[Supplementary Material] Straight to the Point: Fast-forwarding Videos via Reinforcement Learning Using Textual Data

Washington Ramos[1] Michel Silva[1] Edson Araujo[1] Leandro Soriano Marcolino[2]
Erickson Nascimento[1]
[1]Universidade Federal de Minas Gerais (UFMG), Brazil [2]Lancaster University, UK
[1]{washington.ramos, michelms, edsonroteia, erickson}@dcc.ufmg.br, [2]l.marcolino@lancaster.ac.uk

In this supplementary material, we present several qualitative results on recipe videos of the YouCook2 dataset in addition to the ones presented in Figure 4 and Figure 5 of the main paper. In the results, we show the coverage of the selected frames regarding our method and the competitors, *i.e.*, SSFF [2] and FFNet [1]. For each result, we also report the annotated segments.

The results presented in Figure 1 show that the deceleration profile of our method matches the annotated segments of the video. Note that several of the emphasized regions are not perfectly aligned to the ground-truth data. However, it is worth mentioning that there is an emphasized region for each annotated segment of the recipe. The competitors, on the other hand, could not provide the same result.

Even though the FFNet method had been trained in this domain (*i.e.*, using the annotation provided by the dataset), it was capable of only emphasizing a small portion of the annotated segment starting around the frame 3,000. Regarding the SSFF coverage, it is noteworthy the amount and magnitude of temporal gaps in the frame selection. As pointed before, temporal gaps usually lead to visual discontinuity in the final video. One example is the annotated segment starting around the frame 6,000, the frame selection of our method emphasized three large segments, while the SSFF skipped almost the entire segment.

Figure 2 depicts the coverage of the selected frames composing the accelerated video generated by our methodology and the two competitors regarding the annotated segments. We can observe that between frames 4,000 and 4,400, both SSFF and our method decelerate. In the video used in this experiment, the segment portrays the final dish. Since the visual content of the final dish comprises much of the visual content from the recipe's ingredients, it is expected that our method might interpret such frames as relevant – the SSFF method assigned relevance due to the presence of kitchen-related objects. The SSFF method results show more accentuated gaps between frames, resulting in a visual discontinuity in the final video. It is noteworthy that even though FFNet had been trained using the ground truth annotations from the training dataset, the method failed to emphasize frames that contain the recipe instructions.

In Figure 3, when analyzing the first ground truth segment (GT) of the video, we can visualize that our method was able to emphasize frames in both ends of the segment, but not in the segment itself. By analyzing this annotated portion in the original video (see Figure 3-bottom), we see that the frames where our method decelerates, *i.e.*, those surrounding the ground truth segment, depict the actual step of the recipe. On the other hand, the frames included in the ground truth segment itself do not visually represent the instruction, since they are mainly composed of the executor in close-up. We argue that the densely sampled region of the SSFF method can be explained by the frames inside the GT segment having some visual clues that may lead the YOLO extractor to associate them to the kitchen environment, *e.g.*, microwave, bowl, sink, spoon, *etc*.

The third and fourth ground truth segments of the video are annotated as spaced actions, as showed in Figure 3; however, in the original video, they are presented as consecutive frames in a non-cut video clip. The results show that our method emphasizes the entire segment. The SSFF also performed a dense sampling in these segments; however, it is due to the presence of the oven and the object misclassification of the pan as a bowl. It is worth noting that the SSFF method only analyzes the presence of the relevant objects in the scene. In contrast, our method relates visual information of the frames and text from the input instructions document.

References

[1] S. Lan, R. Panda, Q. Zhu, and A. K. Roy-Chowdhury. FFNet: Video fast-forwarding via reinforcement learning. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 6771–6780, June 2018. 1

