Cooking Video Summarization Guided By Matching with Step-By-Step Recipe Photos

Ryo Sobue
Chubu University
ryos@mprg.cs.chubu.ac.jp

Mitsuru Nakazawa, Yeongnam Chae, Björn Stenger
Rakuten Institute of Technology
{mitsuru.nakazawa, yeongnam.chae, bjorn.stenger}@rakuten.com

Takayoshi Yamashita, Hironobu Fujiyoshi
Chubu University
{takayoshi, fujiyoshi}@isc.chubu.ac.jp

Abstract

Cooking videos shared on social media enable users to easily understand the cooking process. However, it is difficult to create such summarization videos, as it requires video editing expertise to create high-quality content. In this paper we propose a semi-automatic system for summarizing cooking videos from a full video of the cooking process and a number of step-by-step recipe images. The system is designed to allow anyone to create shareable cooking videos. We focus on online recipe websites because their recipe images provide good descriptions of the food preparation procedures. Our system first searches for scenes similar to each recipe image using features of an Inception Net, fine-tuned to recipe images. These matches are used as key frame candidates, from which users can then select their favorite scenes. Finally, our system automatically creates a summarized video using editing techniques such as speeding up. User studies showed high satisfaction rates of both editors and viewers, demonstrating the benefits of the proposed system.

1 Introduction

A recent trend on recipe sharing websites is to provide videos summarizing the meal preparation process [1]. The videos show scenes of the cooking process in the correct temporal order, and are edited for increasing their attractiveness on social media. Editing techniques that are commonly used include showing scenes of the finished dishes, and speeding up some of the cooking scenes. Creating such summarization videos from a full video therefore takes time and know-how. Some companies already provide support for creating cooking videos, for example *Cookpad Studio* [2] provides a services to assist the capturing and editing of cooking videos in a fully equipped studio, with advice from professional staff. Similarly, *MoveAsia* provides a paid service for capturing and editing of cooking videos suitable for sharing on social media [3]. However, the costs of such professional services can be a hurdle for adoption by casual users.

Figure 1. **Overview.** *(1) The user selects a recipe, (2) captures the meal preparation process, (3) creates a summary video using the proposed system, and (4) posts the summarized video on a social network service (SNS).*

We propose an interactive video editing system to summarize the cooking process, thereby empowering users with little or no video editing experience. The steps to creating such a video are shown in Figure 1. The user first selects a recipe and records the meal preparation process. The key idea of our summarization system is to use the step-by-step recipe images from the website to guide the video shot selection. We find the top matches in the video and optionally allow the user to select their favorite shot [4]. Finally, these shots of representative scenes are merged by video using effects recommended by professional video editors.

The contributions of this paper are: (1) The design of an interactive system for generating a cooking summarization video from a long video and step-by-step photos as input, and (2) similarity image search using Inception V3 model features, fine-tuned to images of

our target domain. (3) We evaluate the system with user studies.

2 Related Work

This section highlights some of the prior work on video summarization. Smith and Kanade [5] proposed using a combination of visual features such as motion information and speech recognition results to generate a shorter video. Otani *et al.* proposed a clustering approach by projecting segments of a source video into a semantic visual feature space [6]. Zhang *et al.* took a learning-based approach with summarized videos as training data [7]. In an interactive approach, Ana *et al.* proposed an method that continuously prompts users for feedback and updates the summarization result on-the-fly [4].

Some approaches target specific video genres and use their video characteristics for better summarization. For example, for first-person video summarization, Molino *et al.* proposed using gaze information predicted from the source video [8]. For football video summarization, Zhoa *et al.* selected goal scenes as highlights [9]. A number of approaches have been proposed for cooking video summarization. For example, Miura *et al.* detected shots showing repetitive motion, such as chopping food with a knife, using optical flow, and used these as key scenes [10]. Hayashi *et al.* detected redundant scenes within a cooking video using cubic higher-order auto-correlation (CHLAC) features and automatically removed them [11]. In contrast to these approaches, our proposed system makes use of the step-by-step recipe images, and creates videos with the aim of generating high-quality content for social media.

