Kitchen Gesture Recognition with Microsoft Kinect based on Scene Context

Dushyant Goyal, Shubham Bansal, Shubham Gupta, Shubham Khandelwal
The LNM Institute of Information Technology, Jaipur, India
goyal1dushyant@gmail.com, shubbansal27@gmail.com, shubhamgupta5893@gmail.com, skhlnmiit@gmail.com

Abstract

In this paper we propose a novel approach to a challenging problem of daily life cooking gesture/activity recognition task based upon object use and frame sequence tagging. We use a dynamic SVM-HMM hybrid model which combines structural as well as temporal video sequence information to jointly infer the most likely cooking activity labels. Such a context based approach as discussed in this paper can be extended to other fine grain activities domain such as hospital operating rooms in medical practices, agricultural and manufacturing operations, etc.

1. Introduction

Computer-based human activity recognition of daily living has gained a lot of interest in recent years. With a growth of the elder population in the society and with a rapid increase in applications such as human-computer interaction, house-hold robotics, smart homes, computer assisted child care, suspicious activity identification, etc. activity recognition has become a broad area of research. Despite of this fact, activity recognition for various environments is in their nascent stage, especially identification and classification of cooking activities in kitchen scenario because of high variability in action execution, complex cooking motions involved, numerous cooking menus, different cooking styles, etc.

In this paper we address the problem of recognizing human activities in indoor environment with a focus on kitchen scenario. We believe that scene context based gesture recognition can be applied for various uses like real-time analysis of a cooking scene will enable a system to advise a beginner what he/she should do at the next step in a cooking procedure/recipe or may aid to a person with disability to recover his mistakes.

This work was done as a part of a contest "Kitchen Scene Context based Gesture Recognition" to be held during the International Conference of Pattern Recognition (ICPR) -2012.

2. Cooking Video Syntax and Challenges

There are five candidate kitchen cooking menus using the ingredients of egg, milk, and ham namely i) boiled-egg ii) ham & egg iii) kinshi-tamago iv) omelet v) scramble-egg. The RGB-D kitchen scene is captured using a Kinect sensor providing synchronized color and depth image sequences.

Cooking actions performed in the dataset are - breaking, mixing, baking, turning, cutting, boiling, seasoning, peeling, none.

One of the challenges in the kitchen dataset is the amount of variability present in execution of activities as no fix sequence of steps were followed by the subjects to cook a particular recipe. For example, for preparation of "ham-egg" recipe - "mixing" can be performed before cutting as well as after "cutting".

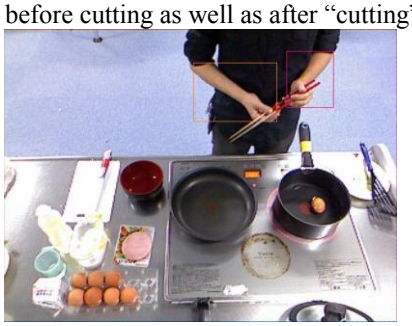


Figure 1. Sample Cooking Video Frame

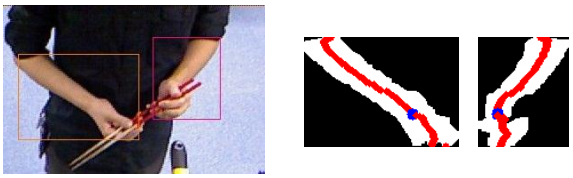
3. Frame Classification – A context Based Approach

3.1. Hand Segmentation and Tracking

The detection of hand region is achieved through color segmentation in YCbCr color space and fusing the result with the depth information to generate a

binary image of the human skeleton with hand segments.

To better understand different cooking actions, it is realized that hand pose can provide strong cues particularly for fine activities like peeling, cutting, etc. Hand pose i.e. angle of the hand with the horizontal is estimated as show in Figure 2 (b).



(a) (b)

Figure 2. Left and right part of hand segmented and angle estimation.

3.2. Object Recognition and Use Identification

Kitchen tools and ingredients are extracted from the training dataset and individual image templates are created. In this scenario we have considered the following 10 objects for recognition as shown in Figure 3. In our approach, for every frame we have tried to identify the objects that are in use. For each object we maintain a binary status of In Use or Not in Use i.e. 1 or 0.

