Learning Cooking Techniques from YouTube

Guangda Li, Richang Hong, Yan-Tao Zheng,
Shuicheng Yan, and Tat-Seng Chua

National University of Singapore, Singapore
{g0701808,eleyans}@nus.edu.sg, {hongrc,chuats}@comp.nus.edu.sg,
yantaozheng@gmail.com

Abstract. Cooking is a human activity with sophisticated process. Underlying the multitude of culinary recipes, there exist a set of fundamental and general cooking techniques, such as cutting, braising, slicing, and sauntering, etc. These skills are hard to learn through cooking recipes, which only provide textual instructions about certain dishs. Although visual instructions such as videos are more direct and intuitive for user to learn these skills, they mainly focus on certain dishes but not general cooking techniques. In this paper, we explore how to leverage YouTube video collections as a source to automatically mine videos of basic cooking techniques. The proposed approach first collects a group of videos by searching YouTube, and then leverages the trajectory bag of words model to represent human motion. Furthermore, the approach clusters the candidate shots into motion similar groups, and selects the most representative cluster and shots of the cooking technique to present to the user. The testing on 22 cooking techniques shows the feasibility of our proposed framework.

Keywords: Cooking techniques, video mining, YouTube.

1 Introduction

Recipes are the natural solution for people to learn how to cook, but cooking is not just about recipes. Underneath the recipe for various culinary dishes, cooking involves a set of fundamental and general techniques, including cutting, filleting, and roasting, etc. These basic cooking skills are hard to learn in cooking recipes, as they only provide textual instructions about certain dish. If user is presented by visual tutorial such as videos for these basic skills, it will be more direct and intuitive for them to understand, such as the examples given in Figure 1.

In recent years, millions of video on the web, make them a source for people to learn some basic cooking techniques. However, web video search engine cannot be automatically applied to visually demonstrate these basic cooking techniques, due to the following reasons. First, search results usually contain irrelevant videos, because textual metadata associated with the video in term of title, tags or surrounding text is not sufficiently accurate to locate videos about cooking techniques. Second, even if the best related video is identified, only part of the video presents this cooking technique.

S. Boll et al. (Eds.): MMM 2010, LNCS 5916, pp. 713–718, 2010.

Fig. 1. Same type of cooking technique has similar motion pattern

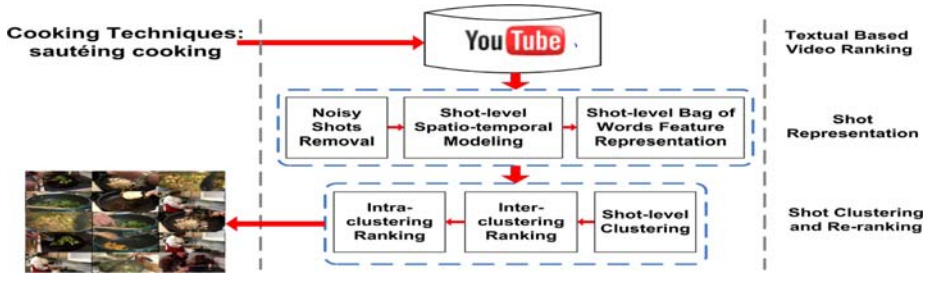

Fig. 2. Overall framework

In this paper, we explore how to mine video tutorials of basic cooking techniques from the noisy web video collection. Our target is to generate a group of representative video shots containing the desired cooking techniques,such as in figure 1. The premise is simple: video shots of the same cooking technique tend to share similar human motion patterns. There are three steps to learn cooking technique patterns. First, a cooking technique, such as "sauteing cooking", is taken as search keyword to submitted to YouTube[1] to find a group of relevant but noisy videos. Then we segment each video into shots. Second, video shots are represented by spatio-temporal bag of words feature. Noisy shots are removed based a set of heuristic rules. Then the cooking technique patterns are mined in an unsupervised fashion. The graph clustering method is utilized to learn the human motion patterns of cooking techniques. Then the most representative cluster is identified by ranking clusters based on their cluster density and scatter. Finally, the most representative shot within this cluster is identified based on its representativeness. The overall framework is given in Figure 2.

2 Related Work

To some extent, our work is to establish multimedia dictionary of cooking techniques. Li et al. [9] and Wang et al. [10] aimed to establish general multimedia

[1] YouTube Website: www.youtube.com

dictionary by leveraging community contributed generated images. However, there are very few works done in cooking domain. Shibata et al. [3] defined the internal structure of cooking videos as three steps: preparation, sauting and dishing up. Then they proposed a framework for acquiring object models of foods from predefined preparation step. Linguistic, visual and audio modalities are utilized to do further object recognition. Hamada et al. [4] considered cooking motions and appearances of foods to be visually important in a cooking video, which contribute to assemble abstract videos. The above two works mainly focus on applying multi-modality features on single video, but the redundant information from different videos is neglected. For example, different cooking videos may involve similar cooking techniques, as shown in Figure 1. With the advance of computer vision, especially in motion analysis, some works have been conducted on extracting spatio-temporal features to represent a certain kind of motion. Dollar et al. [5] used cuboids to extend the 2D local interest points, representing not only along the spatial dimensions but also in the temporal dimension. Ju et al. [2] proposed to model the spatio-temporal context information in a hierarchical way, where three levels of motion context information are exploited. This method outperforms most of others. For this reason, we model the shots by a number of motion features, specifically the sift-based trajectory.