3 Proposed system

3.1 Design guidelines

To better understand what kind of summarized cooking videos are attractive to social network users, we consulted experts with experience of running a recipe sharing website. We gathered the following recommendations:

1. The total playtime should be about 30-60 seconds.
2. Images of the readily prepared food should be inserted both at the beginning and at the end of a video.
3. Short and simple subtitles should be added to enable SNS users to easily understand the ingredients and instructions.
4. Scenes of the cooking processes should be speeded up by a factor of two or more in order for users to quickly grasp the content.
5. The video should be captures from a top-down view.

3.2 Video creation overview

As shown in figure 1, steps for creating the video summary are (Step 1) the user first selects a cooking recipe from a recipe websites. (Step 2) the food preparation is captured continuously from a top-down view, (Step 3) the system searches key frames in the image that are similar to the recipe images on the website. The user can view clips containing matching key frames and select the favorite ones from each set of candidates. Optionally the user can modify the default subtitles taken from the recipe text. (Step 4) the system creates the summarization video using automatic editing.

3.3 Matching recipe images to video frames

Figure 2 shows the flow of processing to extract the feature vectors and measure their similarity. We employ a fine-tuned Inception V3 model for extracting feature vectors from the recipe images and the frames from the cooking video [12]. The Inception V3 model has achieved high recognition accuracy in object recognition tasks. It is possible to extract effective features from the *Mixed 5* layer of Inception V3 which was pretrained on the ImageNet dataset. A fully connected layer is added to the *Mixed 5* layer to output a 128-dimensional feature vector. However original Inception V3 model is for General object recognition, we need achieve high recognition accuracy in cooking scene recognition. thus we use a triplet loss function to measure the distance between two feature vectors as shown in Figure 3. The goal is to reduce the distance between a video frame (positive) and a corresponding recipe image (anchor). On the other hand, the distance between a frame different from the recipe (negative) and the anchor becomes larger. We denote the input feature vector of the anchor, positive, and negative as x_q, x_p, x_n, respectively and denote the margin as t. The error L is expressed by an equation as follows:

$$L(\boldsymbol{x}_q, \boldsymbol{x}_p, \boldsymbol{x}_n) = \max[d(\boldsymbol{x}_q, \boldsymbol{x}_p) - d(\boldsymbol{x}_q, \boldsymbol{x}_n) + t, 0], \quad (1)$$

where d is the L2 distance. Our system learns the feature vector to minimize this error L.

The top 5 frames with the lowest Euclidean distance between feature vectors are selected. We select the best match in each time window of N seconds, carrying out non-maximal suppression in a temporal neighborhood. The key frame selection procedure is illustrated in figure 4. A short clip is generated from each of the top 5 candidate frames, including 10 seconds before and after the selected frame. This clip is presented to the user as a candidate scene of the cooking process.

