The Case for Smartwatch-based Diet Monitoring

Sougata Sen, Vigneshwaran Subbaraju, Archan Misra, Rajesh Krishna Balan, Youngki Lee
Singapore Management University
{sougata.sen.2012,vigneshwaran,archanm,rajesh,youngki}@smu.edu.sg

***Abstract*—We explore the use of gesture recognition on a wrist-worn smartwatch as an enabler of an automated eating activity (and diet monitoring) system. We show, using small-scale user studies, how it is possible to use the accelerometer and gyroscope data from a smartwatch to accurately separate eating episodes from similar non-eating activities, and to additionally identify the mode of eating (i.e., using a spoon, bare hands or chopsticks). Additionally, we investigate the likelihood of automatically triggering the smartwatch's camera to capture clear images of the food being consumed, for possible offline analysis to identify what (and how much) the user is eating. Our results show both the promise and challenges of this vision: while opportune moments for capturing such useful images almost always exist in an eating episode, significant further work is needed to both (a) correctly identify the appropriate instant when the camera should be triggered and (b) reliably identify the type of food via automated analyses of such images.**

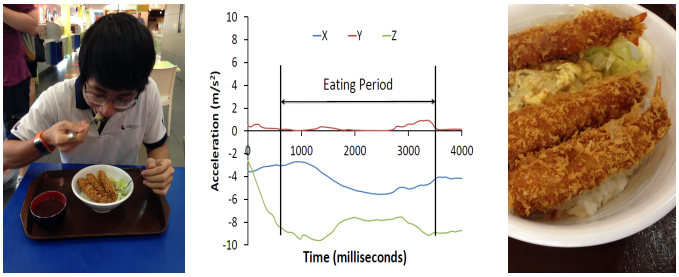
I. Introduction

As eating is one of the most common daily lifestyle activities, the use of mobile or wearable sensors for capturing eating-related behavior has been studied in the past, along two distinct dimensions: (a) to recognize the *gesture of eating* [3], [2] (e.g., using inertial sensors) or (b) *to identify dietary intake* [7] (e.g., using images captured by a smartphone). Motivated by the recent commercial success of smartwatches, this paper explores how far the multiple sensors on a wrist-worn, off-the-shelf smartwatch (worn on the hand used for eating) can help to automatically infer *both* such gestural and dietary context.

More specifically, we envision a novel automated (with very little human effort) food diary system that uses off-the-shelf smartwatches as follows:

1 It first uses the inertial sensors (e.g., the accelerometer and the gyroscope) on the smartwatch to identify an "eating" (hand-to-mouth) gesture, as well as the series of all such gestures that define a complete eating *episode*.
2 Additionally, at appropriate time instants during the episode, the camera sensor on the smartwatch is activated to click a picture of the plate's content, and then (ideally) uses offline image analysis techniques to automatically identify the type and the quantity of the food that is consumed.

This work is supported partially by Singapore Ministry of Education Academic Research Fund Tier 2 under research grant MOE2011-T2-1001, and partially by the National Research Foundation, Prime Minister's Office, Singapore, under the IDM Futures Funding Initiative and the International Research Centre @ Singapore Funding Initiative, and administered by the Interactive & Digital Media Program Office, Media Development Authority. All findings and recommendations are those of the authors and do not necessarily reflect the views of the granting agency, or Singapore Management University.



(a) (b) (c)

Fig. 1: Automated Diet Monitoring: (a) A Smartwatch-wearing User; (b) Accelerometer Pattern for Eating Gesture; (c) Representative Image from Smartwatch Camera

Figure 1 illustrates the concept of the system: Figure 1a shows an individual eating while wearing a smartwatch, Figure 1b shows the corresponding accelerometer trace while consuming one spoonful, while Figure 1c shows an image of the plate captured by the smart-watch embedded camera.