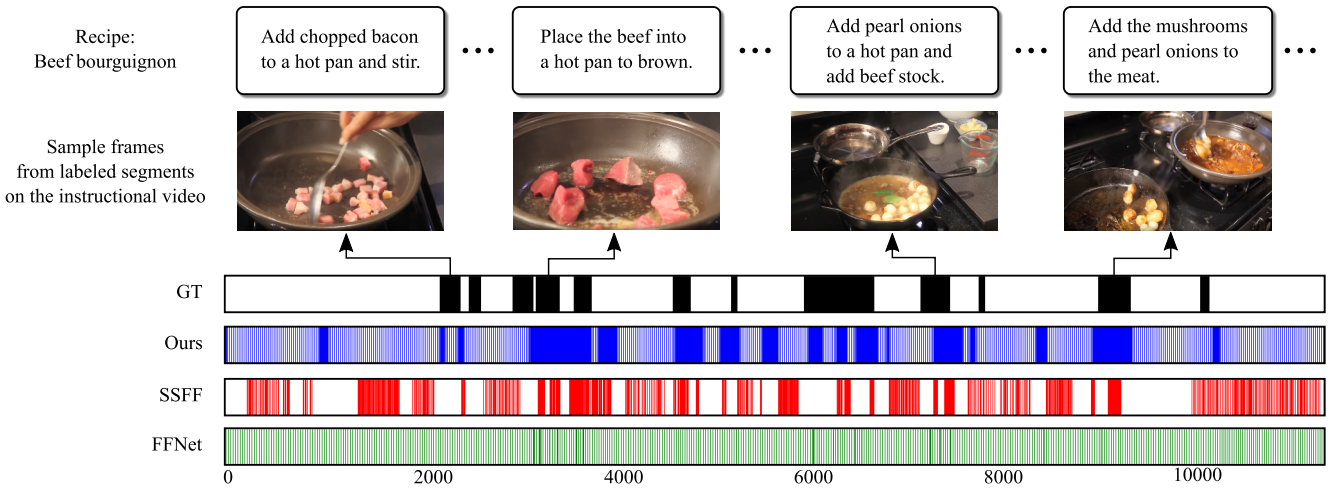

Figure 1. Qualitative results for the compared methodologies in the video for the recipe "Beef Bourguignon" from the YouCook2 dataset. The vertical bars inside the rectangles indicate the selected frames for each method. GT stands to ground truth, and the contiguous black blocks indicate the annotated video segment. The competitors used in our experiments are SSFF and FFNet. In general, our selected frames match most frames from ground truth data.

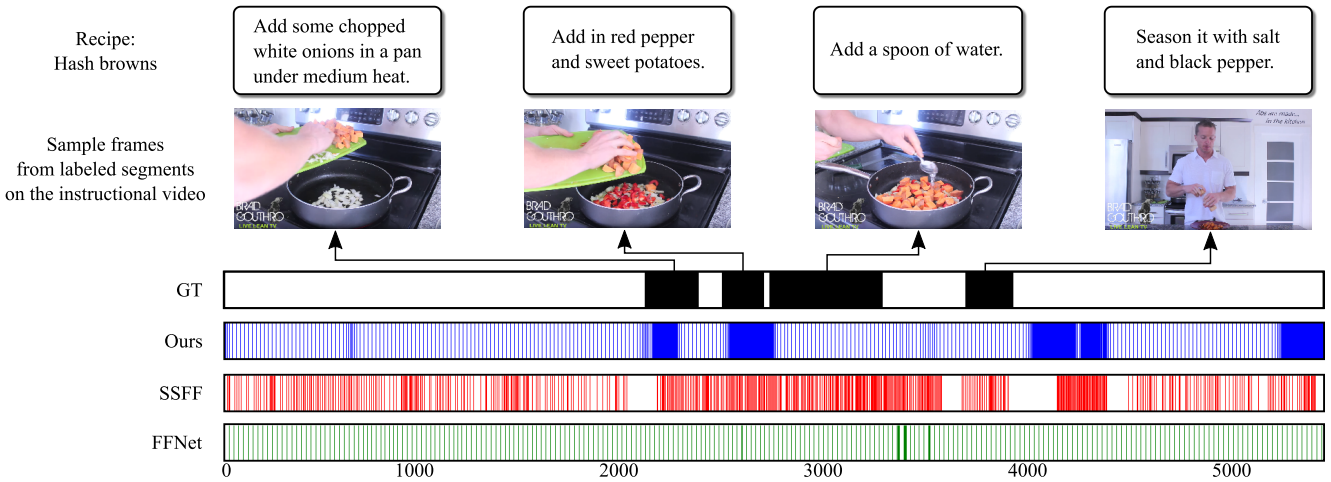

Figure 2. Qualitative results for the compared methodologies in the video for the recipe "Hash Browns" from the YouCook2 dataset. The vertical bars inside the rectangles indicate the selected frames for each method. GT stands to ground truth, and the contiguous black blocks indicate the annotated video segment. The competitors, SSFF and FFNet, present a poor frame selection in terms of GT coverage in comparison to ours.

