To appear in Digital Futures 2012 (3rd Annual Digital Economy All Hands Conference), Aberdeen, UK, October 2012

Towards recognizing food preparation activities in situational support systems

Sebastian Stein and Stephen J. McKenna
School of Computing
University of Dundee
Dundee, United Kingdom
{sstein|stephen}@computing.dundee.ac.uk

ABSTRACT

One way in which technology can help people with cognitive impairments to stay independent for longer is through situational support systems that recognize a person's activities and provide help if needed. While food preparation is one of the most important tasks of daily living, such activities are extremely challenging to recognize automatically as they involve a large number of different objects, gestures and cognitive capabilities. This paper introduces an approach to recognizing food preparation activities using statistical machine learning based on visual and accelerometer data. As a first step towards multi-modal activity recognition, an accelerometer localization algorithm has been developed that provides important information for fusing visual and accelerometer data. We discuss a statistical activity model that will be used to guide the construction of datasets for complex activities and give an outlook on the next steps.

Categories and Subject Descriptors

I.5.5 [**Pattern Recognition**]: Applications; I.2.10 [**Vision and Scene Understanding**]: Video Analysis

General Terms

Algorithms, Human Factors, Measurement

1. INTRODUCTION

Development of new assistive technologies gains in importance with the steadily increasing ratio of people needing personal care to those able to provide care. Situational support systems recognize and track a person's activities, identify situations where help is required and provide support by issuing guiding prompts to the user. Although prototype support systems of this kind exist for individual tasks such as hand-washing [1, 2], there are still many challenges to overcome. For example, systems need to be easily adaptable to a particular person's home and cognitive capabilities for easy deployment. Furthermore, they would need to be able to robustly track and guide through a wide range of activities in order to offer substantial benefit to the user.

Activities in the kitchen environment are particularly complex as they involve a large number of different objects, gestures and cognitive capabilities. Additionally, many individuals maintain unique sets of recipes and personal variations of common recipes. These characteristics among others make it hard to automatically recognize and track through food preparation activities.

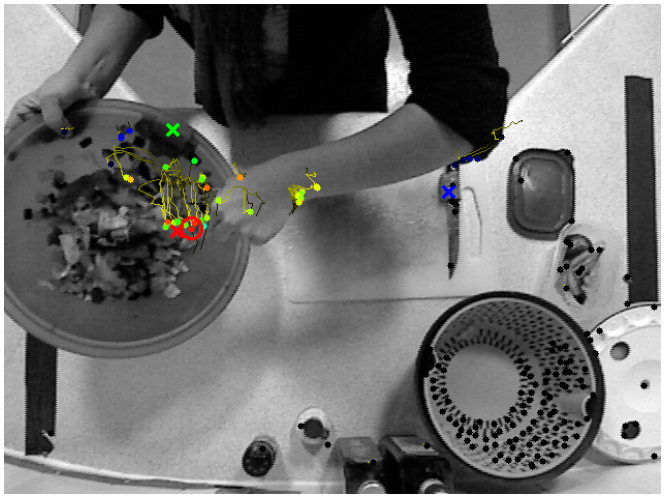

Figure 1: Camera View: Point features (coloured dots) are tracked over time (yellow lines represent trajectories). By measuring similarity of accelerations along these trajectories with accelerometer data (black indicates weakest, and red indicates strongest similarity), the algorithm estimates the accelerometer location (red circle). The red, green and blue crosses mark the ground-truth locations of accelerometers attached to the spoon, the bowl and the knife, respectively.

While previous approaches to activity recognition for situational support make use of either pervasive sensors (e.g., RFID or accelerometers) or visual sensor technology, we propose to combine embedded accelerometers and computer vision techniques to overcome some of the limitations of current systems.

In previous work, Pham et al. [4] developed a method for recognizing food preparation actions such as chopping, peeling, stirring and scooping using accelerometer data from accelerometers embedded in knifes and spoons. Although these motion patterns certainly appear in the context of various recipes, this approach does not recognize the ingredients acted upon (e.g., *what* is being chopped), which is important for tracking progress in a recipe.

We propose to leverage the precise motion data captured by accelerometers and the spatial information that can be

