A Multimodal Benchmark Tool for Automated Eating Behaviour Recognition

V.D. Kakra[1], N.P. van der Aa[1], L.P.J.J. Noldus[1] and O. Amft[2]

[1]Noldus Information Technology, Wageningen, The Netherlands. v.kakra@noldus.nl

[2]Department of Sensor Technology, University of Passau, Passau, Germany.

Abstract

In this paper, we present the multimodal eating behaviour dataset called iEatSet (***i***CareNet Multimodal ***Eat***ing Behaviour Data***Set***). iEatSet shall serve as an algorithm benchmark dataset and aims to facilitate research in automatic dietary monitoring, eating recognition, and activity recognition in general. iEatSet provides multimodal synchronised data streams, including multi-camera vision, inertial motion sensor data, and associated ground truth labelling, recorded from 15 participants over 5 meals in a natural restaurant environment. Recordings included food selection and consumption without scripted protocol, to provide naturalistic behaviour.

Having validated methods and tools to recognize people's eating behaviour could advance research and coaching in applications related to nutrition and dieting. Currently, dieting analysis is done manually and is very tedious. Hence, an automated analysis tool is desired. The current state-of-the-art tools in activity recognition first need to be trained before they can recognize activities. The iEatSet can be particularly useful to benchmark supervised recognition algorithms and serve as reference for unsupervised algorithm analysis.

Introduction

Significant research has been performed in the field of human eating behaviour that includes different aspects like the monitoring of food intake, food selection choices and eating habits of a person. We focus on recognition of a person's food intake activities like the number of bites or sip taken. The counting of number of sips and bites is vital to interpret eating behaviour which can be used to coach the patients and counteract eating disorders that include obesity and anorexia nervosa.

Currently, studies that analyse food intake activities have human observers who use tools like The Observer XT (www.noldus.com /observer), a software package for the collection, analysis and presentation of observational data. In other studies, participants are requested to manually note down details about his/her food consumption. This analysis can be made more efficient if the recognition procedure is automated and standardized.

There have been different approaches developed and sensors used by researchers to automate the recognition of food intake activities as part of daily life activities. For example, Cheng et al. (2011) recorded senior citizens in their homes with a colour camera to recognize different eating activities and walking activities. These video recordings are used to assist in monitoring the health and well-being of the senior citizens. Amft et al. (2005) monitored a person's diet with the help of on-body inertial sensors to recognize eating behaviour for food intake. Stein et al. (2009) used a multimodal approach to recognize food preparation activities. They used inertial sensors along with cameras to detect activities, which improved the activity recognition results. Most of these existing methods for activity recognition use a machine learning approach, where they first train the classifier on training data and subsequently validate the method on test data. Both, training and evaluation of algorithms can benefit from a publicly available benchmark dataset for automatic eating activity recognition.

There are several activity recognition datasets available that use just video cameras or a multimodal sensor setup and some of them also record eating activities among other activities. For example, the Hollywood2 dataset by Marszałek et al. (2009) is a vision-based dataset containing 'eating' actions among 12 actions collected from different Hollywood movies, while the Senior Activity Recognition Dataset (SAR) by Cheng et al. (2011) is in the real world focusing mainly on recognition of eating classes and walking classes of senior persons. These datasets

Proceedings of Measuring Behavior 2014, (Wageningen, The Netherlands, August 27-29, 2014).
Editors: A.J. Spink, L.W.S. Loijens, M. Woloszynowska-Fraser & L.P.J.J. Noldus. www.measuringbehavior.org

focus on detecting different activities rather than recognizing different eating activities. There are also multimodal datasets available for activity recognition by sensor fusion. One such multimodal dataset, which is closely related to what we want to record, is the 50 Salads Dataset by Stein et al. (2013) that records people preparing food. This dataset uses the Kinect v2 sensor (http://www.microsoft.com/en-us/kinectforwindows/) to record RGB-D data from a top view and has accelerometers attached to the objects like the knife, mixing spoon and other objects used for preparing the food. We use the same sensors but our focus is on food intake activity recognition and not on food preparation. The food intake activities involve movements of a hand towards the mouth, which is best captured from the front view with the upper body of the person completely visible. In this dataset, recording is done from the top view with only the hands of the participants visible so that is why the 50 Salads Dataset cannot be used and we recorded the iEatSet for the food intake activity recognition.

