A Study of Temporal Action Sequencing During Consumption of a Meal

Raul I. Ramos-Garcia
Department of Electrical and Computer Engineering
Clemson University
Clemson, SC
gramos@clemson.edu

Adam W. Hoover
Department of Electrical and Computer Engineering
Clemson University
Clemson, SC
ahoover@clemson.edu

ABSTRACT

Advances in body sensing and mobile health technology have created new opportunities for empowering people to take a more active role in managing their health. Measurements of dietary intake are commonly used for the study and treatment of obesity. However, the most widely used tools rely upon self-report and require considerable manual effort, leading to underreporting of consumption, non-compliance, and discontinued use over the long term. We are investigating the use of wrist-worn accelerometers and gyroscopes to automatically recognize eating gestures. In order to improve recognition accuracy, we studied the sequential dependency of actions during eating. Using a set of four actions (rest, utensiling, bite, drink), we developed a hidden Markov model (HMM) and compared its recognition performance against a non-sequential classifier (KNN). Tested on a dataset of 20 meals, the KNN achieved 71.7% accuracy while the HMM achieved 84.3% accuracy, showing that knowledge of the sequential nature of activities during eating improves recognition accuracy.

Categories and Subject Descriptors

I.5 [**Computing Methodologies**]: Pattern recognition

General Terms

Algorithms, Experimentation, Performance

Keywords

Hidden Markov models, gesture recognition, mobile health

1. INTRODUCTION

Advances in body sensing and mobile health technology have created new opportunities for empowering people to take a more active role in managing their health [11]. Obesity, which now afflicts one in three adults and one in six children in the United States [9, 14], has been recognized as a major health problem that could particularly benefit from this approach [8, 12, 24]. Self-monitoring of body weight, physical activity, and dietary intake have been consistently found to be associated with successful weight loss and maintenance [4]. Self-monitoring of dietary intake has specifically been described as "the single most important ingredient to successful dietary change efforts" [13]. However, currently used tools, including food diaries and 24 hour recalls, require users to manually estimate and record energy intake, making them prone to error and difficult to use for long periods of time [23].

Previous work done by our research group studied the tracking of wrist motion as it relates to eating [5, 6]. A method was developed to detect a pattern of wrist motion associated with the action of taking a bite, defined as placing food or liquid into the mouth [7]. The method was shown to be accurate across a wide variety of foods, counting bites with a true positive rate of 86% and a positive predictive value of 82% [7]. Additional research showed that bites, automatically counted using this method, correlated with self-reported caloric intake at the meal level at 0.53 [19]. This paper describes work that builds upon this approach. The original method treated all bites the same, regardless of context, and used a single pattern for detection. The proposed idea in this work is to study temporal sequencing as it relates to eating activities. Specifically, we seek to determine if the recognition of previous activities can be used to improve the recognition of subsequent activities.

Figure 1 demonstrates the idea, comparing our approach to speech recognition. The recognition of each piece of signal can be undertaken independently, in order to determine the word. However, if the recognition results from the previous pieces of signal are known, this can constrain and improve the subsequent recognition of the next piece of signal. In this example, the recognition of the piece of signal labeled "?" can be improved by using the recognition results of the previous pieces of signal ("Hello" "how" "are"). In this example, it may be highly expected that the next piece of signal encodes the word "you". We are pursuing the same idea with respect to recognizing eating activities. Our signal is obtained using accelerometers and gyroscopes to track wrist motion. The recognition of a piece of signal can be done independently, or can be augmented using recognition results from previous pieces of the signal. Figure 2 shows an example. If the actions "inactive" and "manipulate food" have previously been recognized, then it may be highly expected

BCB '13, September 22 - 25, 2013, Washington, DC, USA
Copyright 2013 ACM 978-1-4503-2434-2/13/09 ...$15.00.

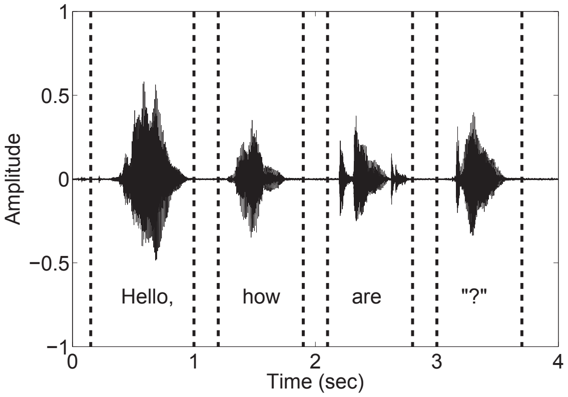

Figure 1: Speech recognition.

