Recipe2Video: Synthesizing Personalized Videos from Recipe Texts

Supplementary Material

Prateksha Udhayanan\textsuperscript{1}, Suryateja BV\textsuperscript{2}, Parth Laturia\textsuperscript{3}\textsuperscript{*}, Dev Chauhan\textsuperscript{4}, Darshan Khandelwal\textsuperscript{5}, Stefano Petrangeli\textsuperscript{1}, and Balaji Vasan Srinivasan\textsuperscript{1}

\textsuperscript{1}Adobe Research; \textsuperscript{2}Avanti Fellows; \textsuperscript{3}Morgan Stanley; \textsuperscript{4}Graviton Research Capital LLP; \textsuperscript{5}Goldman Sachs

udhayana@adobe.com, suryateja@avantifellows.org, dev.chauhan@gravitontrading.com
parthlaturia@gmail.com, darshankhandelwal1218@gmail.com, \{petrange,balsrini\}@adobe.com

1. Introduction

The supplementary material is organized as follows:

1. Example-based explanation of different components of the Recipe2Video architecture
2. Qualitative examples
3. Details on the human evaluation process
4. Demo video

2. Recipe2Video: System Architecture

To build our Recipe2Video system, we adopt the framework of diagrammatic mode from modern multimodal theory \cite{2}, which offers a top-down approach for designing effective multimedia content: (i) decide a communicative goal of the content, also called discourse semantics; (ii) find expressive resources to meet the chosen communicative goals. Food recipe texts are one of the few domains that offer data \cite{3,1} connecting multimedia and communicative goals. We therefore convert recipe texts to videos with the goal of enhancing consumption experiences leveraging these data \cite{3,1}.

2.1. Ranking assets and their combinations

Figure 1 shows an example of using Information Coverage for ranking two assets. These assets are retrieved for the instruction: In a large bowl, whisk together the sugar, flour, cocoa powder, baking powder, baking soda, and salt.

As mentioned before, we score the temporal assets based on 3 questions: (1) Does the image show how to prepare before carrying out the instruction? (2) Does the image show results of the action described in the instruction? (3) Does the image depict an action in progress described in the instruction? The characterization of the temporal aspects into three categories allows us to synthesize video according to specific user preferences. For example, a consumer looking for a succinct summary of the actions might be better served by optimizing the assets for the third question above, while someone who is preparing for a procedure can be better served by optimizing for the first question above.

Figure 2 shows a single asset (a) and a 3-asset (b) example that are likely to get chosen for different video variants.

3. Qualitative Examples

Figure 3 shows a set of frames extracted from the elaborate video variant synthesized by Recipe2Video system for a chocolate cake recipe. Precise illustrations of canola
Figure 2. Part (a) demonstrates a single asset that illustrates the instruction: Top with another layer of ravioli and the remaining sauce not all the ravioli may be needed. Sprinkle with the Parmesan. The image depicts both the action and the result of the action. An asset like this is a potential candidate of being chosen for the succinct variant. Part (b) demonstrates a 3-asset combination chosen for instruction: In a large bowl, whisk together the eggs, water, milk, oil, and vanilla extract. This combination is very likely to be chosen for the elaborate variant.

Figure 3. Frames extracted from the elaborate video variant synthesized by our system. Both frames (a, b) illustrate instructions which could potentially help a novice user understand the prerequisites of making Chocolate Cake recipe.

Figure 4. Two frames corresponding to the same instruction but from different variants synthesized by our system. Frame (c) is extracted from the elaborate variant; Frame (d) is extracted from the succinct variant.

Figure 5. Comparison of frames extracted from videos synthesized by our Recipe2Video method and Audiovisual Slideshows method. Frames (a-d) correspond to our system; Frames (e-h) correspond to Audiovisual Slideshows baseline. Each column represents the same instruction but the assets in our video are more informative, coherent, and thereby enhancing consumer experience as opposed to the baseline video.

4. Human Evaluation

For all MTurk surveys, we set the annotator prerequisites as "MTurk Masters" located in the United States having an
How many instructions/steps are present in this recipe?
- 4 or 5
- 6 or 7
- 8 or 9
- 10 or 11

What could be an appropriate name/title for the recipe shown above?
Respondent chooses one of the four options which contain one correct answer along with three randomly sampled names.

Which of the following ingredients might have been used in the recipe?
Respondent chooses one of the four options which contain one correct answer along with three randomly sampled names.

How enjoyable or boring did you find going through the recipe?
- Very boring
- A little boring
- A little enjoyable
- Very enjoyable

How much of the recipe can you remember now without looking at it again?
- Cannot remember anything
- Can remember very little
- Can remember some of it
- Can remember most

How often did you find yourself going back to previous steps in the text instructions list to understand the recipe?
- Never
- Once
- Sometimes
- Almost after every step

Select the incoherent image in the following set of images. An image is incoherent if it is not related to the recipe.
- Image-1
- Image-2
- Image-3
- Image-4

Were the images/clips within the frames related to the text shown in the frame?
- Not at all
- A little
- Somewhat
- A lot

Do you think the steps displayed in the video along with images and clips were sensible and followed the right order?
- Made no sense at all
- Mostly didn’t make any sense
- Somewhat made sense
- Made perfect sense

Were the images/clips in the video related to the recipe text?
- Not at all
- A little
- Somewhat
- A lot

Suppose you made a mistake while preparing this recipe in your kitchen. What would you opt to view again to correct yourself?
- Video
- Text
- None

Please provide a feedback on this survey.

