AR Cooking Assistant

A Framework and Methods for Automatic Recipe Step Detection

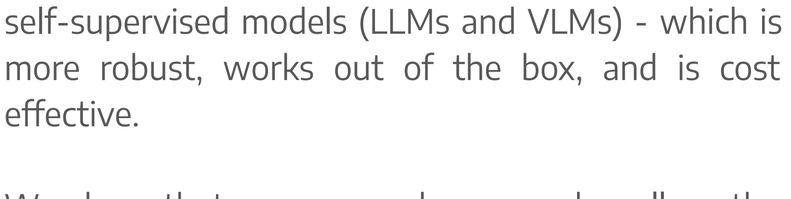
Paul Han, Alvin Le

PAUL G. ALLEN SCHOOL

The Problem

Augmented reality (AR) cooking assistants help users find recipes and cook in real time. A key problem in this application is the automatic recipe step detection problem (ARSD), which is formulated as follows:

Given an input camera stream of the kitchen and a step-by-step recipe, can we identify the exact step that is currently executing?



Related Works

Prior works have used

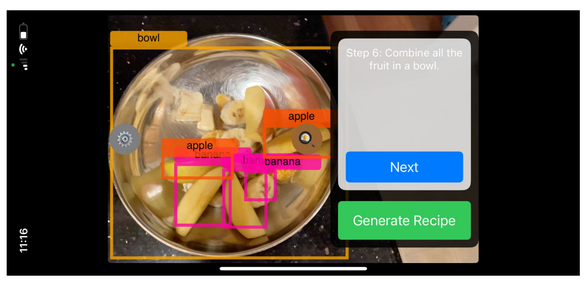
1. Specialized sensors to track objects and probabilistic models to describe actions in the kitchen [2].
2. Supervised methods by training on large datasets of cooking images, text, and videos [3].

Limitations of such methods include costs of fine-tuning and annotation, or brittleness of probabilistic models.

Our Approach

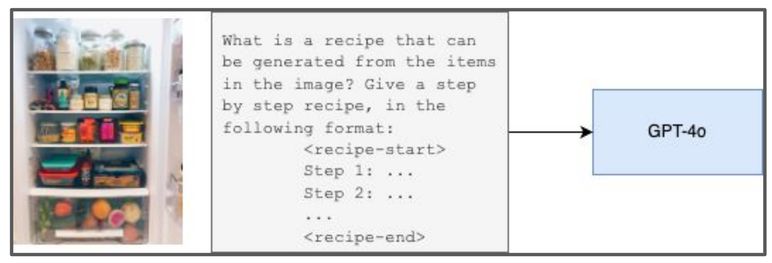
We propose using approaches based on self-supervised models (LLMs and VLMs) - which is more robust, works out of the box, and is cost effective.

We show that our approach can work well on the ARSD task via evaluation and qualitative analysis, by building a working prototype for recipe generation and automated step detection.

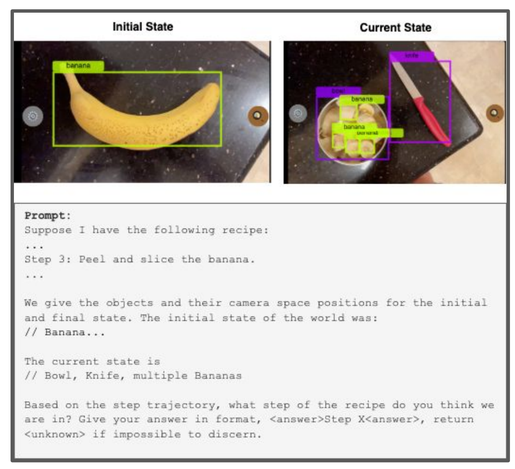


Methods

Recipe generation:
For recipe generation, we use a simple method. On user input, we sample the current camera frame and pass it to a VLM, querying for a recipe based on the ingredients in the image.



ARSD Method 1 (YOLO + LLM):
Our first method consists of three phases.
(1) extract food/kitchen objects using YOLO [5]
(2) build a history of objects and their camera space positions over time via text coordinates
(3) provide the LLM with the recorded object histories and ask to infer the most likely step



```
Prompt:
Suppose I have the following recipe:
...
Step 3: Peel and slice the banana.
...

Based on the image, what step of the recipe do you
think we are in? Give your answer in format,
<answer>Step X<answer>, return <unknown> if impossible
to discern. If there is any doubt, default to
<unknown>.
```

ARSD Method 2 (VLM Only):
Our second method for ARSD is simpler and relies directly on a unified vision-language model (VLM). To determine the current step, we capture the current frame and prompt the VLM to infer the most likely step. Optionally, we provide the recipe generating image as a point of reference.

Evaluation

Evaluation Task:
To quantitatively compare Method 1 and Method 2 on ARSD, we construct a step-by-step fruit salad recipe.

- Each step is represented by multiple images under varying conditions.
- The unknown category includes occluded or misleading images that should not correspond to any valid step.

We evaluate how accurate our two methods are on detecting the correct ground-truth step given the input image. Results show trade offs between cost, latency, and accuracy.

Method	Avg. Latency (s)	Accuracy (%)	Cost per Request ($)
Method 1 (YOLO + LLM)	3.34	88%	0.0007875
Method 2 (VLM)	6.91	100%	0.0028575

Table 1. Performance comparison between Method 1 and Method 2 for ARSD.

References

[1] Mehdi Mekni and André Lemieux. Augmented reality : Applications , challenges and future trends. 2014.
[2] M. Philipose, K.P. Fishkin, M. Perkowitz, D.J. Patterson, D. Fox, H. Kautz, and D. Hahnel. Inferring activities from interactions with objects. IEEE Pervasive Computing, 3(4):50–57, 2004.
[3] Luowei Zhou, Chenliang Xu, and Jason J. Corso. Procnets: Learning to segment procedures in untrimmed and unconstrained videos. CoRR, abs/1703.09788, 2017.
[4] Wes Gurnee and Max Tegmark. Language models represent space and time, 2024
[5] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection, 2016
[6] Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey, 2024.

Conclusions

- Using text-only context is more efficient (Method 1), but using visual modality directly is more accurate (Method 2).
- Combination of approaches may be appropriate, for example simpler recipes are better suited for text based approach, longer recipes better with VLMs.
- Better, more extensive evaluation for ARSD is necessary. We only evaluate on one simple recipe with fixed steps.
- Generative AI has a lot of promising directions for AR assistant applications.