Our work is conceptually similar to other recent research efforts on smartwatch-based detection of daily lifestyle activities (e.g., the RisQ prototype [4] that captures smoking episodes). However, building a practical automated food diary system presents several **unique challenges**:

- *Diversity in Eating Modes:* As compared to activities such as smoking, eating is inherently a more diverse activity—for example, in multi-cultural Singapore, different consumers eat using spoons/forks, with their hands or using chopsticks. The gesture recognition system must thus successfully accommodate each of the resulting distinct set of wrist and hand gestures.
- *Suitable Instant for Image Capture:* Since the plate is not visible during most of the time periods during an episode, we propose to employ a *gesture-triggered* approach, where the camera is activated only at those points in a gesture's trajectory where the system believes that it has the best chance of obtaining a "clear & useful" image of the plate's contents. As shown in [9], even if we have some false triggers in turning *on* the camera, the energy consumption will still be significantly lower as compared to continuous sensing (image capturing). Hence, we must figure out a way to determine *when* the camera must be triggered.
- *Derive Food and Intake Quantity Estimates:* We will need to evaluate the degree to which automatic image analysis can help identify the food type and the quantity consumed, and if needed, we must figure out alternative semi-automated or sensor-augmented ways to enhance the

978-1-4799-8425-1/15/$31.00 ©2015 IEEE

accuracy of such identification.

For this initial exploratory paper, we focus principally on the first two challenges–related to recognition of eating gestures and the feasibility of image capture, with only a very preliminary exploration of the feasibility of automated image analysis. More specifically, we make the following **key contributions**:

- We use a real-world corpus of eating gestures from 6 individuals to establish the distinctive sensor features related to eating, and thus develop a smartwatch-based classifier for detecting eating gestures. While the classification accuracy for detecting individual gestural instances is 92%, we show how a higher-layer outlier elimination technique can lead to improved performance–enabling us to identify individual *eating episodes* with 97% accuracy and the number of bites/spoonful taken with an average error of 17%.
- We show how three common modes of eating {eating with hand, eating using spoon, eating with chopsticks} give rise to distinct gestural patterns, and that gyroscope-generated wrist rotation features can distinguish between gestures of these three modes with a precision of approx. 85.51% (by a person-independent model) and over 90% (by a person-specific model). Moreover, over an entire episode, the mode of eating can be identified with 100% accuracy.
- We show that, for many eating patterns, it is possible for the strap-mounted camera to obtain a clear and usable image of the food content on the plate: approx. 79.05% of the individual gesture instances contain such an image, with the existence of such an image guaranteed at least once for 85.71% of the eating episodes studied.

II. Related Work

Mobile & wearable sensors have been explored for tracking two distinct aspects related to eating: (a) detecting eating episodes (when, and how many spoonful, does a person eat?) and (b) diet habits (what is a person eating?). Early and well known examples of detection of eating episodes include the use of a wearable ear-mounted microphone to detect chewing sounds [1], and the use of inertial sensors [2] to detect eating related gestures. More recently, the Bodybeat system [6] utilized a microphone attached to the neck of the user to identify various body sounds, including eating-related ones (such as swallowing sound). An alternative approach for monitoring the meal intake was explored in the bite-counter work [3], where a watch-like wrist-worn device counts the number of bites the user takes (the number of distinct hand-to-mouth gestures). While the detection of eating episode is quite accurate, this work also demonstrated that the portion of food consumed ('when', and 'how much') does not necessarily strongly correlate to the number of calories ('what') consumed.

Smartphone-based approaches for monitoring the dietary intake of individuals include the DietSense prototype [7], which proposed an automated food diary based on images captured by a mobile device. In this approach, the user is assumed to hang a mobile phone around her neck during eating, with the smartphone clicking images periodically to capture potential "food images" (as well as other spurious

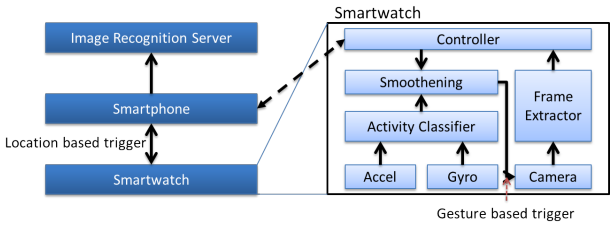


Fig. 2: Architecture of the system

images) that the user could retrieve later to annotate her diet. Besides requiring user intervention to ensure that the phone is properly placed during eating episodes, the continuous capture of images (without any gesture-driven triggering) can lead to high energy overheads.