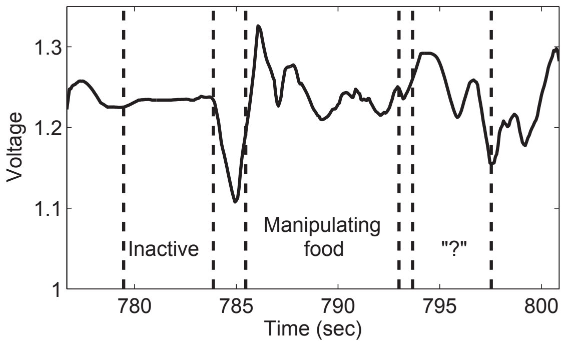

Figure 2: Eating activity recognition.

that the next action is "take a bite".

2. METHODS

2.1 Data

The data used for this study was recorded in the Harcombe Dining Hall at Clemson University [18]. The facility seats up to 800 guests and provides a wide range of foods and beverages. Our group instrumented a space inside the dining hall to record data from four participants simultaneously. Four digital video cameras in the ceiling (approximately 5 meters height) were used to record each participant's mouth, torso, and tray during meal consumption. A custom wrist-worn device containing MEMS acelerometers and gyroscopes [20, 21, 22] was used to record the wrist motion of each participant at 15 Hz. A scale was located under each participant's tray to monitor food weight during eating. Data was collected from 273 eaters. For the work in this paper, 20 of these meals were chosen randomly.

Accelerometer data (AccX, AccY, AccZ) and gyroscope data (Yaw, Pitch, Roll) were smoothed using a Gaussian-weighted window defined by equation 1. Here R_t is the raw data and S_t is the smoothed data at time t. For the sensors we used, the best results were obtained using $N = 15$ (1 second) and with $\sigma^2 = 10$.

$$S_t = \sum_{i=-N}^{0} R_{t+i} \frac{\exp\left(\frac{-t^2}{2\sigma^2}\right)}{\sum\limits_{x=0}^{N} \exp\left(\frac{-(x-N)^2}{2\sigma^2}\right)} \qquad (1)$$

2.2 Language

The problem of eating activity recognition requires the definition of a set of "words" that describe the individual activities. We base our definitions on discernible user intent. The subject's intent is determined by observing the hand wearing the device. The duration of an action lasts from when the intent can first be observed, to when that intent has ended. For this work we define four actions related to eating: *rest*, *utensiling*, *bite*, and *drink*. All other actions for which intent was not defined, including both eating and non-eating activities (e.g. gesturing while talking, cleaning with a napkin, waving at a friend, etc.) are referred to as *other*. We developed a definition for each word consisting of four parts: 1) a description of the activity, 2) the start time of the activity, 3) the end time of the activity, and 4) particular events that should be included or excluded from the word label. The following lists the definitions we used:

- Bite

  1. The subject puts food into their mouth.
  2. Begins when a hand or utensil starts moving towards the mouth.
  3. Ends when the hand or utensil finishes moving away from the mouth
  4. If a bite is interrupted, the word starts after the interruption, when motion towards the mouth resumes. Bites need not begin and end at the plate. Motion towards and away from the mouth should define the boundaries; with food intake taking place in between.

- Rest

  1. The subject's dominant hand has little or no motion.
  2. Begins when subject's hand stop moving.
  3. Ends when subject's hand begins moving again (and moves for at least one second)
  4. A rest may include brief periods of motion (less than one second) such as posture adjustments. Rests should not include motions with other intent, whether eating related or not (such as gesturing while talking, cleaning with a napkin, or waving to a friend).

- Utensiling

  1. The subject uses a utensil or their hand to manipulate, stir, mix or prepare food(s) for consumption.
  2. Starts when utensil or hand moves towards food with intent to manipulate.
  3. Ends when manipulating has finished.
  4. Examples include moving food around the plate, dipping foods in sauces, and cutting foods.

