KITCHEN ACTIVITY RECOGNITION BASED ON SCENE CONTEXT

Shubham Bansal, Shubham Khandelwal, Shubham Gupta, Dushyant Goyal

The LNM Institute of Information Technology, Jaipur, India
shubbansal27@gmail.com, skhlnmiit@gmail.com, shubhamgupta5893@gmail.com, goyal1dushyant@gmail.com

ABSTRACT

In this paper, we propose a novel approach to a challenging problem of daily life cooking 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. We demonstrate that our approach can achieve activity recognition rates for kitchen scenarios of more than 72% on a real-world cooking dataset consisting of 9 cooking activities with significant variations in performance of these activities by different subjects. Such a context based approach as discussed in this paper can be extended to other fine grain activities such as hospital operating rooms in medical practices, agricultural and manufacturing operations, etc.

Index Terms— Kitchen, Cooking Activity Recognition, Frame Classification, Kinect

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. Despite of this fact, activity recognition for various environments is in their nascent stage, especially identification and classification of cooking activities in kitchen scenario. This is 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, which in turn 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. Moreover, we expect that scene analysis and classification of recorded videos can also provide context-based segmentation of image sequences, and facilitate automated scene annotations for video databases. Other applications such as indexing and extracting knowledge could be possible result of such a research. Our system enables efficient recognition of all cooking gestures for various cooking recipes.

The remaining paper is organized as follows. Sections 2 discuss some related work for activity and gesture recognition. Section 3 describes the cooking video syntax, dataset and its challenges. Section 4 proposes the context based approach to activity recognition and explores various features for successful frame classification. In section 5, training and classification algorithms used are described. Section 6 comprises of experimental results.

2. RELATED WORK

Local spatial and temporal interest points based feature have been widely used for action recognition because they can be used without any pre-processing like background modeling, motion estimation, object recognition, etc. These features are quite robust to illumination changes to form a sparse representation of actions and can be effectively integrated with machine learning algorithms. At the same time, many spatial local feature descriptors have been extended to the temporal domain, like the temporal histograms of oriented gradients (HOG) [1], SIFT [2], etc. Optical flow [3] based descriptors have also been proposed to capture the motion information. However, these features have the inability to represent complicated and long-range motions, with a lot of interactions with surrounding objects.

A gesture-based system using a multi-dimensional hidden Markov model (HMM) [4] is also being developed by many researchers. HMMs are employed to represent the gestures and their parameters are learned from the training data. Based on “the most likely performance” criterion, the gestures can be recognized through evaluating the trained HMMs.

Traditionally, most gesture recognition methods focus on motion features only, and assign the gesture label based on the discriminative analysis. There are, however, many gestures which cannot always be uniquely determined by using motion features alone. In this paper we propose a context based approach for activity classification with SVM-HMM sequence tagging based model. The method is motivated by the fact that identification of object cues and

temporal information can effectively serve as a good paradigm for video frame classification.

3. COOKING VIDEO SYNTAX/CHALLENGES

This work was done on the dataset provided for kitchen gesture contest in ICPR 2012 [5]. There are five candidate kitchen cooking menus namely (i) boiled-egg (ii) ham & egg (iii) kinshi-tamago (iv) omelet (v) scramble-egg. Each menu is cooked by seven different actors (five actors in training datasets and two actors in evaluation datasets); i.e. five cooking scenes are available for each training menu. The RGB-D kitchen scene is captured using a Kinect sensor providing synchronized color and 8-bit depth image sequences (depth unit of each level is 8mm). There are a total of 8 cooking gestures in the dataset, namely, breaking (eggs), mixing (mix something in bowl or pan using chopsticks), baking (in frying pan), turning (using turner), boiling, peeling (egg), cutting (cut something using knife) and seasoning. In addition there are frames given no action labels.

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.

4. FRAME CLASSIFICATION - A CONTEXT BASED APPROACH

4.1. Hand segmentation and tracking

In the proposed method the detection of the hand region is achieved through color segmentation, since color is a robust feature that performs with good accuracy even in challenging environmental conditions. Color is invariant to rotation, scaling as well as to geometric variations of the hand, and most importantly it allows a simple and fast processing of the input image.