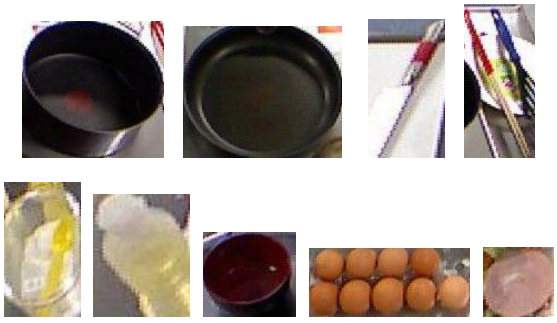


Figure 3. Objects/Ingredients templates (top to bottom, left to right) – Pan, Frying Pan, Knife, Turner & Chopsticks, Salt Box, Vegetable Oil Bottle, Ham, Eggs.

3.3. Features description

The object "Use" or "Not in use" status described in the previous section are used as binary features for supervised classification. A 10-dimensional feature for each frame is described as follows – Turner, Chopsticks, Knife, Salt Box, Bowl, Frying Pan, Pan, Vegetable Oil Bottle, Ham, Eggs, Hand (0 representing No action). A few other spatial and temporal features are also used.

4. Training and Classification

We use contextual information to find the best solution for an entire cooking video sequence at a time using the hybrid multi-class SVM-HMM method proposed in [1]. Each cooking video is a Sequence, features represent the Observations, and 9 different classes represent the States for an HMM model. Each frame in the Sequence is classified into different States/classes based upon the Viterbi Algorithm.

For training the classifier, the original labels and the feature patterns were obtained from the annotated dataset comprising of 25 videos and a training dataset [5] was constructed that consisted of 1,18,848 patterns for all 9 classes combined.

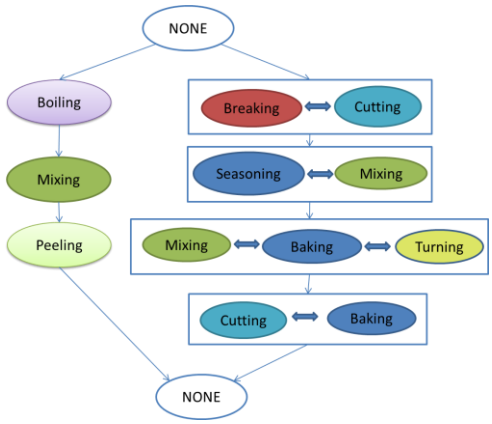


Figure 4. Activity Graph for cooking activities

5. Demonstration

The demonstration will consist of a pre-recorded dataset from a kinect sensor with subjects performing cooking gestures like cutting, peeling, etc. Both the depth and color information is used for action recognition to ensure good results in cluttered environments. In addition, an online working demo of the project is in pipeline and we plan to add additional features to it. For example, various prompts for the user while he is cooking are being incorporated, etc.

5. Conclusion

In this work we have developed a context based gesture recognition scheme for kitchen activity recognition. We made use of the contextual information from the scene both in spatial and temporal domains and designed a robust feature space. Our experiments on the cross-validation set (accuracy of 72%) also prove the competitiveness of SVM-HMM as it is better able to model the frame sequence learning problem.

References

[1] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun, Large Margin Methods for Structured and Interdependent Output Variables, *Journal of Machine Learning Research (JMLR)*, 6(Sep):1453-1484, 2005.

[2] Spriggs, Ekaterina H.; de la Torre, Fernando; and Hebert, Martial, "Temporal Segmentation and Activity Classification from First person Sensing" (2009). Robotics Institute. Paper 324.

[3] I. Laptev and T. Lindeberg. Space-time interest points. In ICCV, pages 432–439, 2003.

[4] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In CVPR, 2008.

[5] Contest on Kitchen Scene Context Based Gesture Recognition (KSCGR), *International Conference on Pattern Recognition (ICPR)*, 2012, Japan.

[6] P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional sift descriptor and its application to action recognition," *ACM Multimedia*, 2007

[7] R. Messing, C. Pal, and H. Kautz. Activity recognition using the velocity histories of tracked keypoints, *ICCV*, 2009.