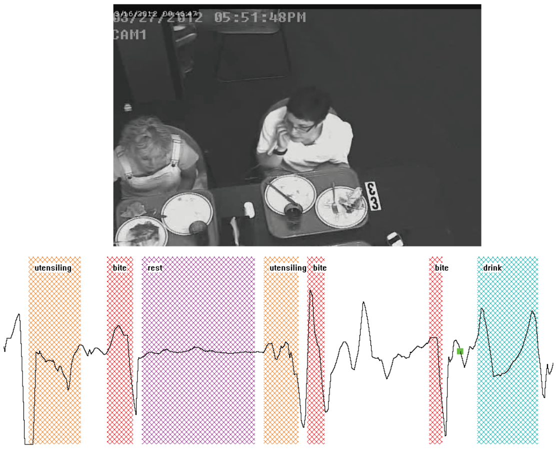

Figure 3: Labeling software.

- Drink
  1. The subject puts beverage into their mouth.
  2. Starts when a hand begins moving a beverage towards the mouth.
  3. Ends when the hand has finished moving away from the mouth.
  4. If multiple sips are taken, each individual sip defines a different drink.

2.2.1 *Ground truth*

A custom tool was developed for reviewing the recorded data to label segments with the word defintions. These hand labels were used as ground truth to evaluate classifier performance. The tool was coded using Microoft Visual C. Video and sensor information are synchronized and displayed as shown in figure 3. Time navigation is done using the keyboard to move forwards, backwards, play and pause. Labeling was done manually by looking at the intent of the instrumented hand of the eater in the video. A word could be labeled by enclosing it within a box using specific keys in the keyboard. A completed label is marked by a color box and a legend on top of the box to discriminate between labels as shown in the bottom of figure 3.

2.2.2 *Other words*

After ground truthing there will be periods of unlabeled time between words due to the transitioning from one word to another (e.g. utensiling to bite). We refer to these as "gaps". Because our sampling rate is 15Hz, some of these gaps could be as small as 67ms. We do not consider these gaps to be the same as larger sections of unlabeled data that correspond to other activities. We therefore devised a strategy to remove them from consideration. Figure 4 shows the distribution of unlabeled sections of data from the 20 meals for periods of up to 15 sec. The figure shows a knee in the curve at 4 sec (dashed line) which we used to indicate where the true distribution for other unlabeled activities overlaps the gap distribution. Based on this analysis, we discarded unlabeled sections of data that were less than 4 seconds, and labeled sections longer than 4 seconds using the word *other*.

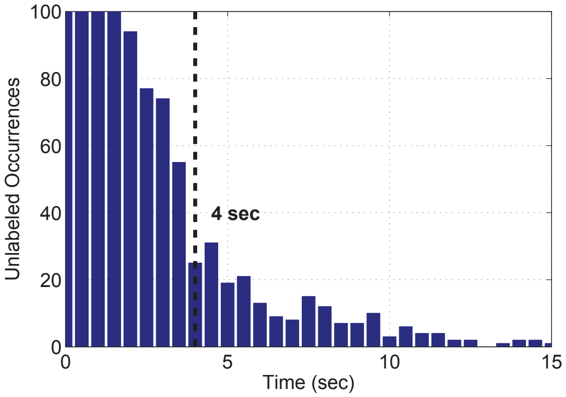

Figure 4: Histogram of unlabeled sections.

Table 1: Total words and time after removing gaps.

Word	Total Count	Total Time (min)
Rest	552	69.0
Utensiling	541	64.5
Bite	854	47.4
Drink	125	16.2
Other	225	29.8
Total	2297	226.8

Table 1 shows the final count of all 5 words, and total time for each word, for the 20 meals. A total of 86.9% of the data was labeled using our 4 word language and 13.1% was considered other activities.

2.2.3 *Inter-rater reliability*

In order to determine the stability of our definitions and ground truthing process, three meals were chosen randomly and labeled independently by five raters. A "meta" ground truth was created for each meal by taking the majority vote of the five raters for every unit time (67 msec, from the 15Hz rate of recording). Figure 5 shows the process graphically. The meta GT is calculated independently for every time unit. If all individual rater labels are different then the meta GT is left unlabeled. The complete labelings of each rater and the meta GT are shown for all 3 meals in figure 6. Visually it can be seen that raters agreed fairly consistently.

