Smart Kitchen: A Multi-Modal AR System

Andrew Mendez
Cornell Tech
aem336@cornell.edu

Zexi Liu
Cornell Tech
zl558@cornell.edu

Eric Jui-Chun Chien
Cornell Tech
jc3256@cornell.edu

Serge Belongie
Cornell Tech
sjb344@cornell.edu

Introduction
Augmented Reality (AR) is a novel technology that will revolutionize how we work, learn, and play. AR utilizes computer vision, computer graphics, spatial and tangible interaction to augment our perception and understanding of the environment. As academia and industry are improving computer vision and computer graphics methods to facilitate better AR use, current display technologies, such as mobile and projector-camera (Pro-Cam) hinder its widespread adoption and usefulness due to considerable usability challenges. We propose a novel multi-modal AR system that combines both mobile and pro-cam display technologies. By combining mobile and pro-cam systems will not only remove each displays limitations, but complement each others usability strengths to provide longer, higher-fidelity AR usage. We apply this contribution in the form of a smart kitchen application, that allows users to be provided instant cooking recommendations and intuitive instructions to prepare dishes. We evaluate our system on several participants, and discuss the potential this system brings for widespread adoption of AR.

Literature Review and Background
We began our project reviewing general usability challenges with general AR usage, and literature to identify work that surveyed various AR display technologies and their usability issues. For issues regarding specific AR display techniques,
we ended utilizing the work from Bimber and Raskar [1] that surveyed several AR display techniques, details about hardware and software implementations, and areas of application.

Milenski identify some challenges and considerations when designing general AR applications and experiences. Challenges include lack of cues to validate illusion and immersion, no standard for effective input or interaction with AR to keep consistent immersion, the issue of adding cognitive load will break immersion, and little or no control of environment.[2]

To identify specific challenges, different display methods and technologies dictate the array of challenges. This article restricts challenges of AR applications to mobile based AR and projector-illuminated AR display techniques. Bimber and Raskar defined mobile AR as a video-see through display method to project virtual 3D objects and information in the environment. This technology combines processor, memory, display, and interaction technology in a single device, and supports wireless and unconstrained mobile handling. Software used to create AR experiences utilizes computer vision and sensor fusion methods to understand the device’s orientation and camera intrinsics to display 3D data aligned with the physical environment.[1]

Pro-cam AR is an AR display technique that applies projection to project images directly onto surfaces within the viewer’s field of view. Pro-Cam systems utilize a camera for image processing and calibration and a projector for augmenting virtual objects and information. Pro-cam systems projects visual content either on flat surface, large walls, or on physical objects. For each AR display technique, Bimber and Raskar’s listed strengths and challenges with each technique that encompassed size of display, fidelity of content, mobility, and length of use.

Mobile AR has the advantage of facilitating experiences both indoor and outdoor, and provide location context[1]. From discussion with our advisors, another advantage is mobile AR facilitates privacy with personal screen to view augmented content. Some challenges with mobile AR experience are discomfort with time of holding device, small field of view, display view is offset from users field of view, and limited time of use with AR on mobile requires intense intensive processing usage. These issues cause significant affect on immersion.

Projector based AR has the advantage of large and unlimited field of view, augments on physical space rather than field of view provides strong immersion, augmentation on surface allows better collaboration with other users, and correct registration and illumination can cause manipulation of perception of object and objects appearance. Several challenges for pro-cam display technique include system is fixed in physical space, projection is view-independent, effective augmentation is dependent on surfaces diffuse properties, and unwanted shadows can occur. All challenges affect immersion.

Synthesizing the usability issues identified from Bimber and Raskar[1], from Miesnieks[2], and discussions from our advisors, we composed a chart in Fig. 1 that depicts the usability advantages and disadvantages of Mobile AR and pro-cam AR display technique.

In the figure, the x denotes the display technique does not have this usability characteristic. The check mark intends to show that the display technique can embody the quality. For other cell values in the table, we denoted a description to offer additional information regarding the usability characteristic of the technique.

We see that opportunities to compliment disadvantages of
one technology with the strength of the other. We bolded
the rows of complementary capabilities, which include mo-

bility, field-of-view (FOV), and engagement period. This
makes sense as mobile AR cannot facilitate long engage-
ment as fatigue in arm will occur, where as pro-cam can
facilitate long time of use as it is hands free. Mobile AR
is limited of displaying content in a small FOV due to its
display, where pro-cam has large FOV to display content as
it projects on large surfaces. Finally, proc-cam AR requires
recalibration of projection for effective use, so it is generally
constrained in a fixed setting. Mobile AR you are able to
move around and the content will adjust accordingly.

