegahapola, and Daniel Gatica-Perez. 2023. Understanding the Social Context of Eating with Multimodal Smartphone Sensing: The Role of Country Diversity. In *Proceedings of the 25th International Conference on Multimodal Interaction*. 604–612.

[7] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. 2017. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950* (2017).

[8] Shian-Ru Ke, Hoang Le Uyen Thuc, Yong-Jin Lee, Jenq-Neng Hwang, Jang-Hee Yoo, and Kyoung-Ho Choi. 2013. A review on video-based human activity recognition. *Computers* 2, 2 (2013), 88–131.

[9] Oscar D Lara and Miguel A Labrador. 2012. A survey on human activity recognition using wearable sensors. *IEEE communications surveys & tutorials* 15, 3 (2012), 1192–1209.

[10] Lizi Liao, Grace Hui Yang, and Chirag Shah. 2023. Proactive conversational agents. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*. 1244–1247.

[11] Rainer Lutze. 2022. Practicality of automatic monitoring sufficient fluid intake for older people. In *2022 IEEE 10th International Conference on Healthcare Informatics (ICHI)*. IEEE, 330–336.

[12] Rainer Lutze and Klemens Waldhör. 2023. Practicality Aspects of Automatic Fluid Intake Monitoring via Smartwatches. In *International Conference on Human-Computer Interaction*. Springer, 67–86.

[13] Maurizio Mancini, Radoslaw Niewiadomski, Gijs Huisman, Merijn Bruijnes, and Conor Patrick Gallagher. 2020. Room for one more?-introducing artificial commensal companions. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–8.

[14] Lakmal Meegahapola, Florian Labhart, Thanh-Trung Phan, and Daniel Gatica-Perez. 2021. Examining the social context of alcohol drinking in young adults with smartphone sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 3 (2021), 1–26.

[15] Laurent Nguyen, Jean-Marc Odobez, and Daniel Gatica-Perez. 2012. Using self-context for multimodal detection of head nods in face-to-face interactions. In *Proceedings of the 14th ACM international conference on Multimodal interaction*. 289–292.

[16] Radoslaw Niewiadomski, Gabriele De Lucia, Gabriele Grazzi, and Maurizio Mancini. 2022. Towards Commensal Activities Recognition. In *Proceedings of the 2022 International Conference on Multimodal Interaction*. 549–557.

[17] Mario Parra, Jesús Favela, and Luis A. Castro. 2024. A Digital Sommelier to Enhance Multisensory Wine Tasting. In *ACM International Conference on Interactive Media Experiences (IMX) Workshop: IMX in Latin America - 2nd International Workshop*.

[18] Mario O Parra, Luis A Castro, and Jesus Favela. 2023. Enhancing Well-being Through Food: A Conversational Agent for Mindful Eating and Cooking. In *Adjunct Proceedings of the 2023 ACM International Joint Conference on Pervasive and Ubiquitous Computing & the 2023 ACM International Symposium on Wearable Computing*. 423–427.

[19] Thanh-Trung Phan, Skanda Muralidhar, and Daniel Gatica-Perez. 2019. Drinks & crowds: Characterizing alcohol consumption through crowdsensing and social media. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 2 (2019), 1–30.

[20] Darshan Santani, Florian Labhart, Sara Landolt, Emmanuel Kuntsche, Daniel Gatica-Perez, et al. 2018. DrinkSense: Characterizing youth drinking behavior using smartphones. *IEEE Transactions on Mobile Computing* 17, 10 (2018), 2279–2292.

[21] Charles Spence, Maurizio Mancini, and Gijs Huisman. 2019. Digital commensality: Eating and drinking in the company of technology. *Frontiers in psychology* 10 (2019), 460197.

[22] Michalis Vrigkas, Christophoros Nikou, and Ioannis A Kakadiaris. 2015. A review of human activity recognition methods. *Frontiers in Robotics and AI* 2 (2015), 28.

[23] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771* (2019).

[24] Yongqin Xian, Zeynep Akata, Gaurav Sharma, Quynh Nguyen, Matthias Hein, and Bernt Schiele. 2016. Latent embeddings for zero-shot classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 69–77.