To evaluate agreement quantitatively, individual rater labelings were compared against the meta GT. The comparison was done by evaluating if each unit of time of a rater' sequence is labeled identically to that of the meta GT. All individual unit times which are labeled the same represent the total time of agreement. The process was repeated for all raters within a meal and the average of total agreement time was computed. The results from each meal were combined to show that the average inter-rater agreement across the three meals was 90.7%. Table 2 shows the confusion matrix.Diagonal entries indicate how often the raters labeled a unit of time with the same label as the meta GT, and can be interpreted as total agreement. Off-diagonal entries indicate how often the raters labeled a unit of time differently from

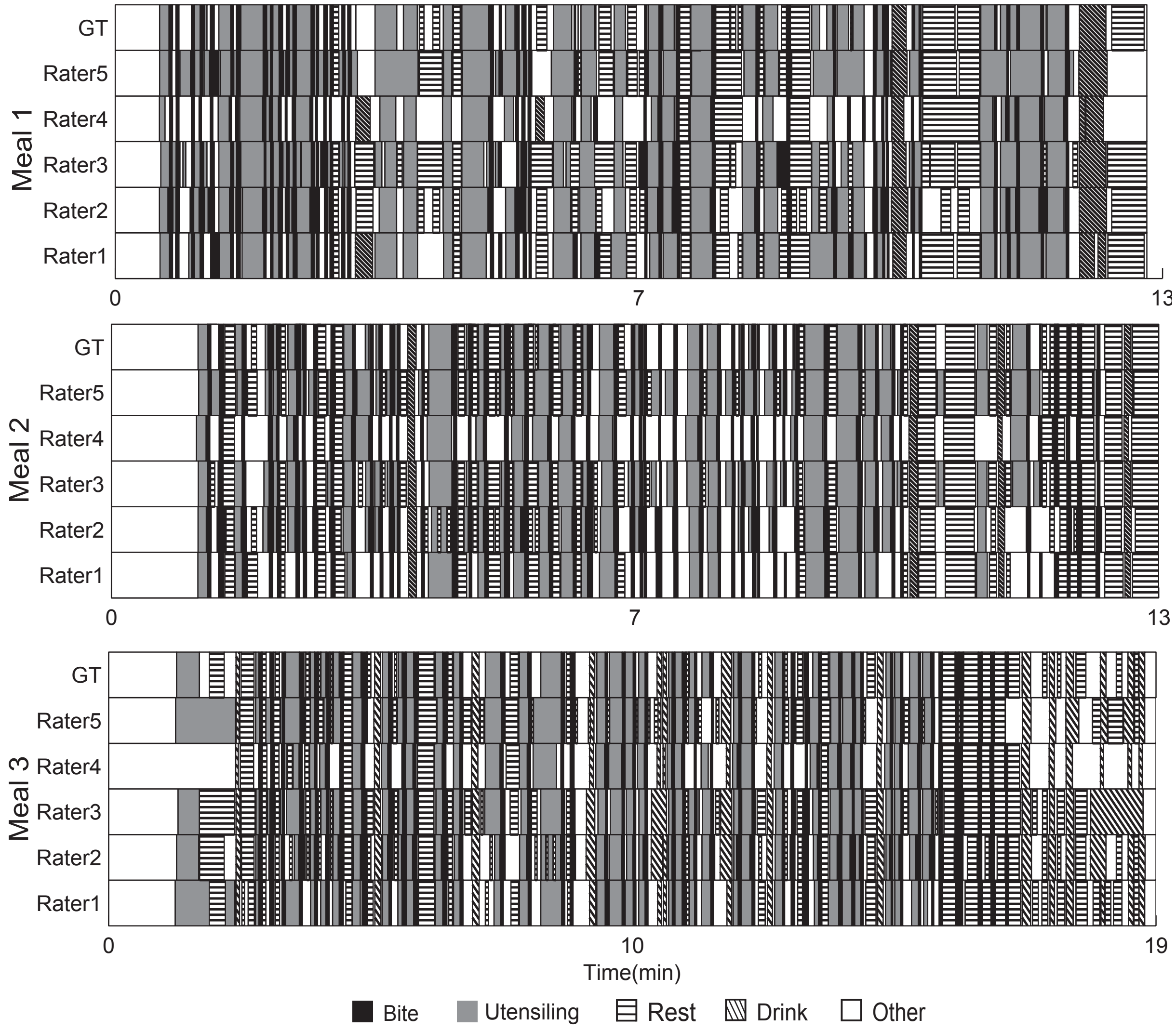

Figure 6: Meta ground truth from multiple raters.

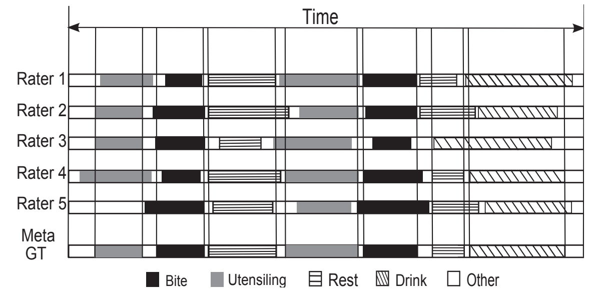

Figure 5: Creating meta ground truth from 5 independent raters.