The proposed YCbCr color space separates luminance from chromaticity components since chromaticity preserves the useful color information. So, if the respective image pixel values of Y, Cb, Cr lie in their determined ranges, the pixel is classified as skin or hand pixel. But, in case of kitchen scenario several other objects like eggs and ham also lie in the same skin range. So, in order to isolate the hand regions from such objects we apply back-ground modeling to the depth images and generate a binary image of the human skeleton with hands i.e. the area changing with each frame. Both of these segmentation results are fused together to obtain the hand segments. To remove outliers, we apply morphological operations like erosion and dilation. Small and unwanted contours are removed using CCA and finally, the frame has 2 white contours as Left and Right Hands. The 3-D hand centroids coordinates (x, y, z) are stored for all further use. It is worth to underline also, that the segmentation results are very good (almost noiseless) Fig. 1(a).

Fig. 1. (a) Hand segmentation (b) Left and right part of hand segmented and angle estimation

4.2. Left and right hand pose estimation

To better understand different cooking actions, it is realized that hand pose, i.e. angle of hand w.r.t horizontal, can provide strong cues particularly for fine activities like peeling, cutting, etc. For instance, when the knife cuts an ingredient, the hand has a particular signature or angle which enables our classifier to identify cutting. In our first person view only the lower hand is visible and is of major concern. So, we segment the hand into two different segments – i) elbow to wrist ii) wrist to finger tips. The corresponding angles w.r.t. horizontal is estimated (Fig 1.b).

4.3. Object segmentation and recognition

Object localization and recognition are the preliminary and important steps for performing action recognition since objects pose a strong association with cooking motions. For extracting features from videos of kitchen recipes one needs to have exact location of each and every object present in every frame of video. Hence, kitchen tools and ingredients are extracted from the training dataset using Active Contour Segmentation method [6] and individual image templates are created.

Object classification was carried out using SVM model. Local features like area, centroid, convex hull, convex area, eccentricity, equiv diameter, euler number, major/minor axis length, mean intensity, depth, orientation, etc. were used for learning the model [7].

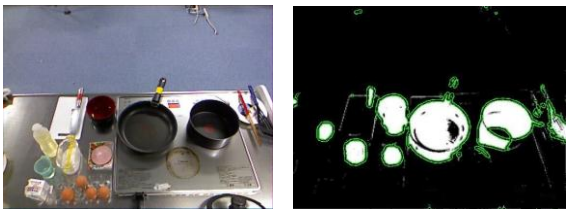

Fig. 2. Object segmentation and classification.

In this scenario we have considered 10 objects for recognition - Turner, Chopsticks, Knife, Salt Box, Bowl, Frying Pan, Pan, Vegetable Oil Bottle, Ham and Eggs. All the objects are identified from the first video frame and their positions are stored. All further tracking or processing on these objects is done starting from their initial positions. The accuracy of Object recognition was approximately 81%.

4.4. Object use identification

Scene context plays a prominent role in the identification of cooking motions being executed. For example, the use of "knife" directly corresponds to the action of "cutting" and use of "bowl" and "chopsticks" together must imply the action being performed as "mixing". Thus, by exploiting semantic relationships between different objects, our approach can detect various cooking motions successfully.

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. 0 or 1. The status of an object is updated to "In Use" when one of the following conditions arises –

4.4.1. Change in object appearance, for example putting egg into frying pan changes the appearance of the pan from black to yellow

4.4.2. Object displacement from initial or stationary position.

4.4.3. Grasping an object i.e. hand approaching the objects or the distance between object and hand reducing to a certain threshold

4.4.4. Continuous object movements

Since, our hand tracking is highly accurate we create a region of interest around the hand co-ordinates and locate the objects which might be possible candidates to be in use i.e. knife, chopsticks, bottle, etc. In future we plan to use a more generalized and robust object recognition algorithm.

4.5. Features description

The object "Use" or "Not in use" status of 10 objects described in section 4.4 are used as binary features for supervised classification. Hand, arm and wrist angles estimated in section 4.2; x, y, z coordinates of both hands; velocity magnitude and direction; distance of hand co-ordinates and angle they make with different objects; are other features which are used. In addition, some spatial and temporal features are also added making the total feature space of size 82.