III. High-Level Architecture

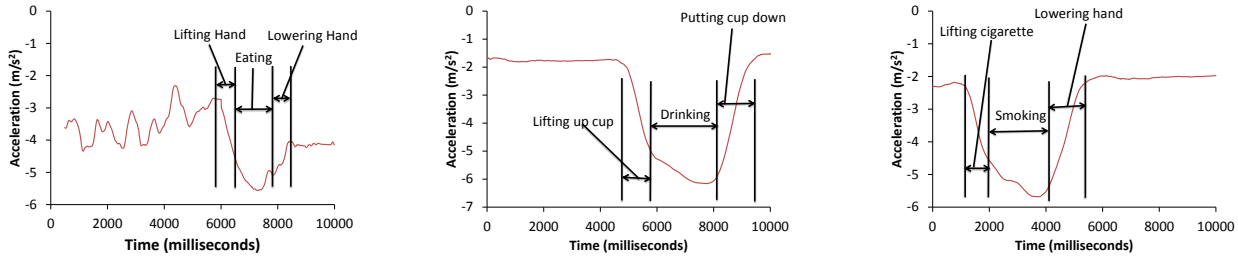
Before delving into the details of the specifics of gesture recognition and context-triggered image capture, we outline (see Figure 2) the architecture of our proposed system, comprising of three distinct components:

- The *smartphone* is responsible for activating the gesture recognition component on the smartwatch as well as acting as the gateway between the smartwatch and the backend. To conserve energy on the smartwatch, the gesture recognition component is activated only when eating is *feasible* for the user–e.g., if the user's location is in the foodcourt. To allow automated estimation of the diet content, the smartphone also relays back images (captured by the smartwatch) to the backend server.
- The *smartwatch* is responsible for recognizing the eating-related hand/wrist gestures and for capturing the appropriate images that helps in identifying the food being consumed. In addition, the smartwatch not only transfers (via the smartphone) the captured images but also summary statistics about the eating episodes (e.g., if the user was eating with chopsticks vs. spoons, the number of mouthfuls consumed, etc.)
- The *backend server* is responsible for both storing the smartwatch-generated data and for identifying the food item (and the portion size) from the uploaded images. identifying images. At present, it is an open question whether such recognition is done automatically by recognition software (e.g., Overfeat [8]) or requires human intervention.

In the rest of this paper, we focus principally on two issues: (a) accurately classifying eating gestures & episodes and (b) ascertaining whether the smartwatch-based camera can feasibly capture useful images of the food content.

IV. Eating Activity Detection

We first investigate the use of inertial sensors on the smartwatch to reliably (a) identify eating gestures, and thus episodes of eating (an episode covers the entire duration of a eating activity and consists of multiple individual hand-to-mouth "eating" gestures along with non-eating gestures), (b) identify the mode of eating, and (c) the number of such hand-to-mouth gestures (which can, arguably, provide an estimate



(a) Eating (b) Drinking Hot Beverage (c) Smoking

Fig. 3: Variation of X-axis of accelerometer (smoothened for 500ms) for various activities

of the quantity of food consumed). To study these properties, we recruited 6 subjects, who performed 21 total episodes (14 eating and 7 non-eating) which were captured using a custom application running on the smartwatch. We utilized 12 out of the 14 eating episodes that, over the episode's entire duration, had at least one suitable image of the plate. A total of 191 eating gesture instances were extracted from the 12 eating episodes and we utilized 151 of them where the food that a person is consuming was visible at least once during the gesture. For the 2 episodes where the food is not at all visible during the eating gesture, our current approach will not be effective. We thus decided not to pursue investigating those episodes at present, but will use them in our future studies where we will utilize other context based triggers from the smartphone to identify eating behavior.