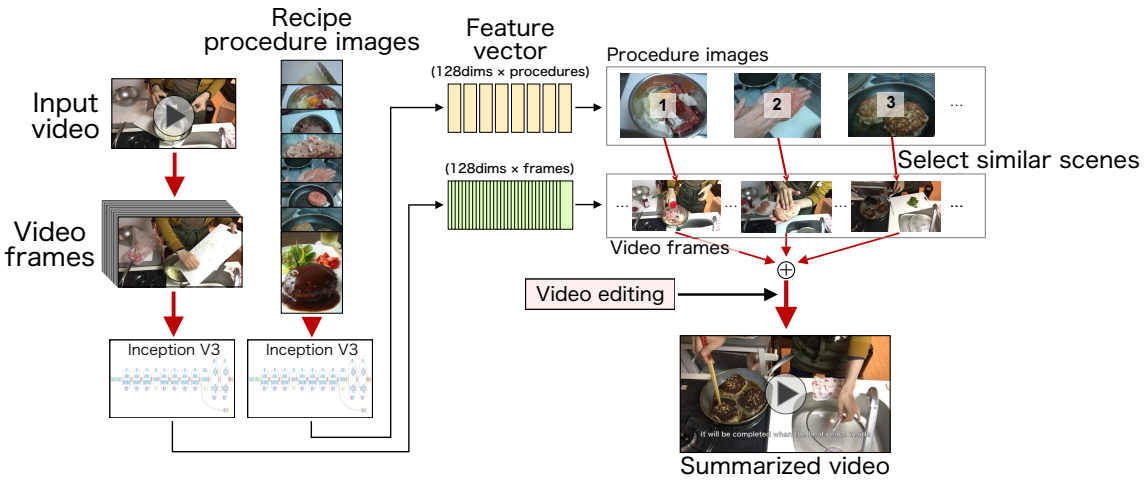

Figure 2. **System Overview.** Recipe instruction images are matched with video frames of the full captured video. Matching is done by taking the Euclidean distance between Inception V3 model feature vectors. The top matches are shown as video clip candidates, which the user can view and select.

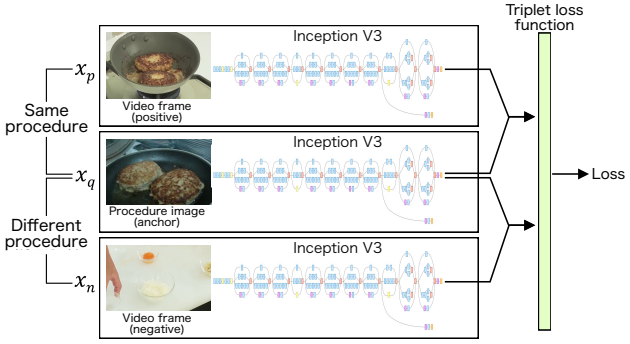

Figure 3. **Learning similarity using triplet loss.** *An Inception V3 network, pre-trained on ImageNet, is fine-tuned on recipe images[13].*

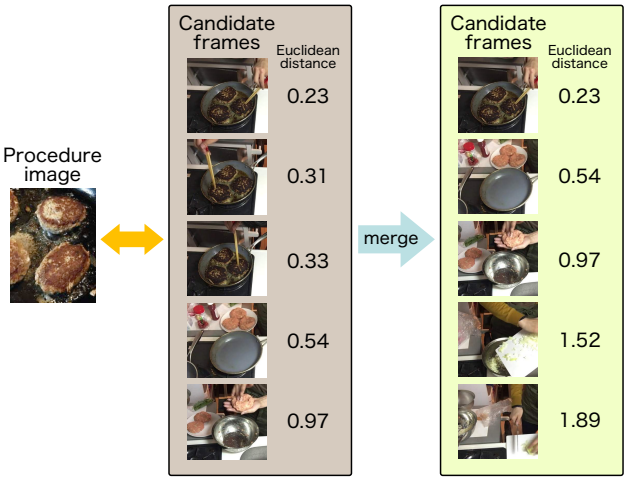

Figure 4. **Key frame selection.** *Frames are sorted in the order of Euclidean distance to the recipe procedure image. If there are multiple candidate matches within N seconds, only the candidate with the closest distance is selected.*

3.4 Video summarization system

Figure 5 shows the user interface of the proposed system. For each cooking process image, 5 short clips are displayed that can be previewed. The user interactively selects the suitable clip as recipe image. The length of each clip is fixed, but can be optionally adjusted by the user. The user can also edit the default subtitles taken from the web recipe image descriptions. When the user finish editing and push the button in the bottom of interface, the system edits clips with several video effects (e.g. speeds up each clips, apply subtitles and connect each clips).