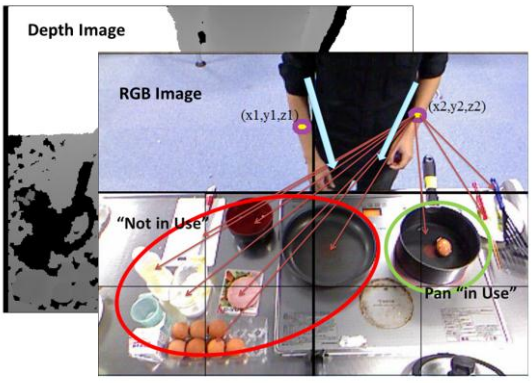

Fig. 3. Illustration of different features using hand centroids.

5. TRAINING AND CLASSIFICATION

Supervised classification was performed to classify the image frames into 9 classes.

5.1. SVM-HMM (Support Vector Machine – Hidden Markov Model)

Frame Classification problem can be modeled as sequential tagging problem. We need to model the task as a stochastic process. Here, we aim to classify a frame or a sequence of frames with respect to one another. We could classify each frame without the contextual and temporal information, ignoring other frames in video sequence using SVM. However, this approach might work for unambiguous case of "Boiling", but context is highly crucial to classify frame sequences like breaking, peeling, mixing. Although these actions are performed using "bowl", but depending on the contextual information of neighboring frames it would be difficult to get accurate results. Kitchen activities are rather dependent on the neighborhood labels (e.g., breaking -> mixing -> baking). Thus, we propose to use contextual information to find the best solution for an entire cooking video sequence at a time.

One of the significant drawbacks in Support Vector Machine (SVM) is that they are inherently static classifiers and do not implicitly model temporal evolution of data. On the other hand, Hidden Markov Model (HMM) has the ability to handle with certain assumptions about stationarity and independence. So, in this paper we use the hybrid SVM-HMM method proposed in [8] - a publicly available toolkit for sequence tagging which takes advantage of inherent properties of both SVM and HMM.

Each cooking video is a sequence, features represent the observations, and 9 different classes represent the States for an HMM model. The transition probability is estimated from the training dataset and emission probability is found using SVM classifier. Finally 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 was constructed that consisted of 1,18,848 patterns for all 9 classes combined. The SVM-HMM classifier was trained using this training dataset [9].

5.2. Output post-processing / anomaly detection

After frame classification, the output label sequences still suffer from certain anomalies which can be fixed. For instance, some frames are classified wrongly among a cluster of correctly classified frames i.e. noise in the output label time series. We apply a smoothing operation to remove such noise and re-classify such frames to their neighborhood

frames by majority voting. In some cases, the classifier gives false results with certain kind of actions like "peeling", "mixing" ("In Use" status of bowl is 1 for both). But, these actions are temporally dependent on previous frames. For example, "peeling" action only occurs after "boiling" has taken place and hence event of "peeling" without "boiling" has been misclassified. Figure 4 depicts a context grammar which aids to better understand this inter-relationship among the actions and helps in identification of such anomalies. The next task is to rectify this wrong classification which is done with a most likely guess. For example, if "peeling" is classified before "boiling" the possible candidates for correct classification would be "mixing" because at that time bowl must be "In Use" and might be wrongly classified.

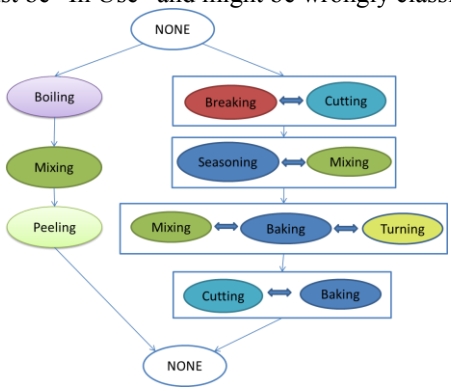

Fig. 4. Context Grammar/ Activity Graph for cooking activities.

6. EXPERIMENTAL RESULTS

We used linear kernel function for both the SVM and the HM-SVM training procedure. For soft margin C we used a range of values [10, 50, 100, 500], where value C=50 produced the best model with the best accuracy. The results are summarized in Figure 5 and Table 1 and they clearly demonstrate the competitiveness of SVM-HMM and our feature space. As expected, SVM-HMM achieved the best result outperforming SVM which validates our approach of using a hybrid classifier.