By combining both technologies through complementation,
we can facilitate small and large FOV AR, immersive and
mobile AR, AR that is single user and collaborative, and
both be usable for short usage and long usage.

Table 1: Advantages and Disadvantages of Pro-cam AR and
Mobile Rendering Techniques

<table>
<thead>
<tr>
<th>Feature</th>
<th>Pro-Cam</th>
<th>Mobile AI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hands-free</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Mobility</td>
<td>Available, food</td>
<td>Always with me, can move around</td>
</tr>
<tr>
<td>Privacy</td>
<td>Relative, public</td>
<td>Private, personal</td>
</tr>
<tr>
<td>Shared/Group activity</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>2D/High-definition content</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Field of view/Large-scale content</td>
<td>Large</td>
<td>Small</td>
</tr>
<tr>
<td>Engagement period/Duration of use</td>
<td>Expected to be long</td>
<td>Short</td>
</tr>
<tr>
<td>Potential level of manipulating AR content</td>
<td>Somewhat lower</td>
<td>Higher</td>
</tr>
<tr>
<td>Ease of interaction</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>Single task/Multitask</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Multi-user</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Interact with physical gestures</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>

By combining both technologies through complementation,
we can facilitate small and large FOV AR, immersive and
mobile AR, AR that is single user and collaborative, and
both be usable for short usage and long usage.

involves instant contextual understanding of your kitchen,
instructed cooking, and easy-sharing of high-resolution 3-
dimensional models of food dishes. As we produced a com-
plete user flow of the experience, we focus our implementa-
tion on the portions of the application where we can inter-
change AR modalities and validate its complementary abil-
ities. This can be seen in Table 3. Within the green area,
you can see the steps scanning ingredients and previewing
3D model of dish complimenting AR systems are leverage
when you interchange from Mobile AR to Pro-Cam.

We conducted extensive usability research on Mobile AR
and Pro-Cam Systems, mapped out the strengths, chal-

lenges, and areas where they could be best applied. Among
our many ideas, Smart Kitchen best utilizes the strengths
of both types of AR, and introduces a novel usage with the
technology.

User Flow of Kitchen Application
Our user flow contains five sections: scanning ingredients
and portions, previewing 3D models of dishes, instructed
cooking, constructing a 3D model of the finished dish, shar-
ning with friends. The last step, sharing with friends, would
potentially loop back to the second step, previewing 3D
models of dishes, since a user could decide what to cook
based on the dishes shared by her or his friends. We want to
validate if our application, utilizing the strengths of both
AR technologies, could incentivize people to try cooking
different dishes through previewing dishes in Mobile AR,
effectively guide people through cooking processes unfa-

miliar to them through understanding the environment and
overlaying instructions with a Pro-Cam System.
Pilot Study
We invited five current Cornell Tech students to test out our application prototype, followed by a qualitative interview to understand each of their habits when it comes to preparing meals.

To help the interviewer walk through the flow smoothly and give more grounded experience for the interviewers, we implement a simple iOS application that runs a food recognition model, provide interface for simple usage guidance and give users a high fidelity 3D view of a recommended dish. Additional to that, we make a simple interactive interface using a pro-cam system using papART, a processing framework for pro-cam AR software. This interface could be projected onto any wide and flat surface. There are area for image, area for video, timer, step by step instructions and some buttons like DONE and SET TIMER. Last but not least, we build up a prototype system. The system contains a projector, a web camera to capture color image, a depth camera to capture depth information, and a lamp frame. In our design, the web camera can take color images that server as inputs to our computer vision model like object recognition, the depth camera enable the system to react to users gesture actions like tapping and swiping. All the projector, two cameras are taped onto the lamp frame and the projector and cameras are faced downward.

Table 2: User flow of smart kitchen application

Table 3: Visualization of complimentary AR with relation to user flow. Green area is where mobile AR compliments procam AR.
The ideal user flow contains five steps: scan ingredients and portions, preview 3D model of dish, instructed cooking, construct 3D model of finished dish and share with friends. However, given the limited time frame, we only simulate the first 3 steps of the flow, leaving the construction of finished dish 3D model and the social sharing function to the users imagination.

