Southeast Asian Journal of Sciences: Vol 4 , No. 2 (2016) pp. 73-83

RECOGNIZING FOOD PREPARATION ACTIVITIES USING BAG OF FEATURES

Nguyen Thi Thanh Thuy and **Nguyen Ngoc Diep**

Department of Information Technology
Posts and Telecommunications Institute of Technology (PTIT)
Hanoi, Vietnam
e-mail: thuyntt@ptit.edu.vn

Abstract

Food preparation activities for cooking in the kitchen involve physical interactions between multiple objects such as hands, utensils, and ingredients. Recognizing these complex activities using sensors embedded in kitchen utensils is challenging. For accurate recognition, it is necessary to design efficient feature representation of sensor data. In this paper, we propose a feature learning method based on bag of features for food preparation activity representation and recognition. The activity model is built using histograms of motion primitives. We experimentally validate the effectiveness of the proposed approach for recognizing ten activity classes. The experiment results show that the proposed approach provided substantially higher accuracy than traditional approaches for food preparation activity recognition using embedded sensors.

1 Introduction

Activity recognition is an active research field with a wide range of potential applications. One important application is situated services for supporting people's lives. In order to provide automatically situated support for people with cognitive impairments in their homes, it is necessary to recognize activities of daily living. Food preparation is one of the essential tasks in daily life and it involves a large number of physical interactions between hands, utensils, ingredients, etc. Recognizing these activities can help to recognize the particular

Key words: Sensor, food preparation activity, bag of features, motion primitive.

ingredients or to reason the intention of the person who is cooking. So that the system can assist people in cooking food if needed. However, because of the complex interactions between multiple objects when processing food and high intra-class variability, it is extremely challenging to recognize food preparation activities.

There are two common approaches to activity recognition which can also be applied for recognizing food preparation activities are computer-vision based and wearable sensor-based. In computer vision-based approach, activities are recognized from video and still images captured by digital cameras equipped in the environment. In wearable sensor-based approach, signal streams from sensors which can be instrumented in the surrounding environment or worn on users or embedded in the utensils are analyzed to detect the activities. For food preparation activity recognition, there are several computer vision-based researches which achieved promising results such as [8, 12, 13]. However, the limitation of this approach is that the system is strictly limited to the area equipped with cameras. Therefore, in this study we focus on sensor-based activity recognition, especially sensors like accelerometers are embedded in kitchen utensils [9]. This approach is more flexible than the former. Moreover, while sensors and cameras equipped in the environments can expose sensitive data and cause privacy concerns among users, there is no such problem with using embedding sensors in utensils.

Sensor-based food preparation activity recognition is a typical time series analysis problem. To recognize activities, sensor data streams are commonly segmented into frames using a sliding window. Then for each frame, features are extracted by transforming the sensor signals into a feature vector. Finally, the feature vector is classified using any classifier such as k-NN or Decision Tree. Like any general purpose pattern recognition problem, to achieve accurate recognition, it is necessary to design appropriate feature representation of sensor data. Good features should be able to clearly separate between different activity classes.

Food preparation activities are very complex since they involve many different utensils and ingredients. Even a simple task like salad and sandwich preparation requires more than ten activities such as chopping, peeling, slicing, dicing, coring, spreading, stirring, scooping, scraping, shaving, etc [9]. In addition, characteristics of these activities are also very complex. For example, some of them have movements in one direction but the others have movements in multiple directions. Some are repetitive or not. Others have lots of changes in movements and directions but some only change a little. Since the characteristics of food preparation activities are too diverse, features commonly used in activity recognition such as mean, variance, entropy, correlation, etc. computed from sensor signals [2] are not efficient enough (see results in [9]).

Another approach to deal with these challenging set of activities is to carefully design a big set of features with more complex ones like kurtosis, skewness

or some important statistical features often used in time series problems (i.e. speech recognition) like zero crossing rate, mean crossing rate or first order derivative, etc. However, such heuristic feature design approach can not guarantee the appropriateness of the features to clearly separate the complex food preparation activities. In our experiments below, we explored a complex set of statistical features to prove this claim.