3 Cooking Technique Learning

3.1 Shot Representation

Following most existing video processing systems [11], we take shot as the basic content unit. After segmenting video into shots, the next step is to model shots within a video by spatial-temporal features. We utilize the trajectory transition descriptor proposed in [8] to characterize the dynamic properties of trajectories within a video shot. Each shot is represented by N trajectory transition descriptors. However, it is infeasible to extract the trajectory transition descriptors for all shots of a video. For example, an 8 minutes MPEG format video which is converted from FLV format downloaded from YouTube will expand to 50M. And the overall number of shots for one video will be above one hundred. To minimize the computational workload, we have to identify some shot candidates from the whole video for the trajectory transition descriptor extraction. Due to our observation, some shots make no contribution for cooking motion techniques discovering. And Hamada et al. [4] pointed that in cooking video, shots with faces are usually less important than shots containing objects and motion. These shots are deemed to be noisy and will not contribute to the mining of cooking techniques. We, therefore, only take the rest of the shots for subsequent processing. Then we construct a trajectory vocabulary with 1000 words by hierarchical K-means algorithm over the sampled trajectory transition descriptors, and assign each trajectory transition descriptor to its closest (in the sense of Euclidean distance) trajectory word.

3.2 Constructing Match Shot Graph

After performing trajectory bag of words model on each shot, this set of shots are transformed to a graph representation by their shot similarity, in which the vertexes are shots. The edges connecting vertexes is quantified by its length. For shots i and j, its length is defined as follows:

$$d_{ij} = \frac{\|d_i - d_j\|}{|d_i| * |d_i|} \tag{1}$$

3.3 Ranking Clusters

As the distance between any two shots for a certain cooking technique has been established, we use clustering to expose different aspects of a cooking technique. Since we do not have a priori knowledge of the number of clusters, the k-means like clustering technique is unsuitable here. Therefore, we use mutual *knn* method to perform the clustering, which connect each shot with k nearest neighbors by their similarity. This is of great advantage over other clustering methods because we can simply control the number k of nearest neighbors to each shots but not the number of clusters. Then all these clusters compete for the chance of being selected to be shown to the user. We use the following criteria to compute a cluster's ranking score, similar to what was done in [6]:

$$RC_k = \frac{\sum_{j \in k, 0<m<K, m \neq k} |S_i - S_j|}{\sum_{j=1}^{n} |S_j - \overline{S_k}|} \tag{2}$$

numerator is inter-cluster distance (the average distance between shots within the cluster and shots outside of the cluster), and denominator is the intra-cluster distance (the average distance between shots within the cluster). A higher ratio indicates that the cluster is tightly formed and shows a motion coherent view, while a lower ratio indicates that the cluster is noisy and may not be motion coherent, or is similar to other clusters.

3.4 Learning Representative Shots

Given the result of clusters ranking, we rank the shots within a cluster to find the most representative shot. The representativeness score RS for shot j is calculated as follow:

$$RS_j = \frac{1}{|S_j - \overline{S_k}|} \tag{3}$$

A higher score indicates that the shot is tighter to the center of this cluster, while a lower score indicate that the shot is far from the center of this cluster.

4 Experiment

To validate the effectiveness of our proposed framework, we assembled a collection of videos posted on YouTube from May 2009 to June 2009, by YouTube API. Based on the cooking category in Wikipedia[2] and manually checking the availability of videos for each concept in YouTube, we got 22 cooking concepts. Then we select the top 20 videos for each cooking technique query.

[2] http://en.wikipedia.org/wiki/Category

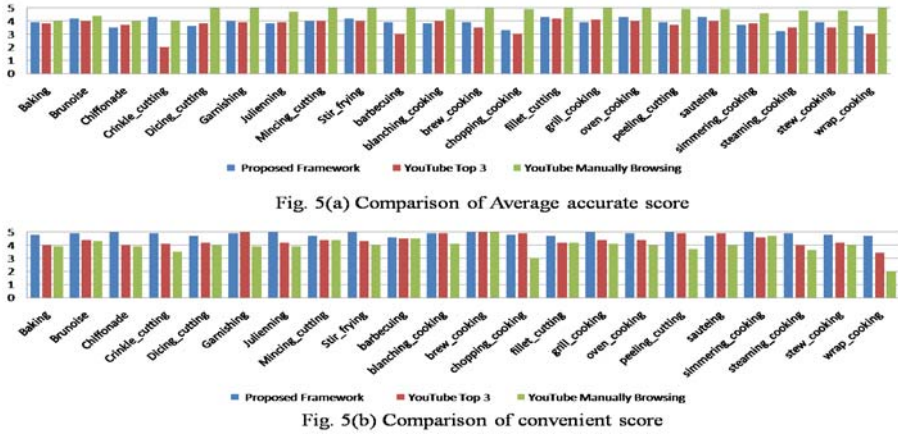

Fig. 3. Same type of cooking technique has similar motion pattern

4 evaluators are involved in our evaluations. Each evaluator was first presented the manually selected shot demonstration as in Figure 1 and the textual explanation for a certain cooking technique. The expectation is that evaluators can understand what a certain cooking technique goes on. After that, the evaluators were shown the top 3 ranked clusters and the top three ranked shots in the first cluster. The evaluators were also asked to search YouTube using cooking technique concepts respectively. They were asked to give satisfaction and convenience score range from 1 to 5 to indicate their preference.