During the pilot study, we asked the interviewers to really pick up the phone and follow the instruction of the mobile AR app and then switch to the pro-cam system interface, we managed to let the users walk through the flow without too much intervention from us. After the whole process, the interviewer was asked a few questions and fill out a questionnaire, in some case, we also have a little chat with the interviewer about the feeling of our system.
User Feedback

The results of the qualitative interviews showed that most of our interviewees are willing to cook their meals rather than purchase at a restaurant. We learned from our users that the cooking frequency is evenly distributed from 2 times a week to everyday. The reasons for cooking often are, 1) There are not a lot of restaurants around the living neighborhood, 2) The users are more used to their own cooking style, 3) It is much cheaper if cooking at home. The reasons for not cooking are, 1) The users have no time for cook, 2) The users are too lazy for cooking, 3) The users love cooking but do not want to prepare food ingredients and wash dishes after cooking.

The interviews show the biggest difficulties the users might run into when cooking are the following: 1) When users follow the recipe to make a dish, sometimes there are too many steps and users have to check the laptop a few times, 2) It is annoying when users have to touch the phone with ingredients on their hands, 3) Some instructions on the recipes are vague, they would often watch videos in order to understand better, 4) Users do not have many new ideas, if they find a new dish they would cook for the whole week or until they get tired of it, 5) They only know some dishes, sometimes they are short of the ingredients or they do not know what to cook with existing ingredients.

The exciting news is that 80% of our interviewers think Smart Kitchen could be time saving and useful for them. With the recommendation system based on the ingredients or users preference, Smart Kitchen encourage users to try new dishes with the same ingredients or even new dishes with totally different ingredients. Also, users pointed out the instructions during the pilot study are not clear enough, however, they can still see how such instructions make cooking easier and with more fun.
Voice Control Feature

Table 6: Screenshot of pro-cam AR. This system contains projection of interactive content, voice conversational interface, and paper tracking for prioritized display.

After doing the pilot study and analyzing the users feedbacks, we felt the need to augment the current system with voice control system. Voice control system has been integrated into smart home design long ago and have seen strong adoption rates in the last few years. With the boost of performance of several main voice control systems like Alexa, Siri, Cortana and Google Assistant, voice control smart home is getting more and more useful and versatile. Our project can also serve as a pioneer in exploring how voice control can help in a kitchen.

Voice control system is really a good fit into the SmartKitchen scenario. Our current system makes use of a projector AR system to allow people to free their hands. So that people can save the time washing their hand repeatedly to control the electronic devices and enjoy the continuity of cooking. However, most of the interaction between our system and users are through the projected surface. We display information onto the surface and users take actions like tapping and swiping on the surface, most importantly, users have to receive the information via reading and watching, which could be a huge distraction when the oven is running, the water is boiling and the user is cutting something, too many things are happening at the same time. Voice control system comes to rescue. It is more direct and natural for people to receive voice information. For reading and watching, people are actively doing such actions, but for listening, people are passively receiving information.

After discussion, we decide to perform our voice function using Amazon Alexa. Because the Amazon eco-system allows developer to easily customize skills that Alexa can perform. We build up a simple Alexa skill called SmartKitchen that should able to communicate with our mobile system and the projector system. After user selecting the dish on the mobile end, the selection will be sent to the Alexa backend, so when users ask Alexa to invoke the SmartKitchen skill, Alexa has already known what the user want to cook and begins to tell the user what to do next. Users can obtain a seamless experience switching from mobile to Alexa. From the users’ star point, the mobile system, the projector-camera system and the voice system are merged and mixed together.

As for the detail implementation, there are several choices to host an Alexa backend. Because finishing the whole Alexa skill is far beyond the range of the work in a semester, we did not choose to build up our own sever to host the Alexa backend. Instead, we use the serverless service from Amazon, AWS Lambda. After the skill is intrigued from the Alexa using some keywords, the intent will be transformed into JSON data stream and sent to a python script hosted on AWS Lambda. The backend python script will
Paper Tracking Technique

Voice control system is one inspiration from the user feedback, another improvement to the system is a Smart Cutting Board. The Smart Cutting Board would be a dedicated area for the Pro-Cam system to project instructions or illustrations on. It is an important hardware element which can help Smart Kitchen accommodate users with different kitchen environments. In our previous design, we assume the users would have an adequate open space to prepare ingredients, and we use this area to be our projecting surface. However, at least from the user feedback, this assumption is broken. For the whole system to handle various kitchen environments, we can make a specific area that is capable to host the projection instead of projecting the information on wherever. However, this dedicated area must be irrelevant to the environment and should not be a fixed point. These constrains lead us to a cutting board. Cutting board is detached from the environment and it is moveable. Plus, a cutting board is usually a flat plain surface, which is perfect for projections.