A possible solution to this problem is using multilevel features which have shown good recognition performance on complex activities in recent activity recognition works [3, 4, 16]. The features are extracted from sequential data based on feature learning using bag of features, which can automatically discover meaningful representation of data to be analyzed. Bag of features is often used in text categorization and image classification [6,14]. First, local features are extracted from small segments of each activity frame. Then these local features are grouped to form motion primitives in order to generate higher level features using a clustering algorithm like k-means. The use of motion primitives combines with the bag of features create histogram feature vectors which form the activity model. This approach has resulted in significant advances in activity recognition. Therefore in this paper, we propose a motion primitive-based model using bag of features for recognizing food preparation activities.

The rest of this paper is organized as follows. Section 2 gives a brief survey of recent works on food preparation activity recognition using sensors embedded in utensils. Section 3 introduces the dataset and the proposed method. Section 4 presents the experimental results and compares the performance of the proposed method and the heuristic feature design approach. Section 5 is the conclusion.

2 Related Work

There is a lot of works on activity recognition using wearable sensors but just a few on recognizing food preparation activities. The recent work of Stein and MacKenna [13] follows a multi-model approach using both accelerometers attached to kitchen objects and an RGBD-video camera, which help to combine generic and user-specific data from multiple sensor modalities. The multi-model approach can take advantages of complementary information but require presettings in the environment and may raise privacy concerns form users. Another work by Pham et al. [9] which based on accelerometers embedded in utensils has proposed to use some popular statistical features extracted from the temporal domain. Even the recognition result can demonstrate that a broad set of food preparation actions are able to reliably recognized using sensors embedded in kitchen utensils but it is not feasible to use in practical.

For the approach of motion primitive based model using bag of features for

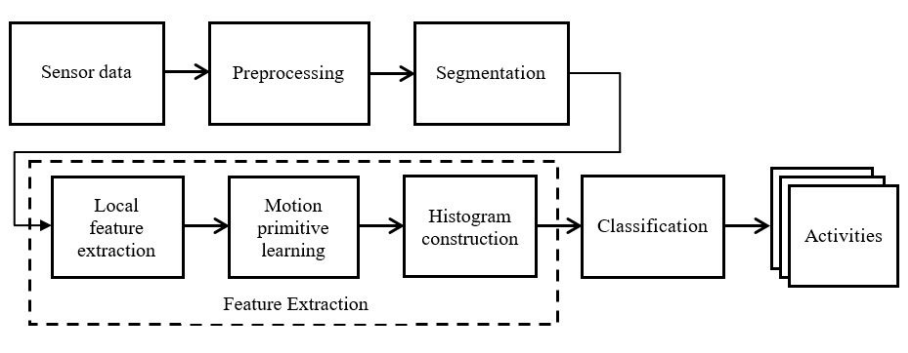

Figure 1: Schema for bag of features based food preparation activity recognition using accelerometers

activity recognition, there are several works used recently like [3, 4, 16]. The basic principle of this approach often consists of three steps: local feature extraction, motion primitive learning and motion histogram construction. Local features are extracted from small segments of sliding windows. In this step, local characteristics of activity signal are captured. Motion primitives can be learned by using a clustering algorithm like k-means or Gaussian mixture model to group local features into clusters. Then each cluster center forms a motion primitive. And motion histogram is constructed based on the occurrence of motion primitives in an activity frame. In this study, we follow this approach to efficiently recognize a fine-grained food preparation activities. In additional, a set of statistical features used in local feature extraction is designed (presented in section 3.3) to capture local characteristics of food preparation activities. This feature set is more efficient than the simple local feature set with mean and variance, which is often used in similar motion primitive based model for activity recognition [4, 16].