4 Evaluation

We evaluate the proposed system in a user study. In order to train feature vectors, we collected 199 recipes from a recipe website, containing a total of 1014 step-by-step process images. We also collected 17 cooking videos from the website for training. We manually annotate the pairs of corresponding cooking video frames and the process recipe images. Each process image is set as an anchor, and all frames within N seconds before and after in the cooking video that correspond to a recipe process image are selected as positive examples. We randomly select frames N seconds or more before and after the corresponding frame as negative examples. The threshold value N for narrowing down candidate frames is set to 5 seconds. We set of anchor, positive, and negative as one set, and prepared 10140 sets to train. We also implemented the model in Tensorflow and train fine-tuning Inception V3 model, it

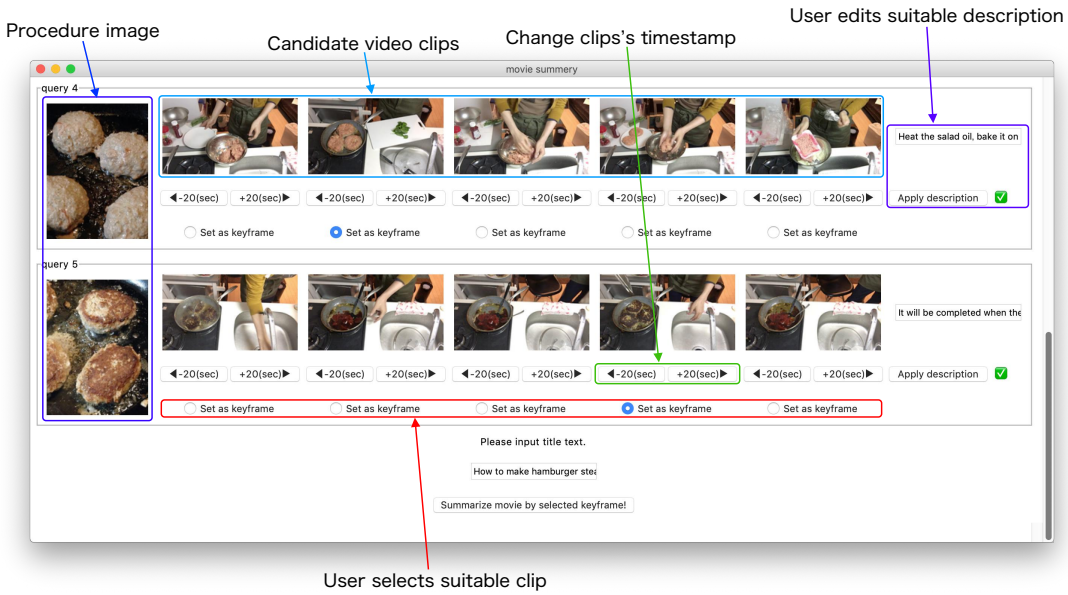

Figure 5. **Interface of the proposed system**. *Multiple candidate clips that match the recipe procedure image are shown to the user, who can select their favorite among them as well as optionally modify the subtitles.*

takes 100 epochs. The cooking video for a hamburger recipe [14] is used and summarized in the evaluation. As shown in Figure 1, a camera is mounted on the wall of the kitchen, captures entire cooking process over 42 minutes.

4.1 Evaluation of system use

9 participants used the proposed system to create a summarized movie. They were asked to answer following questions:

Q1 Could you easily create a summarized video with this system?

Q2 How much time did it take to complete the video?

For Q1, they evaluate using a five-point scale ('strongly agree'/'agree'/'undecided'/'disagree'/'strongly disagree').

Only two out of nine users did not find the system easy to use. Positive comments in the free description were 'it is easy to select a clip of the process from the candidates'. However, since in a few cases no suitable clip was displayed as a video clip candidate, another comment is 'I wondered which frame to select'. Because the top 5 accuracy of clip selection basis of the fine-tuned Inception V3 is about 61.1%, further improvement in order to decrease time to change clip's timestamp is needed. For Q2, the longest time taken to create a summarized video is less than 15 minutes, and the average is about 7 minutes.