The iEatSet provides a dataset with 15 participants, recorded 5 times having different meals for 5 days and ground truth labelling is provided to use for both testing and validation of future algorithm development for eating behaviour as well as other activity recognition algorithms. The dataset includes uncompressed data from 3 RGB cameras that record the person from the front, side and top views, the calibrated data from the 4 inertial sensors that are attached to both the upper and lower arms and the data from the Kinect v2 depth sensor from the front view. All the data is time synchronized and the time synchronization is also provided.

iEatSet (*i*CareNet *Eat*ing Behaviour Data*Set*)

Choice of Sensors

Food intake activities include actions like taking a bite or sip that involves movement of the hand from the plate/table towards the mouth. This movement can best be captured through vision by capturing the upper body of the person with the hand and face of the person visible. That is why we recorded the RGB data from the front view. Also, the distance between either the food or cutlery or food containers and the mouth can be used to recognize whether the person is eating or not. This information can be captured with a depth sensor. The top view cannot be used here because the person's mouth isn't visible and if the person leans over to consume food, only the head of the person would be visible while most other things would be occluded. That's why we choose to use the Kinect v2 to capture the RGB-D data from the front view. An additional Kinect v2 sensor cannot be used to capture the depth data from the side or top view because this would cause interference while recording.

One big disadvantage of using cameras is that there can be occlusion at any given time during the recording. This occlusion problem can be tackled by recording from different field of views using multiple cameras. The person can be captured eating from the side view and the top view. The movement of hands towards the mouth can be captured best from the left side for a right-handed person and on the right side for a left-handed person. This way the camera can get a clear view of the hands, cutlery and the table and the eating activities can be recognized from the side view. The top view can be used to observe the food items on the plate and the table as well as the movement of the hands when a person eats. This information can be used to recognize food intake activities. The top view also provides researchers a way to investigate food identification. That is why we used 2 IP cameras to capture RGB data from the side view and top view that includes the table and participant while eating. Each of the three camera views can be used separately to recognize eating activities or a multi-camera view approach can be used.


Figure 1. Frontal view of person eating.

On-body sensors like the inertial sensors can also be used to recognize the eating activities. There has already been research in this area where the Inertial Measurement Units (IMUs) or inertial sensors have been used to recognize eating activities. Amft *et al.* (2005) demonstrated inertial body-worn sensors can be used to detect eating and drinking gestures from other gestures where the sensors were placed on both the lower and upper arms of the person. When a person eats, he/she holds the food or cutlery in a specific way which is reflected in a specific orientation of the hand which can best be

captured by a gyroscope. Then the person moves hands with the food towards the face and the speed of the hand can be captured by the accelerometer. These features can then be used to recognize eating activities. The orientation and acceleration can be captured when the sensors are placed on the wrists or lower arms and upper arms of the participant.

Sensors Description

Kinect v2: The Kinect v2 records the RGB data with a resolution of 1920×1080 pixels @30fps and the depth sensor with a resolution of 512×424 pixels, 13-bit depth. The RGB data and depth map are synchronized internally by the Kinect v2. The Kinect v2 recorded the person eating from the front view as shown in Figure 1.

RGB camera: Axis M1054 IP cameras were used to record the person from the side and top views with a resolution of 1280×800 pixels @25fps. The calibration between all the three cameras was done by using a checkerboard pattern and the intrinsic and extrinsic parameters are made available. One camera was placed above the table and little in front of the person so that the plate was visible at all times even if the person leans forward to eat. The other camera was placed on the left side if the person was right-handed and on right side if the person was left-handed.

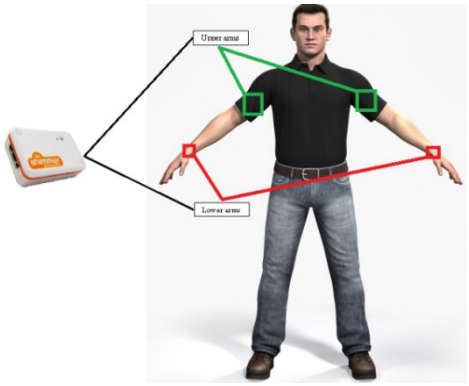

Figure 2. Shimmer3 sensor placement.

Inertial Sensors: Shimmer3 (www.shimmersensing.com /shop/shimmer3) is a state-of-the-art wearable IMU sensor which has an in-built accelerometer, gyroscope and magnetometer. The size of Shimmer3 is similar to that of a wrist watch so it can easily be mounted. Data can be recorded on the SD card slotted on the Shimmer3 sensor or streamed directly to the machine via Bluetooth. Data was recorded on the SD card of each Shimmer3 sensor. The sensors were mounted on the upper and lower arms of the person. The Shimmer data is 48-bit data from accelerometer, gyroscope and magnetometer @51.2Hz. The 4 devices are synchronized using the synchronization software (Multi Shimmer Sync for SD v1.2) where a master-slave pairing is established between the sensors. The Shimmer3 sensor is shown in Figure 2.

Sensor Network Synchronization

For synchronization between all sensors, the participant was asked to clap his hands twice: before the start and at the end of eating meals. This was because the clap is clearly visible through vision (Kinect v2 & RGB camera) and that would be used as a reference to start and end the recordings. A clap represents a sudden movement and acceleration of the hands which is clearly visible in the acceleration data from the Shimmer3 sensors. The synchronization is achieved using the timestamps generated by each sensor.