3 Proposed Method

This section presents the schema for food preparation activity recognition using accelerometers with four steps: preprocessing, segmentation, feature extraction and classification. The most important step is feature extraction using bag of features model to extract appropriate features for recognizing food preparation activities. Figure 1 shows the proposed schema for for food preparation activity recognition.

3.1 Preprocessing

In preprocessing step, data is acquired using sensors embedded in kitchen utensils. Some sensors can provide multiple values (i.e. multiple directions) or multiple sensors are jointly sampled. Each of the sensors have it own sampling rate and the sampling rates of the different sensors can differ. Some sensors may change their sampling rate for some reasons, for example, for power saving. For this reason, in preprocessing step, it is required to synchronize the input sensor data to prepare for feature extraction step. In addition, data are often missing or being distorted due to noise or calibration effects. To alleviate these unwanted effects, it is necessary to denoise and fill the gap for the input data. It is important to notice that preprocessing step need to keep characteristics of sensor signal containing activities of interest.

3.2 Segmentation

In this step, the input signal stream is segmented into frames or windows by using a sliding window. The size of sliding window influences performance of the system [5]. In our problem of recognizing food preparation activities, empirical study for finding best window size on the experimental dataset has been performed for segmentation step.

3.3 Feature Extraction

Good features are crucial to improve the classification accuracy of any activity recognition system. Previous approaches like [9] often extracted feature values from activity frames using features heuristically designed. In this study, we use bag of features approach for automatically learning appropriate features to improve the accuracy. Therefore, each frame is then divided into smaller segments with overlap using another smaller sliding window. These segments are much smaller than the frame and are called slices. Size of the slices is also very important to the performance of recognition. Since there is not many studies about effective size for window slice, in this work we discover the best window slice size via tuning in the experiments. From each slice, we extract feature values using a set of statistical features to form a local feature vector.

Notice that, the chosen features should be simple enough so that they can reliably be calculated from slices which are very small segments. Moreover, the feature set need to be carefully selected to capture multiple characteristics of food preparation activities. In this work, we propose a set of several simple statistical features including mean, variance, standard deviation, and correlation. In which, mean is effective for differentiating postures like sitting, standing, lying [1] (based on calculating gravity vectors) so that it is useful for distinguishing food preparation activities using a chopping board or not. Variance shows the level of movement in the signal [5] and it may help to

differentiate chopping, dicing, slicing as these activities are similar but different in movement intensity. Standard deviation can tell about the spread out from the average and is often used in combination with mean and variance [2]. Correlations between pairs of acceleration axes can help to separate activities related to one direction and multi-direction movements [11] and hence it can differentiate activities like dicing, chopping, slicing, scraping and activities like stirring, scooping, coring.

After that, local feature vectors from all training activity frames are pools together and similar local feature vectors are grouped in same clusters to form motion primitives. Set of all motion primitives is called vocabulary. Common clustering algorithms like k-means [3, 7, 16] or Gaussian Mixture Model [16] can be used in this process. In the next step, we combine motion primitives learned with bag of features model in order to find higher level features. In which, occurrences statistics of motion primitives are calculated on each frame to form histogram feature vector. These feature vectors are taken as input features for classification algorithms.

3.4 Classification

In order to show the advance of proposed feature extraction method based on bag of features, a very popular classifier C4.5 decision tree - which is widely used in classification problems [15] is chosen for classifying fine-grained food preparation activities. C4.5 are built from a set of training data based on the concept of information entropy. At each node of the tree, C4.5 chooses the attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other using the normalized information gain criteria. The output classifier can accurately predict the class to which a new case belongs [10]. C4.5 is also used in Pham's work [9] with the same dataset.