We compare users' satisfaction score for the proposed system and two different strategies of using YouTube. We can see that in Figure 3(a), the overall satisfaction about accuracy of the proposed framework is better than only screening the top 3 videos from YouTube, but not as good as manually browsing and picking a retrieved videos from YouTube. This is because any automatic system cannot perform better than manual methods. In Figure 3(b), the comparison shows that user deemed our proposed our system was more convenient for them to learn the cooking techniques than the other two strategies. This is because evaluators

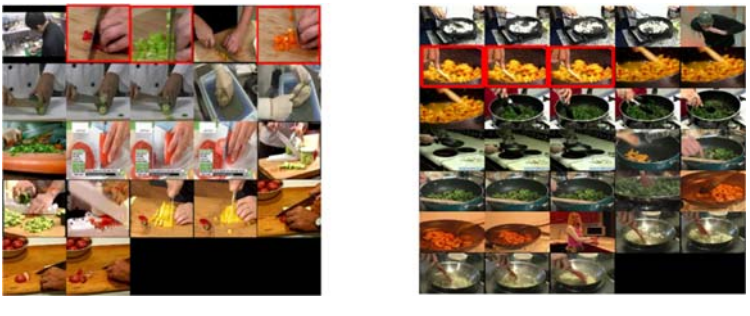

Fig. 4. Same type of cooking technique has similar motion pattern

need to click several times on retrieved videos to discover what they want using YouTube. However, our proposed framework eases the learning workload.

Figure 4 shows the generated group for two cooking techniques. The top three ranked shots are marked with red square. We can observe that some shots are from the same video, which could be caused by two reasons. First, the shot segmentation tool is sensitive to scene change, so continuous shots should be merged. Second, trajectory features for some shots are not discriminative enough for the graph clustering method.

5 Conclusion and Further Work

In this paper, we explore how to leverage YouTube video collections as a source to automatically mine videos of basic cooking techniques. The proposed approach take a simple premise: video shots of the same cooking technique tend to reveal similar human motion patterns. The future works are to mine cooking techniques on web videos utilizing the motion features and the embedded rich metadata, to facilitate the multimedia Question-Answering (QA) system as described in [1], [8] and [7].

References

1. Chua, T.-S., Hong, R., Li, G., Tang, J.: From Text Question-Answering to Multimedia QA. To appear in ACM Multimedia Workshop on Large-Scale Multimedia Retrieval and Mining, LS-MMRM (2009)
2. Sun, J., Wu, X., Yan, S., Cheong, L.F., Chua, T.-S., Li, J.: Hierarchical Spatio-Temporal Context Modeling for Action Recognition. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), Miami, Florida, US (2009)
3. Shibata, T., Kato, N., Kurohashi, S.: Automatic object model acquisition and object recognition by integrating linguistic and visual information. In: ACM'MM (2007)
4. Hamada, R., Satoh, S., Sakai, S.: Detection of Important Segments in Cooking Videos. In: Proceedings of the IEEE Workshop on Content-Based Access of Image and Video Libraries (Cbaivl 2001), p. 118 (2001)
5. Dollar, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior recognition via sparse spatio-temporal features. In: VS-PETS 2005, pp. 65–72 (2005)
6. Kennedy, L.S., Naaman, M.: Generating diverse and representative image search results for landmarks. In: Proceeding of the 17th international Conference on World Wide Web, Beijing, China, pp. 297–306 (2008)
7. Li, G., Ming, Z., Li, H., Chua, T.-S.: Video Reference: Question Answering on YouTube. In: ACM'MM (2009)
8. Hong, R., Tang, J., Tan, H.-K., Yan, S., Ngo, C.-W., Chua, T.-S.: Event Driven Summarization for Web Videos. In: ACM Multimedia Workshop on Social Media, WSM (2009)
9. Li, H., Tang, J., Li, G., Chua, T.-S.: Word2Image: Towards Visual Interpretation of Words. In: ACM'MM 2008 (2008)
10. Wang, M., Yang, K., Hua, X.-S., Zhang, H.-J.: Visual Tag Dictionary: Interpreting Tags with Visual Words. In: ACM Multimedia Workshop on Web-Scale Multimedia Corpus (2009)
11. Tang, S., Li, J.-T., et al.: TRECVID 2008 High-Level Feature Extraction By MCG-ICT-CAS. In: TRECVID Workshop, Gaithesburg, USA (2008)