Table 2: Inter-rater confusion matrix (units are % time during 3 meals). Diagonal elements are agreement; off-diagonal elements are confusion.

Words	Rest	Utensiling	Bite	Drink	Other
Rest	13.1	0.2	0.1	0.2	2.9
Utensiling		20.0	0.5	0.0	2.4
Bite			8.2	0.0	1.0
Drink				4.0	0.5
Other					38.9

the meta GT, and can be considered confusion. As the table shows, the largest amount of confusion was with the word *other*, which is to be expected given it is a catch-all term for all undefined activities.

2.3 Classification

2.3.1 Hidden Markov model

Hidden Markov models (HMMs) are particularly well-suited for classification problems having temporal data dependence [3, 15]. We designed an ergodic HMM to model the sequential nature of eating activity words as shown in figure 7. Specifically, the HMM is composed of five states which are rest, utensiling, bite, drink, and other. To train the HMM, we used the data from the 20 meals to compute transition and prior probabilities. Transition probabilities are calculated as shown in equation 2 as a ratio of the total number of transitions going from word s_i to word s_j, where $i, j = 1, 2, 3, 4, 5$. Prior probabilities are calculated as the ratio of the total number of word s_i over the total number of word samples in the data base as described in equation 3. The values we found for the transition probabilities and prior probabilities are shown in tables 3 and 4, respectively.

Table 3: Transition probabilities.

To From	Rest	Utensiling	Bite	Drink	Other
Rest	0.089	0.290	0.397	0.083	0.140
Utensiling	0.172	0.006	0.786	0.015	0.022
Bite	0.347	0.335	0.169	0.026	0.214
Drink	0.224	0.232	0.200	0.112	0.232
Other	0.382	0.276	0.182	0.156	0.004

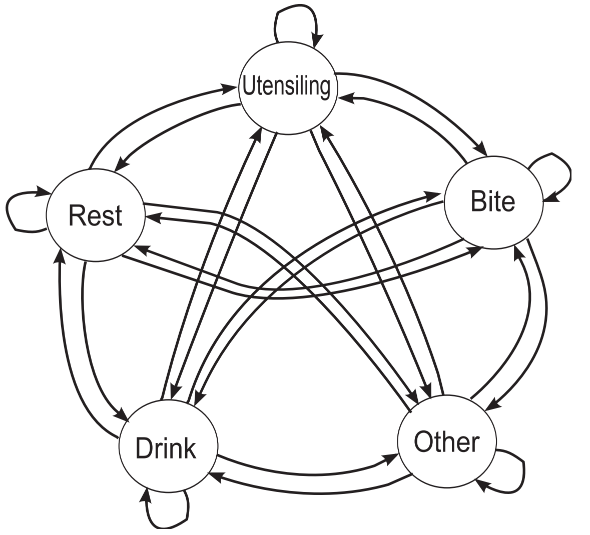

Figure 7: HMM (prior and transition probabilities given in tables).

Table 4: Prior probabilities.

Word	Probability
Rest	0.240
Utensiling	0.236
Bite	0.372
Drink	0.054
Other	0.098

$$a_{ij} = \frac{\text{Total \# of transitions from word } s_i \text{ to word } s_j}{\text{Total \# of transitions from word } s_i} \quad (2)$$

$$\pi_i = \frac{\text{Total \# of word } s_i}{\text{Total \# of words}} \quad (3)$$

Words can be modeled as continuous observations using Gaussian mixture models through the EM algorithm [17]. The Gaussian mixture model for word s_i is given by equation 4, where $c_{s_{ik}}$, $\mu_{s_{ik}}$, and $\Sigma_{s_{ik}}$, represent the mixture weight, the mean, and the covariance matrix of the kth Gaussian, respectively. We calculated these models using the HMM toolbox in MATLAB[1]. The number of Gaussians M chosen is discussed below in section 2.4 during feature selection.

$$G_{s_i} = \{c_{s_{ik}}, \mu_{s_{ik}}, \Sigma_{s_{ik}}\}, \text{ where } k = 1, ..., M \quad (4)$$