4 Experiments and Results

4.1 Dataset

We conduct experiments on Ambient Kitchen dataset created by Pham et al. [9]. This dataset contains data streams from tri-axial accelerometers embedded in kitchen utensils. Sensor readings are sampled at the rate of 40 Hz. Participants in the experiments performed various food preparation activities in the kitchen using different utensils with different ingredients. In details, the food ingredients used for salad and sandwich preparation task include potatoes, tomatoes, lettuce, carrots, onions, garlic, kiwi fruit, grapefruit, pepper, bread and butter. Four kitchen utensils embedding accelerometer used in food preparation task are big knife, knife, small knife and spoon. Twenty subjects

are required to freely perform any actions for salad and sandwich preparation under no limitations. The number of activities in the dataset is eleven consisting of chopping, peeling, slicing, dicing, coring, spreading, eating, stirring scooping, scraping and shaving. The activities were manually assigned and provided with the dataset. Because eating is not a food preparation activity, we remove this activity from the dataset for our experiments below. Example of some activities are shown in figure 2.

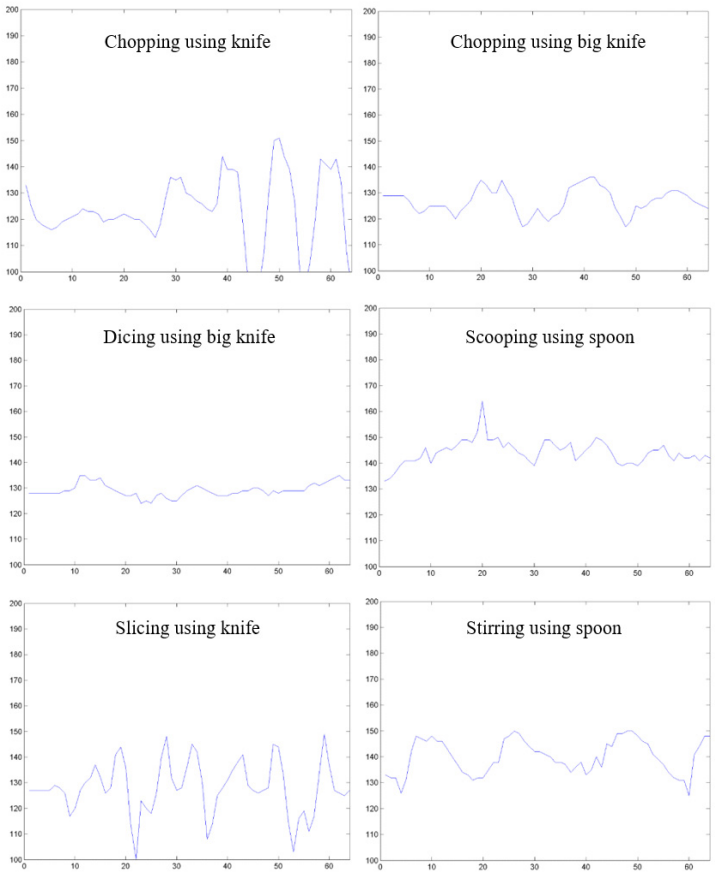

Figure 2: Examples of some food preparation activities in the dataset

Empirical study for sliding window size has been performed and the results shown that the best performance of the proposed algorithm can be achieved with size of 128 value points (3.2 seconds) and an overlap of 50%. Therefore in the experiments we use these value to segment the data streams into frames.

4.2 Experimental Settings

In the experiments, overall accuracy is used as the performance metric to evaluate recognition accuracy of the proposed method and compare with other approaches. It is widely used in activity recognition [2]. Overall accuracy is computed as ratio of number of frames correctly classified over the total number of frames. With n_i is the number of test frames of an activity a_i , $Acc(a_i)$ is the accuracy of activity and M is the total number of test frames for all activities, the formula of overall accuracy is represented as follows:

$$Overall\ Accuracy = \frac{\sum_i n_i \times Acc(a_i)}{M} \quad (1)$$

All methods and settings are evaluated with 10-fold cross-validation protocol. For each fold, 10% of the training set is held out and used as the validation set for tuning slice size and use this parameter for the test set. Empirical experiment for the number of clusters k is performed and the result showed that the best one is 200. This is also the number of motion primitives. We follow this number for our experiments below.