4.2 Evaluation of summarized video

39 participants (including the 9 who created summarized videos) answered the following questions after watching the summarized video. They also state whether it satisfied each definition of suitability for SNS sharing in Section 3.1.

Q3 Could you imagine the finished meal from the image at the beginning?

Q4 Could you understand the cooking process?

Q5 Do you think the video satisfied each definition of suitable for sharing on social networks? ('yes' or 'no' for each one)

For Q3, Q4, viewers evaluated using the same five-point scale as in the previous section.

As shown in Figure 7 (Q3, Q4), the total for 'strongly agree' and 'agree' for Q3 was 94.8%. Having a completed scene at the beginning of the video make it easier to grasp the image of the completion of cooking, which should be effective for attracting users' interest. The total for 'strongly agree' and 'agree' with the question for Q4 was 94.8%. An example of cooking recipe video summarized by using the proposed system is shown in the Figure 6. The figure shows video clips including a matching frames for each recipe procedure image. Extracting important scenes in the cooking process from the procedure image and connecting them enables understanding without viewing the entire cooking process.

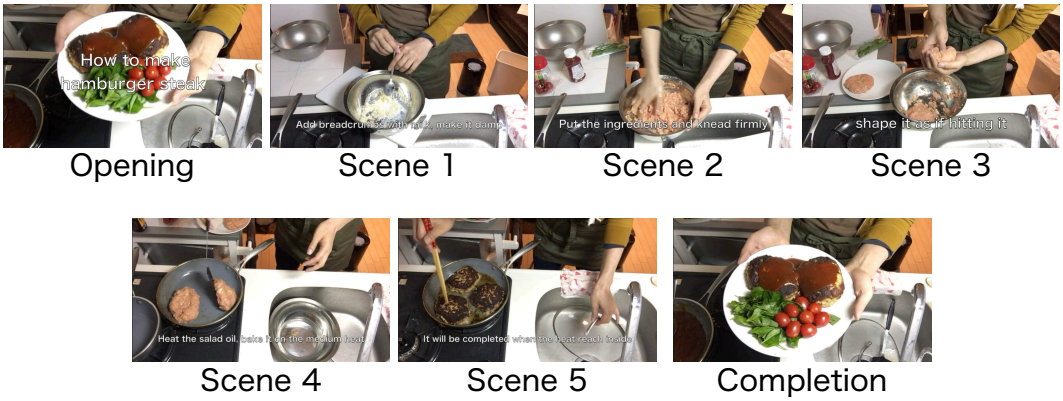

Figure 6. **Example output.** Key frames from a cooking summary video created using the proposed system.

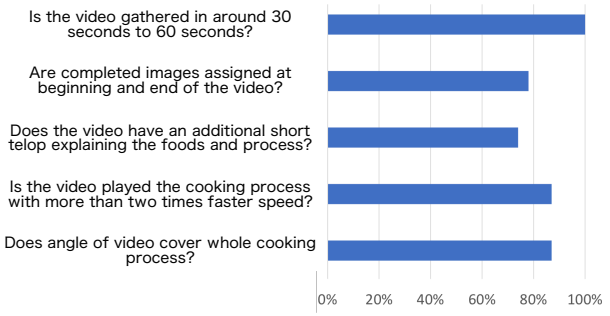

Figure 8. User feedback on the criteria for shareable videos.