2.3.2 K-nearest neighbor

For comparison with a technique that does not account for temporal dependency, we classified the data using a KNN classifier [16]. A KNN assumes that the data is in a feature space, or metric space, which has a notion of distance. Each vector in the training data is associated with a class label. The process of classification places an unlabeled sample x in the feature space among the previously labeled samples,

[1] http://www.cs.ubc.ca/ murphyk/Software/HMM/hmm.html

and uses an exhaustive computation of distances between x and all labeled samples. The classification searches for the K closest labeled samples to x. The number of occurrences of each label are calculated among the K closest samples. The highest label occurrence is then assigned as the label of the unlabeled sample x. A Euclidean distance (equation 5) is used for classification, where x is the test sample and y is a training sample. In this work the optimal number of neighbors K is estimated in the feature selection section (see next).

$$D_E(x, y) = \sqrt{\sum_{j=1}^{d} (x_j - y_j)^2} \tag{5}$$

2.4 Feature and model parameter selection

The wrist-worn device provides six analog signals from the accelerometer (AccX, AccY, AccZ) and gyroscope (Yaw, Pitch, Roll) sensors which are used to compute features. Features were calculated for each word using the manually segmented data from the 20 meals. We were uncertain which features would be most descriptive of the chosen words, and so calculated a large number of features. A total of 29 features were calculated for each of the 6 sensors, yielding 174 features. We also calculated correlations for all possible combinations of the 6 sensors. Lastly, the duration of a manually segmented word was used as a feature. Collectively, this resulted in a total of 190 features. This feature vector was reduced to a reasonable count of those most useful for classification using a feature forward selection method. This is an iterative process in which the data is trained and tested on different subsets of the feature space, selecting the best feature set at each iteration. The strategy then was to do a forward feature selection using the entire available data as training and test data to accomplish two purposes: 1) select the model order for the sequential and non-sequential classifier, and 2) build a feature vector with a smaller dimension.

For the HMM a forward feature selection was done by training the data as continuous observations using Gaussian mixture order models of $M = 1$ to $M = 9$. Figure 8 shows the result after training for each model order up to a feature vector candidate of dimension 30. It can be noticed that after $M = 3$, higher orders model tend to produce similar results, and although there are slight increases in accuracy as M grows, this is likely a consequence of overfiting the data with more Gaussians. We therefore selected a model order of $M = 3$. It can also be seen that after a feature size of 15, the accuracy tends to be fairly similar. We therefore selected a 15 element feature vector. Table 5 shows the list of features for the HMM.

This process was repeated for the KNN classifier. A forward feature selection was done using the 190 features for $K = 1, 3, 5, 7, 9, 11, 13, 15$ neighbors. The data was split in two halves, i.e., 10 meals for training and 10 meals for testing. Figure 9 shows the accuracy of feature vector candidates up to 30 elements. From the graph it can be seen that after $K = 5$ accuracy tends to be concentrated for higher values of K, indicating that 5 neighbors are sufficient. It can also be seen that after a feature vector of size 20, accuracy rises abruptly, and we can see this rise appear sooner for higher values of K. We believe this is a result of overfitting, and so selected a feature vector of size 15. Table 6 shows the list of features for the KNN test.

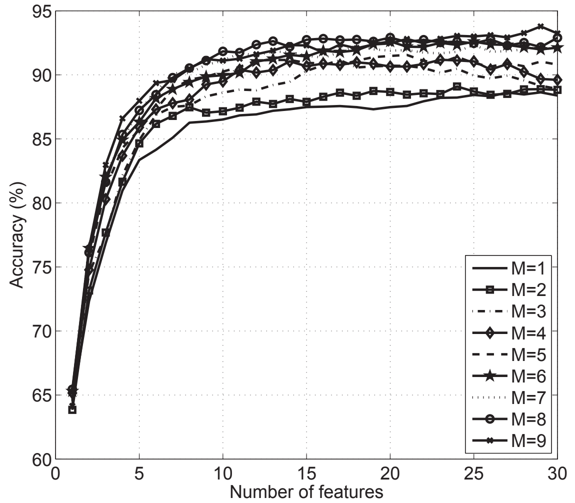

Figure 8: HMM forward feature selection.

Table 5: HMM features.