A. *Detecting eating gestures*

To detect an eating activity, we captured the accelerometer and the gyroscope sensor data from the smart-watch. To be practically useful, the gesture recognition model needs to distinguish between eating and non-eating hand movements which are similar to eating (such as smoking, drinking tea, washing one's face, shaving, applying makeup etc.). Fig 3 shows the plot for the variation of the x-axis reading of the accelerometer (sampled at 100 Hz, and smoothed using a 50-sample running average) for a set of eating and representative non-eating activities. As the figures show, an eating gesture is quite distinct from the other gestures. In case of the eating gesture, a person holds his hand to the mouth for a shorter duration than 'smoking' or 'drinking a hot beverage', and longer than 'splashing water on the face' (not shown in the figure).

Past work (e.g., [5]) has used accelerometer-based features to identify a set of such gestural activities. Based on our inspection of the sensor patterns for eating vs. non-eating activities, we utilize accelerometer features (listed in Table I). We collected various activities which involved hand-to-mouth gestures from 3 different participants and extracted features from 5-second long frames. Utilizing the $F1$ subset of features (see Table I) we built a supervised model, using the J48 decision tree classifier in Weka. On applying this classifier to our collected data, we obtain a classification accuracy of approx. 92%. Since it is highly unlikely for an user to perform the eating gesture simultaneously with the other gestures (Smoking, Washing face etc.), we performed a smoothing of the predicted label over a time window: if the class label predicted for the current frame is different from the labels for the prior and successive two frames (e.g., 5 continuous frames have been predicted as Eating, Eating, Non-Eating, Eating, Eating$\cdots$), it is re-labeled with the label of the preceding frame. With such smoothing, the accuracy increases to 97%.

	Accuracy (Eating vs. Non-Eating)	Eating mode accuracy (per-instance)	Eating mode Accuracy (Episode-level)
Personalized	NA	90.74%	100%
Non-Personalized	97%	85.51%	100%

TABLE II: Accuracy of Identifying Eating activity and Mode of Eating at instance level (and episode level)

B. *Identifying Eating Mode*

We next focus on identifying the *mode* of eating–namely, whether the person was using (i) their hands (ii) chopsticks, or (iii) spoon. Identifying the mode of eating can be useful in eliminating identifying certain food items. For example, it is highly unlikely that a person will use chopsticks to consume soup. From Figure 4, it is easy to see that the gyroscope values for *roll* and *pitch* (corresponding to the rotation of the wrist) are quite distinct for each of these three modes.

To identify the mode of eating, we performed a two level classification: (a) Level 1: we identified whether an instance in the episode is an eating instance or not; (b) Level 2: subsequently, all instances identified as eating were run through a 3-ary classifier to identify the mode of eating (spoon, chopstick or bare hand). Features used in building this classifier included the Features $F1$ and $F2$. A leave-one-episode-out cross validation was performed to compute the classification accuracy at level 1. However due to the limitation of number of episodes corresponding to each eating type, for level 2 we performed a 10 fold cross-validation. With this two-level hierarchical classifier, the accuracy of identifying the modality of eating was 85.51% for a non-personalized model and 90.74% for a personalized model (tested only on one user who had eating episodes for each of the 3 modes). Table II summarizes the accuracy of identifying the modality of eating.

Moreover, as none of our participants switched modes (switching between chopstick, hand and spoon) *within* a single eating episode, we applied a dominant set labeling technique

Feature	Count	Description	Feature Set
Accel_mean	3	The mean of the 3 axes of accelerometer for 50 frames	$F1$ and $F2$
Gyro_mean	3	The mean of the 3 axes of gyroscope for 50 frames	$F2$
Accel_variance	3	The variance of the 3 axes of accelerometer for 50 frames	$F1$ and $F2$
Gyro_variance	3	The variance of the 3 axes of gyroscope for 50 frames	$F2$
Accel_covariance	3	The co-variance between the 3 axes of accelerometer for 50 frames	$F1$ and $F2$
Gyro_covariance	3	The co-variance of the 3 axes of gyroscope for 50 frames	$F2$
Accel_correlation	3	The correlation of each of the three axes of the accelerometer with the other two axes	$F1$ and $F2$
Gyro_correlation	3	The correlation of each of the three axes of the gyroscope with the other two axes	$F2$
Duration_of_gesture	1	The time duration of the *hand-to-mouth* gesture	$F1$
Axes_LinearToRotation	3	The ratio between the linear and rotational motion for the three axes	$F2$
LinearToRotation	1	The ratio of sum of three axes of linear and rotational motion	$F2$
$n-ContinuousEating$	1	Boolean value denoting if the previous n instances were classified as eating	$F2$