Table 1. Here are a few sample questions that were presented to the annotators for human evaluations.

<table>
<thead>
<tr>
<th>Question</th>
<th>Category</th>
<th>Experiment(s)</th>
<th>Metrics</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>How many instructions/steps are present in this recipe?</td>
<td>Sanity Check</td>
<td>Task; 1, 2, 3</td>
<td>none</td>
<td>4 or 5, 6 or 7, 8 or 9, 10 or 11</td>
</tr>
<tr>
<td>What could be an appropriate name/title for the recipe shown above?</td>
<td>Sanity Check</td>
<td>Task; 1, 2, 3</td>
<td>three randomly sampled names</td>
<td>one correct answer with three options</td>
</tr>
<tr>
<td>Which of the following ingredients might have been used in the recipe?</td>
<td>Sanity Check</td>
<td>Task; 1, 2, 3</td>
<td>three randomly sampled names</td>
<td>one correct answer with three options</td>
</tr>
<tr>
<td>How enjoyable or boring did you find going through the recipe?</td>
<td>Enjoyable</td>
<td>1, 2</td>
<td>none</td>
<td>Very boring, A little boring, A little enjoyable, Very enjoyable</td>
</tr>
<tr>
<td>How much of the recipe can you remember now without looking at it again?</td>
<td>Retainable</td>
<td>1, 2</td>
<td>none</td>
<td>Cannot remember anything, Can remember very little, Can remember some of it, Can remember most</td>
</tr>
<tr>
<td>How often did you find yourself going back to previous steps in the text instructions list to understand the recipe?</td>
<td>Jarring</td>
<td>1, 2</td>
<td>none</td>
<td>Never, Once, Sometimes, Almost after every step</td>
</tr>
<tr>
<td>Select the incoherent image in the following set of images. An image is incoherent if it is not related to the recipe.</td>
<td>Task; 1, 2</td>
<td></td>
<td>four images</td>
<td>Image-1, Image-2, Image-3, Image-4</td>
</tr>
<tr>
<td>Were the images/clips within the frames related to the text shown in the frame?</td>
<td>Intra-Coherence</td>
<td>2</td>
<td>none</td>
<td>Not at all, A little, Somewhat, A lot</td>
</tr>
<tr>
<td>Do you think the steps displayed in the video along with images and clips were sensible and followed the right order?</td>
<td>Inter-Coherence</td>
<td>2</td>
<td>none</td>
<td>Made no sense at all, Mostly didn’t make any sense, Somewhat made sense, Made perfect sense</td>
</tr>
<tr>
<td>Were the images/clips in the video related to the recipe text?</td>
<td>Relevance</td>
<td>3</td>
<td>none</td>
<td>Not at all, A little, Somewhat, A lot</td>
</tr>
<tr>
<td>Suppose you made a mistake while preparing this recipe in your kitchen. What would you opt to view again to correct yourself?</td>
<td>Self-correction</td>
<td>3</td>
<td>none</td>
<td>Video, Text, None</td>
</tr>
</tbody>
</table>

Figure 6. Average preference of respondents on a Likert Scale, when they were shown recipe texts in Expt (1) and asked if there were actions where they might have preferred visuals. Note that the total number of responses is $N \times 5 = 75$, since we evaluated on $N = 15$ RecipeQA texts.

Table 1 lists a set of example questions that were presented to the annotators for human evaluation. We also indicate the metrics and experiments each question correspond to. For instance, [Enjoyable; 1, 2] means that this question is used in experiments (1) and (2) to gauge the enjoyability of the displayed modality.

Figure 6 shows the average number of respondents who prefer to consume actions via visuals. This further verifies our hypothesis that the consumption of a procedural text is better carried out via visuals than text.

As described before, in Expt (3), we ask the respondents to opt for their preferred modality (Video/Text/None) of consumption under different scenarios. Figure 7 shows the aggregated responses across both RecipeQA (23 recipes) and Tasty Videos (25 recipes) datasets for different scenarios. To capture the respondent bias of opting for "Video" regardless of the situation, we calculate the average across all scenarios and variants for a particular system and display it as a horizontal line (125.2 for Recipe2Video, 55.7 for Baseline). Respondents generally opt for our elaborate
Figure 7. The bar graphs show the number of respondents who chose “Video” for each of the questions listed in the x-axis, aggregated over both datasets. Note that each bar here corresponds to 

\[(23 + 25) \times 5 = 240\] responses, since we solicit five responses for each (text, video) task pair in Expt (3).

video variant over the procedural text when they would need to share it with a novice who is new to cooking. Similarly, respondents opt for our succinct variant to use as a reference for the future. Also, there is a significant gain (+24.8) in opting for the elaborate variant (150) over the average (125.2) of opting for videos of Recipe2Video system. Similarly, succinct variants offer a gain of 19.8 over the average. These results show that the synthesized variants meet the intended communicative goals.

5. Demo Video

A working demo video of our end-to-end system Recipe2Video can be found in the following link -
https://bit.ly/3s11wp3 The video file is titled Recipe2Video_Demo.mp4. The video contains a demonstration of the working of the system - it takes an input document and generates 2 variants of videos catering to different user preferences. The demo also explains the working of the ranking modules with several examples.

References