No.	HMM Features
1	Interquartile range of Roll
2	Difference mean of absolute value (MAV) of Roll
3	Standard deviation of Yaw
4	Root mean square (RMS) of AccZ
5	Skewness of AccY
6	Zero crossings (ZC) of Roll
7	Max time between ZC of AccY
8	Skewness of AccZ
9	Mean of Roll
10	Difference MAV of AccY
11	Median time between ZC of AccZ
12	Max time between ZC of Yaw
13	Correlation(AccX,Yaw)
14	Correlation(Pitch,Roll)
15	Difference MAV of Pitch

3. RESULTS

For the HMM, recognition of words was done for an observation sequence by obtaining the state sequences through Viterbi decoding. The estimated state sequence was then compared against the true state to determine the accuracy. For the KNN, a simple Euclidean distance was computed from a given test sample to each training sample. All the data was evaluated using 10-fold cross validation, where 18 meals were used to train the classifiers for testing on the other 2 meals. The KNN achieved an accuracy of 71.7% while the HMM achieved an accuracy of 84.3%. Figure 10 shows the improvement of the HMM over the KNN for each of the words. The overall improvement using sequential context information is approximately 13%.

4. CONCLUSIONS

Amft and colleagues pioneered the idea of recognizing eating activities using body-worn sensors [1, 2, 10] and used

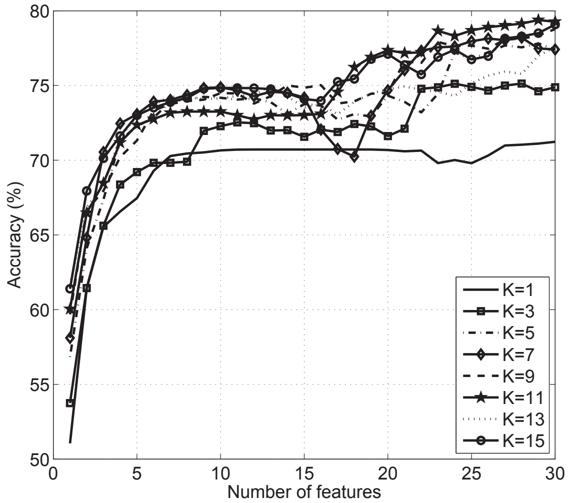

Figure 9: KNN forward feature selection.

Table 6: KNN features.

No.	KNN Features
1	Interquartile range of Yaw
2	Ratio RMS to MAV of AccZ
3	Difference MAV of Roll
4	Variance of Yaw
5	Ratio RMS to MAV of AccY
6	Variance of Roll
7	Difference MAV of Yaw
8	Ratio RMS to MAV of Roll
9	Ratio RMS to MAV of Pitch
10	Ratio RMS to MAV of Yaw
11	Difference MAV of AccY
12	Variance of Pitch
13	Difference MAV of Pitch
14	Ratio RMS to MAV of AccX
15	Standard deviation of Yaw

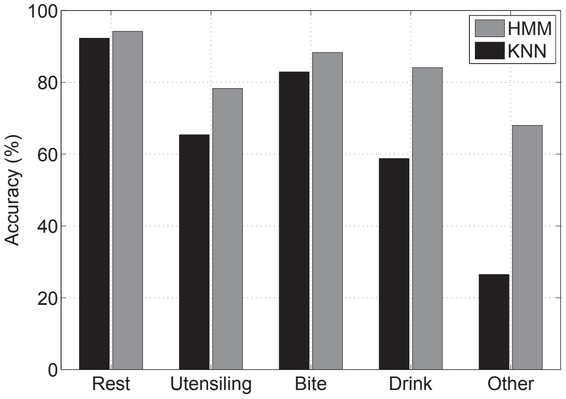

Figure 10: Results.

HMMs as classifiers. However, to our knowledge, our work is the first to study the significance of temporal sequencing of actions during eating. In this paper, four eating activities were defined and an other category is considered to model all activities which do not belong to our four main eating activities. To this point the study has been developed at the meal level, therefore sequential dependency of other daily activities are not included. Inter-rater reliability was found to be 90.7% showing that our definitions are fairly objective. Although four words is a fairly limited set, we found that they comprised on average 86.9% of the time during a meal. A confusion matrix shows that the greatest ambiguity resides in the remaining 9.3% of other activities, suggesting that additional words may be useful (e.g. cleaning hand or mouth with napkin). The comparison of an HMM and KNN classifier on our data showed an approximately 13% improvement using the HMM, showing that the temporal context of the preceding action improves recognition accuracy. Future work will seek to determine if a more complex HMM further increases accuracy, by modeling temporal context across multiple word sequences and by increasing the vocabulary. Our method could also prove useful for analyzing an individual's eating habits. For example, obesity has been treated using behavioral change methods. Thus, it is plausible that an HMM trained to an individual could show habits that affect the individual's eating behavior and help the individual to correct bad habits. It is envisioned that our work will provide a tool that estimates energy intake automatically. Thus, by analyzing eating patterns, it could be possible to determine the type of food that the individual is consuming and adjust the prediction of energy intake.