[2] M. Silva, W. Ramos, J. Ferreira, F. Chamone, M. Campos, and E. R. Nascimento. A weighted sparse sampling and smoothing frame transition approach for semantic fast-forward first-person videos. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, pages 2383–2392, June 2018. 1

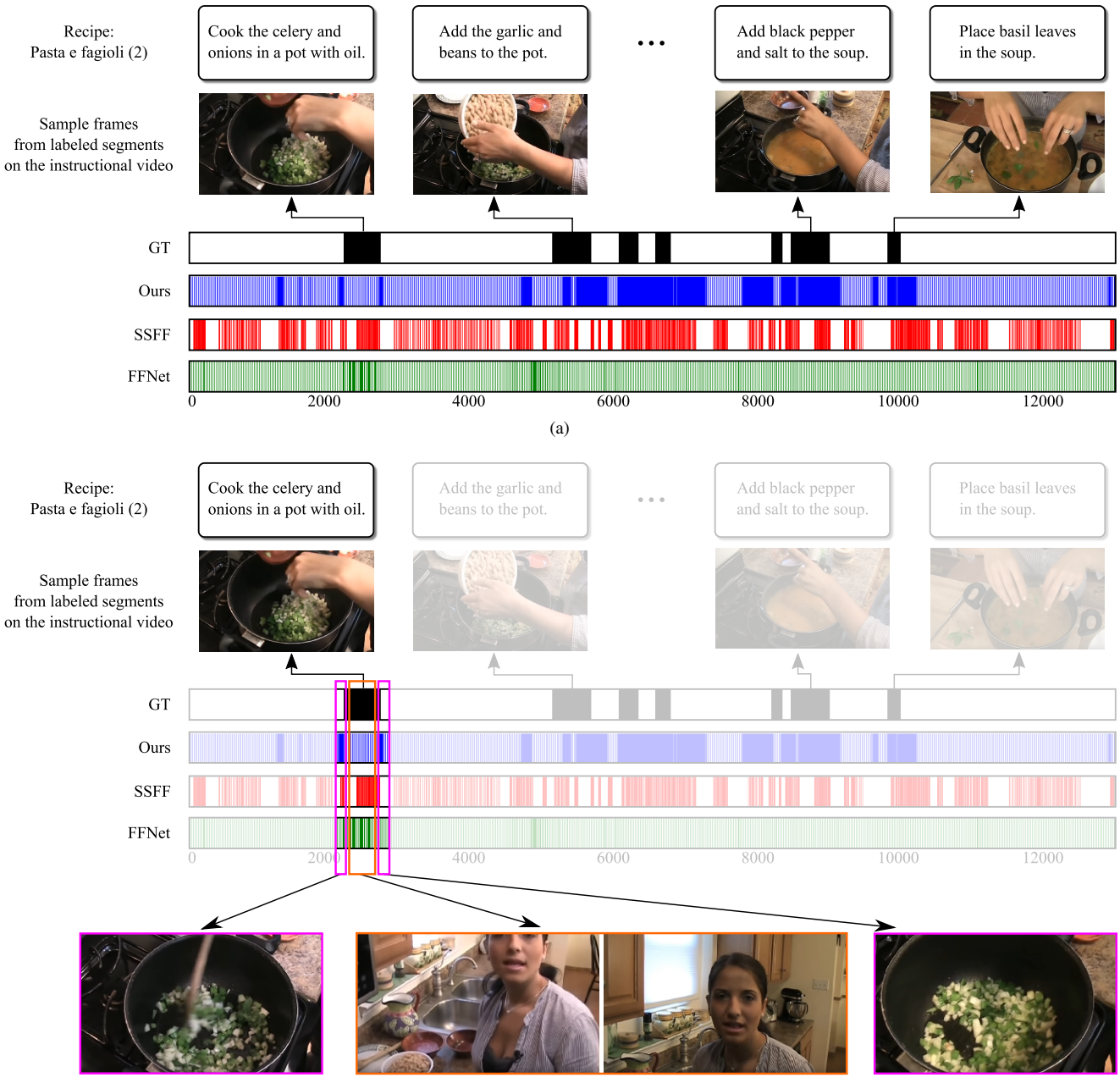

Figure 3. Qualitative results for the compared methodologies in one of the videos for the recipe "Pasta e Fagioli" from the YouCook2 dataset. The vertical bars inside the rectangles indicate the selected frames for each method. GT stands to ground truth, and the contiguous black blocks indicate the annotated video segment. The competitors used in our experiments are SSFF and FFNet. Analyzing the SSFF frame sampling, we notice the temporal gaps resulted from the adaptive frame sampling performed by this technique. The highlighted region in (b) shows that, in the first annotated segment, only the very beginning and the final of this segment shows the recipe. The majority of the frames shows a close-up of the instructor narrating details and curiosities about the original Italian recipe.