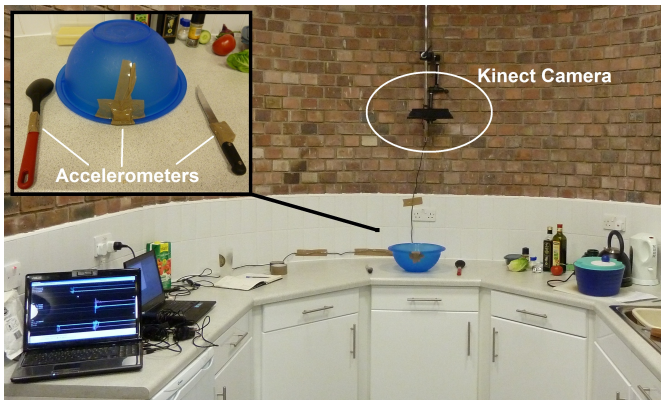

Figure 2: Experimental setup

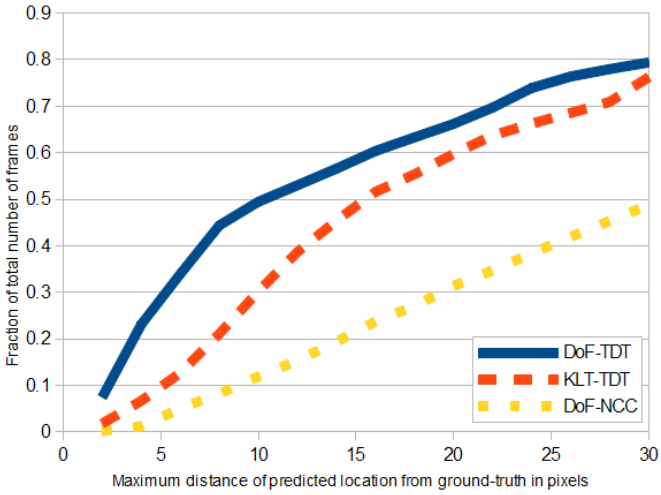

Figure 3: Accelerometer localization accuracy

extracted from visual data by combining these modalities for recognizing complex utensil-ingredient interactions. With this approach we hope to close the gap between recognizing utensil actions and tracking through a complex food preparation task.

2. SENSOR FUSION

Data from different sensor types can be combined at various stages of the recognition pipeline, for example before classification in feature space (early fusion) or after separately recognizing actions from different types of sensor data (late fusion).

Late fusion is easy to implement as an incremental extension of activity recognition algorithms that use either of the different types of sensor data. If recognition results can be interpreted as probability distributions over action classes, classifiers may be combined using simple schemes such as the product-rule or the sum-rule [3].

Early fusion using simple schemes such as concatenating feature-spaces generally does not produce satisfactory recognition performance. Combining data from different sensor types in a more clever way, i.e., exploiting dependencies between data streams and domain knowledge, however, has the potential to significantly outperform late fusion methods.

We identified the localization of an accelerometer in the camera's field of view as a potentially powerful mechanism for fusing vision and accelerometer data. Localizing an accelerometer in the visual field may allow the modeling of activities involving utensil-ingredient interactions by focusing visual attention on the image region that is in spatial proximity to the accelerometer location. Actions such as cutting, peeling and scooping could be combined with information about the object acted upon, defining activities such as peeling an apple.

We have developed a prototype in which accelerometers are localized by comparing accelerometer data with acceleration along point trajectories extracted from video data (see Figure 1). The similarity between individual point trajectories and accelerometer data is measured incrementally by comparing both visual and device acceleration to a threshold with temporal decay. For technical details we refer the interested reader to [5].

To evaluate this method, we collected data of a person preparing a mixed salad. The experimental setup is depicted in Figure 2. Three accelerometers were attached to a knife, a spoon and a bowl. Ingredients did not carry any sensors or tags. A camera was mounted to have a top-down view of the work surface. In order to evaluate the performance of our accelerometer localization algorithm quantitatively, we annotated the locations of the three accelerometers in all frames by hand. Figure 3 shows the fraction of predicted locations within some radius of the ground truth accelerometer location for different algorithm configurations. Method DoF-NCC is the baseline method for comparison using normalized cross-correlation, and KLT-TDT and DOF-TDT are the new threshold method with KLT tracking and dense optical flow, respectively. The method of [5] using dense optical flow worked the best.

3. DATASETS OF COMPLEX ACTIVITIES

For statistical machine learning algorithms it is crucial that the data represent the target domain well. In the context of recognizing simple actions such as *running, jumping, boxing* and *waving*, variation in video data is mainly due to different people performing those actions differently, and in some cases due to changing background clutter. When observing a person performing multi-step activities interacting with a number of different objects, different orderings in which these steps are carried out induce strong variation on the configuration and appearance of objects in the scene. In the context of preparing a mixed salad, for example, the scene looks different after preparing the dressing, depending on whether the ingredients of the salad have been cut and mixed already. In order to build robust activity models for recognition it is convenient to have a *balanced* dataset that contains roughly the same number of examples for all likely task-orderings.

In practice it is costly to acquire annotated video data of a large number of people performing the same multi-step activity. Additionally, the task-orderings the recorded sample population chooses naturally are potentially highly imbalanced. Therefore, we propose to sample task-orderings from a statistical activity model and ask participants to follow the steps of a recipe in orderings generated by the model. An example of a statistical activity model for preparing a mixed salad is illustrated in Figure 4.