For activity recognition, we compare our method with the state-of-the-art vision approach (RGB only), spatial-temporal interest points (STIP) [4] bag-of-words model. STIP does not require hand tracking (difficult to do in general) and is the basis of many recent works on action recognition. We find that the accuracy of STIP is 35% and F-Score is 0.47, while our context based approach, which is much simpler in nature and faster to compute, achieves a higher accuracy of 64% with an F-Score of 0.61. This shows the advantage of explicitly tracking hands, objects even though hand positions are not always accurate (in particular the separation of hands from objects-in-hand using color). The benefit of using SVM-HMM approach is that it inherently combines spatial as well temporal context for frame classification. By feeding post processing (PP) based upon Context Grammar (Figure 4) into action recognition, the accuracy is further increased to 72% and F-Score to 0.68.

Table 1. Average Recognition Accuracy, F-Score, Precision and Recall for 9 cooking gestures for the test datasets

Method	Accuracy	F-Score	Precision	Recall
STIP	35.72	0.47	0.35	0.41
Context Based – SVM	49.66	0.52	0.55	0.54
Context Based SVM-HMM	64.28	0.61	0.62	0.63
Context Based SVM-HMM (Post Processing)	71.79	0.68	0.68	0.68

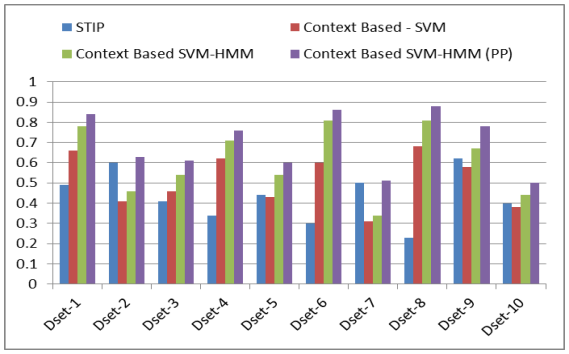

Fig. 5. Recognition F-score by different methods on 10 test datasets

7. CONCLUSION

In this work we have developed a context based gesture recognition scheme for kitchen activity recognition. We make use of the contextual information from the scene both in spatial and temporal domains and designed a robust feature space. However, the proposed features are dependent on the dataset, but the idea is quite robust. Our experiments also prove the competitiveness of SVM-HMM since it has the ability to model the frame sequence learning problem.

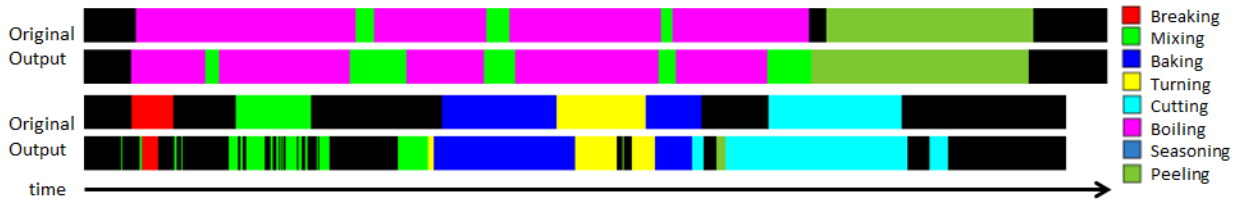

Fig. 6. Time series representation of Ground Truth vs Proposed Output for two testing cooking video datasets.

8. REFERENCES

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

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

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

[4] Yang, Jie, Yangsheng Xu, and Chiou S. Chen. "Human action learning via hidden Markov model." Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on 27.1 (1997): 34-44.

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

[6] Chan, T.F.; Vese, L.A. "Active contours without edges," Image Processing, IEEE Transactions on , vol.10, no.2, pp.266-277, Feb 2001

[7] Jianguo Zhang, Marszalek, M, Lazebnik, S, Schmid, C. "Local Features and Kernels for Classification of Texture and Object Categories: A Comprehensive Study," International Journal of Computer Vision June 2007, Volume 73, Issue 2, pp 213-238

[8] 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.

[9] http://www.cs.cornell.edu/people/tj/svm_light/svm_hmm.html