4.3 Experimental Results

In this section, experiments are performed in order to compare recognition accuracy of our method when using features extracted based on bag of features and the other methods using heuristic feature design. The feature set combines several popular statistical features that shown their performance across a variety of activity recognition problems [1, 11]. It consists of mean, variance, standard deviation, correlation, zero crossing rate, mean crossing rate, first order derivative, FFT coefficients in order to afford the complexity of the ten fine-grained food preparation activities in the dataset. In which, mean shows mean of acceleration value and being used to differentiate postures when using utensils for preparing food. Variance shows the level of movement in the signal and it may help to differentiate chopping, dicing, slicing as these activities are similar but different in movement intensity. Correlations between pairs of acceleration axes can differentiate activities related to one direction and multi-direction movements. FFT coefficients are effective in differentiating some repetitive activities so that it can be good for many food preparation activities like chopping, slicing, stirring, etc. The rest of features are important statistical features often used in many time series problem like speech recognition. In addition, we also re-implemented Pham's work [9] using sliding window size of 128 data points and kept the best parameter values as report in his paper. The classifier used in all experiments is C4.5 decision tree.

Table 1 summarizes overall accuracies of three methods. All methods achieve high accuracy (over 82%). The accuracy improvements of our proposed method over the two other methods are significant (nearly 4% compared

Table 1: Comparison of overall accuracies of all three methods

Method	**Overall Accuracy (%)**
Pham's method	82.9
Heuristic feature design	84.7
Our proposed method	88.3

Table 2: Comparison of accuracies of each activities on the experiment dataset for Pham's method, heuristic feature design method and our proposed method

Activity	**Pham's method (%)**	**Heuristic feature design method (%)**	**Our proposed method (%)**
Chopping	85.50	95.39	98.36
Peeling	93.22	90.16	93.44
Slicing	31.97	36.84	39.47
Dicing	26.26	40.54	35.14
Coring	79.60	77.27	86.36
Spreading	50.41	52.63	55.56
Stirring	83.76	82.35	85.85
Scooping	89.75	83.02	94.12
Scraping	63.38	65.45	69.09
Shaving	64.90	65.85	70.73
Average Accuracy	**66.88**	**68.95**	**72.81**

with heuristic feature design and 5.4% compared with Pham's work). Heuristic feature design methods has higher accuracy compared with Pham's work. This result shows that by selecting appropriate features, we can improve the recognition accuracy.

In table 2, detailed results for each class and the average accuracy are represented for all methods. There are 5 activities with high recognition accuracy for all methods including chopping, peeling, coring, stirring, scooping. It is obvious because each of these activities are rather different with other activities. Some activities with low recognition accuracy are slicing, dicing, spreading. These activities are similar in term of signal so that it is hard to differentiate them clearly. Or even in real life, people sometimes can not differentiate between dicing and slicing. When using heuristic feature design, accuracy of some activities like chopping, slicing, dicing, spreading, scraping and shaving increased, especially slicing and dicing. However accuracy of other activities decreased. But for our method based on bag of features, the accuracy of all activities are improved compared to the two other methods, except dicing.

The next experiment is designed to evaluate the influence of different local

Table 3: Comparison of overall accuracy of the two local feature sets

Local feature set	Overall Accuracy (%)
Feature set 1	86.9
Feature set 2 (our proposed features)	88.3

feature sets. The first local feature set consists of mean and variance, which are often used in recent work based on motion primitive based model using bag of features such as [4, 16]. The second feature set used in previous experiments consists of features which are designed to capture local characteristics of food preparation activities. For this experiment, all parameters used in previous experiment are kept. Table 3 summarizes the accuracies achieved when using each feature set on the experiment dataset. The results show that the second feature set provide better accuracy than the first feature set.