TABLE I: Features Extracted and Used for activity/modality classification and point of interest identification

to identify a single eating mode over the entire episode. With this approach, we were able to correctly identify the eating mode for each of our 12 episodes. However, in a real world setting, a full meal might involve multiple eating modalities (using soup spoon for the soup, the chopstick for noodles and hand for the dessert). Instead of having a dominant set labeling for the entire episode, the episode can be divided into frames and the dominant set labeling can be applied at frame level.

C. Estimating the number of hand-to-mouth actions

We finally investigate the ability to accurately count, during a single episode, the total number of hand-to-mouth gestures (as this number may provide an estimate of the quantity of food consumed). Empirically, we found that a complete eating cycle (hand ascending, putting food in mouth and hand descending) lasts for an average of 2.36 seconds with a standard deviation of 0.6 seconds. An initial strategy of dividing the duration of eating gestures during the entire episode by this mean gesture duration(2.36 second) turned out to have a low accuracy, as each individual was observed to have a different eating speed (which itself can vary during the course of the meal).

Instead, we took the individual eating instances and then applied an additional smoothing technique to eliminate spurious/noisy elements in this sequence. Two heuristics were applied in the smoothing: (1) if a *hand-to-mouth* gesture started at time T_0, the next one could not start within 2500 milliseconds of this gesture and, (2) within a block of 1000 milliseconds, a person cannot have more than 2 state transitions. Using the above two heuristics, we find that, on average, while each eating episode had a total of 24 hand-to-mouth gestures (the ground truth), our classifier estimated a total of 29 such gestures. In other words, we currently tend to overestimate the total count by approx. 17% (the worst case error for our episodes was 26%), falsely classifying a set of non-eating gestures, that occur during an eating episode (such as a person raising her hand but without a spoon), as 'eating'. Note that using the gyroscope data to more microscopically disambiguate eating gestures may improve the accuracy of this estimate.

V. Image-based Food Identification

The second stage of our envisioned system involves the capture and analysis of images (using the smartwatch's camera) of the food being consumed. As mentioned before, we need to address two separate challenges: (i) determine an opportune time when an image should be captured (that provides clear visibility of the food item) and then, (ii) apply automated analysis to identify the food being consumed.

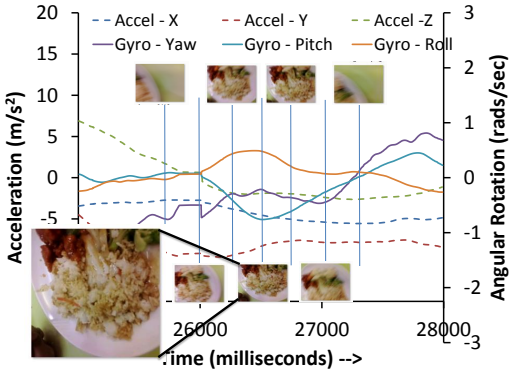


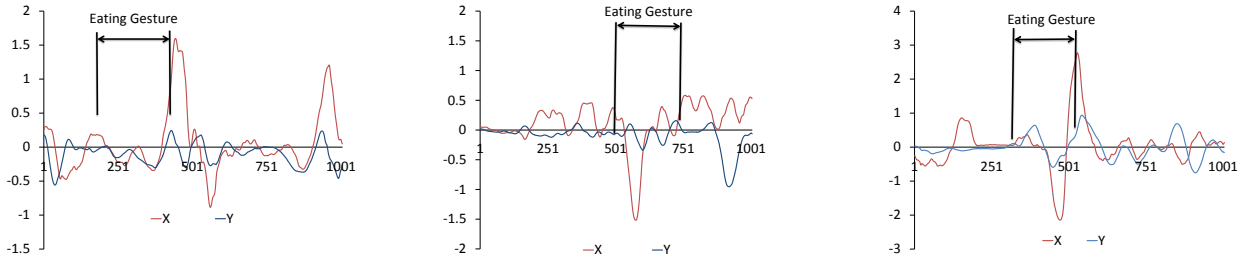
Fig. 5: Variation of image quality with variation of 50 frame smoothed accelerometer and gyroscope data