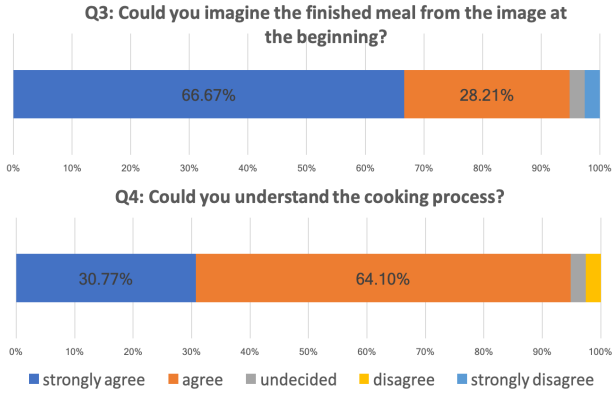

Figure 7. Questionnaire results (Q3, 4).

Figure 8 shows the questionnaire results evaluating five criteria for attractive shareable videos. As shown in Figure 8, more than 75% of the participants answered 'yes' to all criteria. As a result, it shows that a summarized video created using our system can be attracting SNS users' interest, and we confirmed it was possible to show the effectiveness of summarized cooking recipe video.

5 Conclusion

We proposed a new semi-automatic system for summarizing a cooking video in order to attract the interest of social media users. We match recipe process images with video frames, and allow the user to select among the top matches. A description of each cooking process is inserted as subtitle text, and the playback speed is increased. In user studies, the system received positive feedback, both on the ease of creating summarized videos, and on the attractiveness of the summary cooking video.

References

[1] Do house, Inc. `https://www.dohouse.co.jp/news/research/20160705/` (in Japanese, accessed Dec 14, 2018).

[2] Cookpad, Inc. Cookpad studio. `https://cookpad-studio.com/` (in Japanese, accessed Dec 14, 2018).

[3] REAPRA Pte. Ltd. Moveasia. `http://www.moveasia.co/` (accessed Dec 14, 2018).

[4] Ana Garcia del Molino, Xavier Boix, Joo-Hwee Lim, and Ah-Hwee Tan. Active video summarization: Customized summaries via on-line interaction with the user. In *Proc. 31st AAAI Conference on Artificial Intelligence (AAAI'17)*, pages 4046–4052, 2017.

[5] M. A. Smith and T. Kanade. Video skimming and characterization through the combination of image and language understanding techniques. In *CVPR*, pages 775–781, 1997.

[6] Mayu Otani, Yuta Nakashima, Esa Rahtu, Janne Heikkil, and Naokazu Yokoya. Video summarization using deep semantic features. In *Proc. 13th Asian Conference on Computer Vision*, pages 361–377, 2016.

[7] Ke Zhang, Wei-Lun Chao, Fei Sha, and Kristen Grauman. Video summarization with long short-term memory. In *Proc. 14th European Conference on Computer Vision*, pages 766–782, 2016.

[8] Jia Xu, Lopamudra Mukherjee, Yin Li, Jamieson Warner, James M. Rehg, and Vikas Singh. Gaze-enabled egocentric video summarization via constrained submodu-

lar maximization. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 2235–2244, 2015.

[9] Zhao Zhao, Shuqiang Jiang, Qingming Huang, and Qixiang Ye. Highlight summarization in soccer video based on goalmouth detection. In *Asia-Pacific Workshop on Visual Information Processing (VIP'06)*, 2006.

[10] Koichi Miura et al. A motion based automatic abstraction of cooking videos. In *IPSJ Trans. Computer Vision and Image Media CVIM*, pages 21–29, 2003.

[11] Yasuhiro Hayashi et al. Cook-log video summarization by removing temporal redundancy. In *IEICE technical report*, pages 55–60, 2012.

[12] Christian Szegedy et al. Rethinking the inception architecture for computer vision. In *IEEE Conf. Computer Vision and Pattern Recognition, CVPR*, pages 2818–2826, 2016.

[13] Reqruit, Inc. Zexy kitchen. `https://zexy-kitchen.net/recipes/923` (Japanese, accessed Dec 14, 2018).

[14] Rakuten, Inc. Rakuten recipe. `https://recipe.rakuten.co.jp/recipe/1860001109/` (Japanese, accessed Dec 14, 2018).