5 Conclusion

In this paper, we have presented a feature learning method based on bag of features for fine-grained food preparation activity recognition. Based on bag-of features, general motion primitives has been identified. Then the activity model is built using histogram of those motion primitives. The proposed approach has been evaluated on AK dataset containing complex food preparation activities. The experiment results show that the proposed approach provided substantially higher accuracy than existing approaches for food preparation activity recognition.

A weakness of this approach is it only bases on the distribution of motion primitives and can not keep temporal correlations of the activity streams. In our future work, we will consider other powerful models like HMM or CRF for this direction. Deep learning based techniques are also promising approach for our activity recognition problem.

References

[1] Ling Bao and Stephen S. Intille, *Activity recognition from user-annotated acceleration data,* in Pervasive computing, pages 117. Springer, 2004.

[2] Andreas Bulling, Ulf Blanke, and Bernt Schiele. *A tutorial on human activity recognition using body-worn inertial sensors,* ACM Computing Surveys (CSUR), **46**(3):33, 2014.

[3] Tâm Huynh, Ulf Blanke, and Bernt Schiele, *Scalable recognition of daily activities with wearable sensors,* in Location-and context-awareness, pages 5067. Springer, 2007.

[4] Tâm Huynh, Mario Fritz, and Bernt Schiele, *Discovery of activity patterns using topic models,* in Proceedings of the 10th international conference on Ubiquitous computing, pages 1019. ACM, 2008.

[5] Tâm Huynh and Bernt Schiele, *Analyzing Features for Activity Recognition,* in Proceedings of the 2005 Joint Conference on Smart Objects and Ambient Intelligence: Innovative Context-aware Services: Usages and Technologies, sOc-EUSAI 05, pages 159163, New York, NY, USA, 2005. ACM.

[6] Pranam Kolari, Akshay Java, Tim Finin, Tim Oates, and Anupam Joshi, *Detecting spam blogs: A machine learning approach,* in Proceedings of the National Conference on Artificial Intelligence, volume 21, page 1351. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2006.

[7] Andreas Krause, Daniel P Siewiorek, Asim Smailagic, and Jonny Farringdon, *Unsupervised, dynamic identification of physiological and activity context in wearable computing,* in 2012 16th International Symposium on Wearable Computers, page 88. IEEE Computer Society, 2003.

[8] Jinna Lei, Xiaofeng Ren, and Dieter Fox, *Fine-grained kitchen activity recognition using rgb-d,* in Proceedings of the 2012 ACM Conference on Ubiquitous Computing, pages 208211. ACM, 2012.

[9] Cuong Pham and Patrick Olivier, *Slice&Dice: Recognizing food preparation activities using embedded accelerometers,* Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 5859 LNCS:3443, 2009.

[10] J Ross Quinlan, *C4. 5: programs for machine learning,* Elsevier, 2014.

[11] Nishkam Ravi, Nikhil Dandekar, Preetham Mysore, and Michael L Littman, *Activity recognition from accelerometer data,* in AAAI, volume 5, pages 15411546, 2005.

[12] Marcus Rohrbach, Sikandar Amin, Mykhaylo Andriluka, and Bernt Schiele, *A database for fine grained activity detection of cooking activities,* in Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, pages 11941201. IEEE, 2012.

[13] Sebastian Stein and Sj McKenna, *Combining Embedded Accelerometers with Computer Vision for Recognizing Food Preparation Activities,* Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp 2013), Zurich, Switzerland, pages 110, 2013.

[14] Jason Weston, Samy Bengio, and Nicolas Usunier, *Wsabie: Scaling up to large vocabulary image annotation,* in IJCAI, volume 11, pages 2764 2770, 2011.

[15] Xindong Wu, Vipin Kumar, J Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J McLachlan, Angus Ng, Bing Liu, S Yu Philip, and Others, *Top 10 algorithms in data mining,* Knowledge and Information Systems, 14(1):137, 2008.

[16] Mi Zhang and Alexander A. Sawchuk, *Motion primitive-based human activity recognition using a bag-of-features approach,* in Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium, pages 631640. ACM, 2012.