A. Capturing the Image

To first understand the possibility of using an image-based approach for capturing the type of food, we had our users perform their regular eating activities, but with the camera (which is located on the top strap of the smartwatch, just above the 12 o' clock mark) on the smartwatch set to capture images (at 16 fps), while the user performed the eating activity. Based on the manual inspection of the captured data, we found that *85.7%* episodes had at least one image that provided a clear and focused picture of the food being consumed. More microscopically, on analyzing each individual instance (2 second frames) of each eating gesture, we found that 79.05% of such *hand-to-mouth* gestures had at least one clear image that a human expert could use to identify the food item.

The challenge, of course, is to determine the *right time instant* when the camera should be triggered to obtain a useful image. Based on analyses of our user traces, we found that (see Figure 1c for an illustrative example), for the Samsung Galaxy Gear™ smartwatch, the camera usually points towards the plate at the time instant when the person has her hand near her mouth (at the zenith of the hand-to-mouth gesture). However, the exact moment to capture image varies based on the eating mode (as the orientation of the wrist is different for different eating modes).

We plotted the temporal variation of the accelerometer and gyroscope values for the eating gestures to see if there is any indicative pattern which might suggest the correct instance to capture the picture. Figure 5 shows the sensor variation of one such hand-to-mouth instance from an activity and corresponding images captured (the best captured image has



(a) Eating with spoon (b) Eating with hand (c) Eating with chopstick

Fig. 4: Variation of the roll and pitch of gyroscope for various modalities of eating

been highlighted in the figure). However, no immediate feature for identifying such an opportune moment is discernible.

Subsequently, we have plotted the probability of a captured image being 'useful' (i.e., provides a clear view of the food content) as a function of the time when the image was captured, relative to the overall duration of the gesture. Figure 6 shows the plot of this probability as a function of the time, with the time being expressed as a percentage of the overall gesture duration (the duration being recorded from when the food was picked up until the hand returned to the same position). We can see that the probability of getting a useful image is highest around the mid-point of the gesture (when the hand is at its zenith).

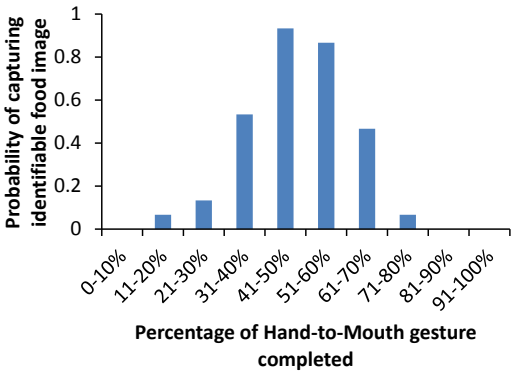


Fig. 6: Probability of 'useful' image capture vs. fraction of the eating gesture completed

B. Automatic Food Identification

As a preliminary exploration of the possibility of automatic image classification, we used the OverFeat image recognition software [8] to try and classify the food images captured by the smartwatch. OverFeat is a classifier that returns a likelihood value for different possible items. For our system, we found that when we supplied the best image in an episode to OverFeat, only 10% of the images were classified correctly and with a confidence value higher than 0.5. These images were easily identifiable by a human being though. Initial evidence suggests two possible reasons for the less-than-satisfactory performance: (i) Most of the food items consumed by our participants consisted of more than 1 item in the plate (rice with veggies, rice with meat, bread with veggies etc.). OverFeat performed poorly with such mixed images; its best performance was recorded when only a single food item (a bowl of soup) was present; (ii) The OverFeat corpus was not specifically restricted to food items. Thus, for example, it sometimes identifies images of noodles as earthworms! Clearly, further investigations into the use of an appropriate corpus of food images is required to better identify the food items. Also, modifying the image segmentation approach as well as tuning parameters such as the zoom of image, the focus etc can be done to capture images which might be more suitable for the image recognition software.