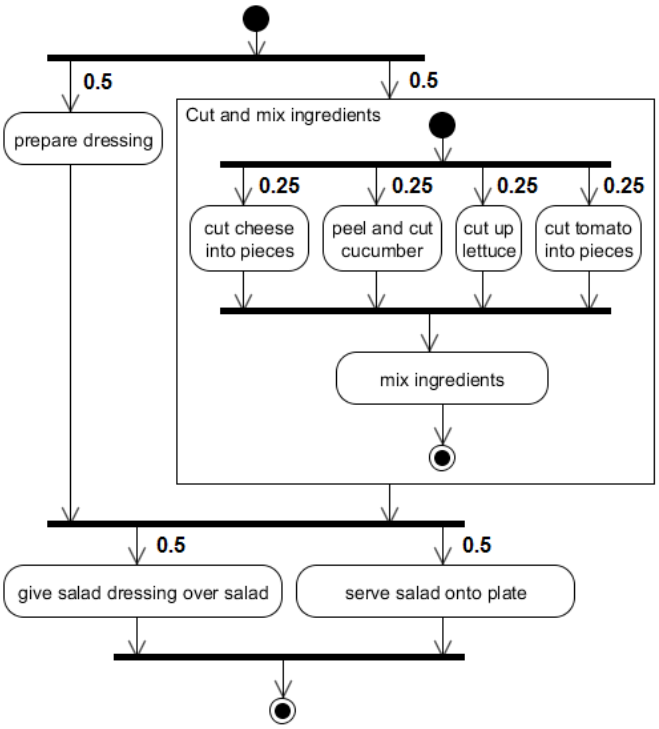

Figure 4: Activity model

The model is based on *Activity Diagrams* used in computational process specification and analysis. Every *choice*-node of the diagram (represented by a horizontal bar with multiple outgoing arcs) is augmented by a probability distribution over all options representing the probability of choosing each option when that choice-node is reached. The probabilities in Figure 4 are set to be uniform to ensure that the dataset generated using task-orderings sampled from this model is balanced.

Using the statistical activity model to generate orderings of tasks reduces the risk of misrepresenting an activity with the idiosyncrasies of a small number of participants. In addition to creating balanced datasets of complex activities, the statistical activity model may potentially serve as a reference structure for tracking through such an activity. In this case probabilities associated with choice-nodes should reflect prior knowledge about the probability of choosing one option over others.

4. DISCUSSION & FUTURE WORK

With the statistical activity model introduced in the previous section we will construct a dataset containing annotated accelerometer and video data of people performing food preparation activities. We will use this dataset to evaluate (i) machine learning algorithms for recognizing food preparation activities based on accelerometer and video data individually, in order to analyze the relative strengths of different sensor types, and (ii) recognition models based on the combined data in order to investigate different sensor fusion techniques for multi-modal activity recognition. In this context we plan to compare late fusion combining classifier results and an early fusion technique that specifically takes the accelerometer location into account to guide visual attention.

The combination of computer vision with embedded sensors could enable recognition of complex interactions and tracking through food preparation tasks, which is essential for providing situational support in the kitchen. While previous work focussed on recognizing utensil actions based on embedded sensors, the approach presented here puts those actions into the context of the recipe and the ingredients acted upon. Although for an implementation with people with dementia further challenges remain, the current work is an important step towards being able to provide support to help people live longer in their homes and assist with tasks of daily living.

5. ACKNOWLEDGMENTS

This research is funded by RCUK Digital Economy Research Hub EP/G066019/1 SIDE: Social Inclusion through the Digital Economy, from December 2010 until May 2014. We would like to thank Patrick Olivier and his group for providing accelerometers and for useful discussions and Vicki Hanson and her group for their feedback.

6. REFERENCES

[1] J. Hoey, T. Ploetz, D. Jackson, A. Monk, C. Pham, and P. Oliver. Rapid specification and automated generation of prompting systems to assist people with dementia. *Pervasive and Mobile Computing*, 7(3):299–318, 2010.

[2] J. Hoey, A. v. Bertoldi, P. Poupart, and A. Mihailidis. Assisting persons with dementia during handwashing using a partially observable markov decision process. In *Proceedings of the International Conference on Computer Vision Systems (ICVS 2010), Bielefeld, Germany*, 2007.

[3] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20:226–239, 1998.

[4] C. Pham, T. Plötz, and P. Oliver. A dynamic time warping approach to real-time activity recognition for food preparation. *Ambient Intelligence, LNCS*, 6439:21–30, 2010.

[5] S. Stein and S. J. McKenna. Accelerometer localization in the view of a stationary camera. In *Proceedings of the 9th Conference on Computer and Robot Vision (CRV'12), Toronto, Ontario, Canada*, pages 109 – 116, 2012.