Study Design

The iEatSet consists of 15 participants where each participant was recorded when he or she was having lunch for 5 days. We expected 8-10 eating instances and 6-8 drinking instances per meal from each person. That was why we chose to record the same person for 5 days to have enough eating and drinking instances for training and testing for a single person's behaviour recognition. We chose to record 15 people because this gave us enough instances for training and test data for inter-person recognition.

Each day the participant was served different food like sandwiches, hot meal, soup and yoghurt and different types of drinks like soft drink, orange juice, tea and coffee to gather as many instances and variety of eating gestures. The participant was asked to eat either with their hands or cutlery depending on the food served to ensure all the types of eating gestures were covered. Also, different drinks were served in a glass, a bottle or a cup to ensure each day different types of drinking activities were recorded.

Eating Gesture Lexicon & Ground Truth Annotation

The eating activities were annotated according to the classes provided by Koenderink *et al.* (2014). The annotation is two-layered: the first layer broadly classifies all the eating activities into 6 main classes and the second layer consists of 22 eating activities that are further grouped under each of these 6 classes. We modified this lexicon where we didn't annotate the gestures that happen inside the body and couldn't be detected by the sensors. With the two layers, researchers can investigate methods either for inter-class or intra-class eating activity recognition or a combination of them for both layers. The 6 main classes and the 22 sub-classes classified under these 6 classes are mentioned in Table 1.

Manual annotation is subject to human errors which could lead to the annotation being diverged from what it should be in multiple ways as described by Mathet *et al.* (2012). Each manually annotated element can diverge either by being annotated in the wrong location, the category to which it is annotated is not correct, is a false positive or a false negative annotation. In order to keep these errors to a minimum, we used a methodology whereby all experts label a share of the data and 20% in parallel. For the parallel annotated data, we provide iterator reliability metrics and timing coherence metrics. The Observer XT is used as the event logging software.

Layer 1	Non-eating gestures	Prepare food gestures	Prepare consumption gestures	Consume gestures	Finalize consumption gestures	Other
Layer 2	Play (moving food, plate); Wipe clothes, hand, mouth or table; Pick up cleaning aid; Stack waste or tableware;	Cut; Open package; Pick up package; Place spread on bread; Pour; Put down package; Stir;	Pick up cutlery; Move to mouth; Pick up food or drink;	Take bite; Take sip;	Move away from mouth; Put down cutlery; Put down food;	Lick finger, cutlery or lips; Pick teeth; Express with hands;

Table 1. The two layers of the eating gesture lexicon to be used in the annotation

Discussion

This dataset is made available to the research community. It contains (1) uncompressed and synchronised RGB videos of the recordings from the IP cameras and the Kinect, (2) the calibration data of each camera including the intrinsic and extrinsic parameters, (3) 13-bit depth data from the Kinect, (4) the raw, calibrated and synchronised 48-bit data from all the 4 IMUs, (5) the labelled annotations, (6) the timestamps generated by all the sensors and (7) accompanying software to read the data. With this benchmark, we hope to advance research in eating behaviour recognition and activity recognition in general.

Acknowledgement

This project was supported by the EU FP7 Marie Curie Initial Training Network iCareNet under grant number 264738.

References

1. Amft, O., Junker, H., & Troster, G. (2005). Detection of eating and drinking arm gestures using inertial body-worn sensors. *Wearable Computers, 2005. Proceedings. Ninth IEEE International Symposium on* (pp. 160-163), October 2005.

2. Cheng, H., Liu, Z. Zhao, Y. & Ye, G. (2011). Real world activity summary for senior home monitoring. 2011 IEEE International Conference on Multimedia & Expo. (ICME), vol., no., pp.1-4, 11-15 July 2011.

3. Koenderink, N.J.J.P., Top, J.L., van Doorn, A. (2014). Unravelling the language of eating, Workshop on Smart Technology for Cooking and Eating Activities, CEA2014.

4. Marszałek, M., Laptev, I. & Schmid, C. (2009). Actions in context. IEEE Conference on *Computer Vision and Pattern Recognition, pp. 2929-2936, 2009.*

5. Mathet, Y., Widlöcher, A., Fort, K., François, C., Galibert, O., Grouin, C., & Zweigenbaum, P. (2012). Manual corpus annotation: Giving meaning to the evaluation metrics. *Proceedings of the International Conference on Computational Linguistics (COLING 2012)*, pp. 809-818, December 2012.

6. Stein, S. & McKenna, S.J. (2009). Combining Embedded Accelerometers with Computer Vision for Recognizing Food Preparation Activities. *The 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp 2013), Zurich, Switzerland, 2013.*