Once the idea is decided, the remaining task is to make it happen. Tracking the cutting board in real time requires the supports both from hardware and software. The camera in the Pro-Cam system was used for detecting users’ actions before, but we only use its deep camera function. Here we also make use of its RGB camera function, serve as an eye of the system to keep track of the cutting board. We don’t have time to implement computer vision methods like objects recognition and moving objects tracking technique, but there is always a work around. We pasted markers on the cutting board and the mechanism behind the tracking is no different than QR code. With the support of hardware, the backend of our system now knows where the content should be projected. However, because the cutting board is moving, the workload of real time calibration is fully on the software side. The calibration problem is not completely solved yet, however, our system is still able to perform good enough to make a demo and receive another round of pilot study.

Validation with User Research

After the implementation of Voice Control and paper tracking features is complete, we performed a second-round of user research in order to validate the advancement of our guided cooking function. The user research consists of a user test with the new system, followed by a list of qualitative interview questions that investigate whether the added features contribute to a smoother work flow, and create a better way to present voice instructions and visual content to support the cooking process. The interviewees of this user research include people who were part of the first pilot study, as well as new users who have never tried out Smart Kitchen in the past.

For the user test, the interviewees are presented with some food ingredients and a predetermined dish that they will cook, which in this case are apples and apple pie. The interviewees start by launching the Smart Kitchen mobile app on a phone, and uses the camera to recognize and record the ingredients they have. Next, the users are presented with a few different suggested recipes of dishes they can cook with the existing ingredients. Under our instruction to opt for the apple pie recipe, the users can then preview a 3D model of the finished apple pie in AR. By tapping on the phone to place and fixate the model on any surface the phone detects, the users can then walk around the model to take a closer look, or view from a different angle. After the users tap the "Start Cooking" button on the phone,
the process is handed over to the Pro-Cam system and the Alexa voice control system.

The users then initiate the guided cooking process by saying "Alexa, start cooking with Smart Kitchen." This voice prompt is Alexa’s cue to start walking the users through with voice instructions. Alexa begins by telling the users the portion they need for each ingredient, and it can also repeat the instructions when needed, so that the users don’t need to refer to the recipe step by step. In the rest of the cooking process, Alexa and the Pro-Cam system work together to provide the users visual and voice instructions and content necessary to complete each task. For example, when asking the users to dice an apple, Alexa would provide detailed descriptions to execute, while the Pro-Cam system would project visual illustrations on the Smart Cutting Board at the same time. After the users put the apple pie into the oven to bake, the system would continue to track the time elapsed, and remind the users to take out the pie when it is fully baked. The same kind of mechanism applies to other tasks in the cooking process.

The feedback from the interviewees were positive, and all of them believed the addition of voice control and instructions will benefit the system greatly. According to the interviewees, the main reasons they favor the feature is because: 1) voice control interface provides a way to communicate with and construct an input or request to the system, 2) talking is faster, more intuitive and convenient than tapping on visual content projected on surfaces since the users’ hands are usually busy in the cooking process 3) voice control prevents the kitchen from being stained by touches, as the users’ hands could be greasy or holding ingredients, 4) voice control is more hygienic as it minimizes the contact of the users’ hands and other objects, but the downside is the potential transmission of bacteria through saliva.

The responses regarding the Smart Cutting Board were also affirmative. Interviewees who cook frequently responded that the Smart Cutting Board is a great dedicated medium for viewing videos, display timer, recipes, cooking tips, illustrations for cutting. They also suggested that the Smart Cutting Board can display the sought-after texture for ingredients, or work as an electronic scale to weigh ingredients.

Discussion and Future Plans
We have a few suggestions for researchers who would like to continue on this research in the future. First of all, for AR to work in a kitchen setting, better hardware is needed. For example, the projector in the Pro-Cam system needs to have higher luminosity, and preferably shorter throw projection. One alternative is to use ultra-short throw projectors. Secondly, communication between each separate device in the system can be improved. For now, we have limited network communication from the mobile phone to the Pro-Cam system and Alexa when the user begins the guided cooking process. In order to achieve a seamless user flow, the hand-off should be robustly implemented. Thirdly, we need more advanced computer vision technology to help the Pro-Cam system understand the kitchen environment, to be able to provide more assistance to the user. For example: precisely recognize the position of different ingredients on the kitchen counter, or even a pot of boiling water. Additionally, more extensive user research should be conducted before developing further features of the Smart Kitchen. Our user research was not comprehensive enough to reflect the needs of the mass population, therefore we encourage future researchers to include more interviewees and conduct contextual inquiries to really understand their pain points.
Acknowledgements
We would like to thank Benedetta Piantella and Harald Harrison for their wonderful support for the project. Benedetta and Harald provided valuable insights and feedback that greatly improved the project.

References and Citations