VI. Discussion and Future Work

The experiments above indicate that (a) the detection of eating activity can be reliably achieved using a smartwatch and (b) that, at certain points in a person's eating gesture, the smartwatch camera can provide useful and un-occluded view of the food content. However, the development of a full-fledged diet monitoring system has several additional issues and aspects that must be considered.

Adaptation to Changing Watch/Camera Position: The experiments conducted here required the participants to wear the watch on their preferred eating hand, and were conducted using only a single smartwatch model. In reality, during various episodes, an individual may have the watch in slightly different positions on the wrist (e.g., a loose-fitting strap will cause the watch to rotate around the wrist), resulting in slightly different gestural trajectories and slightly different optimal time instants for image capture. Moreover, different watch models may have the camera in slightly different positions. All of these artifacts suggest that the process for determining the best instant for image capture needs to be *adaptive*. One possibility is to apply an intra-episode adaptation strategy, where the image is first captured and then analyzed (e.g., using standard edge detection algorithms) to see if the plate (or food item) is visible. If not, then the time instant for image capture can be adjusted (i.e., moved earlier or later along the gesture trajectory) until an appropriate capture instant is discovered.

Diversity in Eating Styles: Our limited user studies show the existence of two predominant eating *styles* across individuals: (i) one where the user lifts her spoon to the mouth to consume the food, and (ii) one, where the user moves the mouth closer to the food bowl, as well as lifts the bowl (e.g., while consuming soupy dishes). While our current studies principally involved users exhibiting the first style, it is likely that the approaches for triggering image capture may not be effective when the

dish and the wrist/hand are both moving. Moreover, users may choose to exhibit different eating gestures at different times–e.g., when seated vs. standing. To address such diversity, one possibility may be to sense additional contexts provided by the user's smartphone (e.g., the level of crowd occupancy in the place or whether the user is standing vs. sitting), and use such context along with context obtained from internet (e.g., whether the place provides plastic chopsticks or metallic ones or whether the place is specific to a particular food type) to apply context-specific gesture recognition models.

Improved Accuracy of Image-based Diet Recognition: Initial studies (using a generic image analysis library) suggest that the automated recognition of specific food items may not be feasible in the near-term. We propose to evaluate a variety of approaches to potentially overcome this limitation. To improve the accuracy of automated recognition, one possibility may be to segment the image first into distinct regions (where each region consists of a single food item), and attempt recognition separately for each region. Additionally, the *modality* of eating may also be used to be used to narrow the search space of food items—e.g., knowing that a person is using chopsticks may eliminate or reduce the likelihood of the food item being a bowl of soup or a slice of bread. Finally, we may need to resort to a semi-automated approach: image analysis may limit the uncertainty to a small set of food items, and either the user or a mTurk worker may be asked (e.g., once every night) to manually annotate the corresponding image with one of these small set of labels.

Personalized vs. Generic Models: A classic challenge in activity recognition relates to the need for building *personalized classification models*. Clearly, person-independent models are much more practical (as they avoid the need for a person-specific training phase), but often suffer from unacceptably poor recognition accuracy. Our studies have shown that person-independent models are adequate (achieve over 97% accuracy) for identifying *eating vs. non-eating*. Similarly, the classification accuracy for eating mode (spoon vs. chopstick vs. hands) was 85.51% for a person-independent model, compared to an average of 90.74% for person-specific models. However, on closer inspection, we found the classification accuracy with the person-independent to be low for a person who was left-handed (while all the other users were right-handed), suggesting that we need to build separate models for left vs. right-handed individuals.

Privacy: Our approach of capturing images for use in diet identification can give rise to two distinct privacy concerns: (i) Due to errors in the gesture recognition system, the system might trigger during various potentially-sensitive non-eating activities (e.g., washing hands in the restroom); (ii) During meals, an erroneously triggered camera might capture images of other people at the table (instead of the meal being consumed). While these concerns *may* be mitigated by automated image preprocessing (e.g., blurring human faces in the captured images), the practical feasibility of these approaches, and their acceptance by consumers remains an open question.

VII. Conclusion

The use of wearable devices (specifically a wrist-worn smartwatch) for automatic recognition and profiling of various activities of daily living (ADLs) is an exciting possibility. In this paper, we have explored the possible use of the smartwatch (and its various sensors) for automatically capturing an individual's *eating* activity, and for potentially automatically identifying the type and quantity of food consumed. We have proposed and evaluated a cascaded classification approach where: (a) at the lowest level, the accelerometer sensor is used to distinguish eating gestures from similar non-eating activities–empirical studies show a classification accuracy of over 92%; (b) at the next level, once an eating episode is detected, additional information on the pattern of wrist rotation (captured by the gyroscope) is used to distinguish the mode of eating (spoon, hand or chopsticks)–empirical studies show accuracy of approx. 85.51%, and (c) at the final level, the smartwatch camera is used to snap an image (of the food item) at an appropriate point of the hand-to-mouth gestural trajectory–empirical studies show that a clear, usable image has a strong (more than 85.7%) chance of being captured at least once during an eating episode.

However, our early studies also show that there at least two major unsolved hurdles in achieving our vision of a practically useful and comprehensive diet monitoring system. First, we will need to automatically derive the right time to click an image. Second, we need to develop an appropriate automated (or semi-automated) mechanism to accurately identify the food type within the captured image. Additionally, we will have to empirically test the energy overheads of the implemented system to ensure that it does not drain the smartwatch battery excessively.

References

[1] Amft, O., Stäger, M., Lukowicz, P., and Tröster, G. Analysis of chewing sounds for dietary monitoring. *Proceedings of the 7th International Conference on Ubiquitous Computing*, UbiComp'05, pages 56–72, Berlin, Heidelberg, 2005. Springer-Verlag.

[2] Amft, O. and Tröster, G. Recognition of dietary activity events using on-body sensors. *Artificial Intelligence in Medicine*, 42(2):121–136, 2008.

[3] Dong, Y., Hoover, A., Scisco, J., and Muth, E. A new method for measuring meal intake in humans via automated wrist motion tracking. *Applied psychophysiology and biofeedback*, 37(3):205–215, 2012.

[4] Parate, A., Chiu, M.-C., Chadowitz, C., Ganesan, D., and Kalogerakis, E. Risq: Recognizing smoking gestures with inertial sensors on a wristband. *Proceedings of the 12th Annual International Conference on Mobile Systems, Applications, and Services*, MobiSys '14, pages 149–161, New York, NY, USA, 2014. ACM.

[5] Park, T., Lee, J., Hwang, I., Yoo, C., Nachman, L., and Song, J. E-gesture: A collaborative architecture for energy-efficient gesture recognition with hand-worn sensor and mobile devices. *Proceedings of the 9th ACM Conference on Embedded Networked Sensor Systems*, SenSys '11, pages 260–273, New York, NY, USA, 2011. ACM.

[6] Rahman, T., Adams, A. T., Zhang, M., Cherry, E., Zhou, B., Peng, H., and Choudhury, T. Bodybeat: A mobile system for sensing non-speech body sounds. *Proceedings of the 12th Annual International Conference on Mobile Systems, Applications, and Services*, MobiSys '14, pages 2–13, New York, NY, USA, 2014. ACM.

[7] Reddy, S., Parker, A., Hyman, J., Burke, J., Estrin, D., and Hansen, M. Image browsing, processing, and clustering for participatory sensing: lessons from a dietsense prototype. *Proceedings of the 4th workshop on Embedded networked sensors*, pages 13–17. ACM, 2007.

[8] Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., and LeCun, Y. Overfeat: Integrated recognition, localization and detection using convolutional networks. *International Conference on Learning Representations (ICLR 2014)*. CBLS, April 2014.

[9] Subbaraju, V., Sen, S., Misra, A., Chakraborty, S., and Balan, R. Using infrastructure-provided context filters for efficient fine-grained activity sensing. *Proceedings of the 2015 IEEE International Conference on Pervasive Computing and Communications*, PerCom'15.