5. ACKNOWLEDGMENTS

We gratefully acknowledge the support of the NIH via grant 1R41DK091141-A1.

6. REFERENCES

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

[2] O. Amft and G. Tröster. On-body sensing solutions for automatic dietary monitoring. *IEEE Pervasive Computing*, 8(2):62–70, 2009.

[3] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[4] L. E. Burke, J. Wang, and M. A. Sevick. Self-monitoring in weight loss: A systematic review of the literature. *Journal of the American Dietetic Association*, 111(1):92–102, 2011.

[5] Y. Dong. *Tracking Wrist motion to Detect and Measure the Eating Intake of Free-Living Humans*. PhD dissertation, Electrical and Computer Engineering Department, Clemson University, SC., 2012.

[6] Y. Dong, A. Hoover, and E. Muth. A device for detecting and counting bites of food taken by a person during eating. In *IEEE Int'l Conf on Bioinformatics and Biomedicine*, pages 265–268, 2009.

[7] Y. Dong, A. Hoover, E. Muth, and J. Scisco. A new method for measuring meal intake in humans via

automated wrist motion tracking. *Applied Psy-chophysiology and Biofeedback*, 37(3):205–215, 2012.

[8] A. Ershow, A. Ortega, J. Baldwin, and J. Hill. Engineering approaches to energy balance and obesity: Opportunities for novel collaborations and research: Report of a joint National Science Foundation and National Institutes of Health workshop. *Journal of Diabetes Science and Technology*, 1(1):96–105, 2007.

[9] K. Flegal, M. Carroll, B. Kit, and C. Ogden. Prevalence of obesity and trends in the distribution of body mass index among us adults, 1999-2010. *J. American Medical Association*, 307(5):491–497, 2012.

[10] H. Junker, O. Amft, P. Lukowicz, and G. Tröster. Gesture spotting with body-worn inertial sensors to detect user activities. *Pattern Recognition*, 41(6):2010–2024, 2008.

[11] S. Kumar, W. Nilsen, M. Pavel, and M. Srivastava. Mobile health: Revolutionizing healthcare through transdisciplinary research. *Computer*, 46(1):28–35, 2013.

[12] B. McCabe-Sellers. Advancing the art and science of dietary assessment through technology. *Journal of the American Dietetic Association*, 110(1):52–54, 2010.

[13] B. McCann and V. Bovbjerg. Promoting dietary change. *The handbook of health behavior change*, edited by S. Shumaker, J. Ockene and K. Riekert, New York, Springer Publishing Company, 2009.

[14] C. L. Ogden, M. D. Carroll, B. K. Kit, and K. M. Flegal. *Prevalence of obesity in the United States, 2009-2010*. US Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics, 2012.

[15] L. Rabiner and B. Juang. An introduction to hidden markov models. *IEEE ASSP Magazine*, 3(1):4–16, 1986.

[16] L. Rabiner, S. Levinson, A. Rosenberg, and J. Wilpon. Speaker-independent recognition of isolated words using clustering techniques. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 27(4):336–349, 1979.

[17] D. Reynolds and R. Rose. Robust text-independent speaker identification using gaussian mixture speaker models. *IEEE Transactions on Speech and Audio Processing*, 3(1):72–83, 1995.

[18] J. Salley. Accuracy of a Bite-Count Based Calorie Estimate Compared to Human Estimates With and Without Calorie Information Available. Master thesis, Psychology Department, Clemson University, 2013.

[19] J. Scisco. *Sources of Variance in Bite Count*. PhD dissertation, Pshycology Department, Clemson University, 2012.

[20] STMicroelectronics. Lis344alh: Mems inertial sensor. `http://www.st.com/stonline/products/literature/ds/14337.pdf`.

[21] STMicroelectronics. Lpr410al: Mems motion sensor. `http://www.st.com/st-web-ui/static/active/en/resource/technical/document/datasheet/CD00254123.pdf`.

[22] STMicroelectronics. Lpy410al: Mems motion sensor. `http://www.st.com/st-web-ui/static/active/en/resource/technical/document/datasheet/CD00254136.pdf`.

[23] F. Thompson and A. Subar. *Dietary Assessment Methodology*. Academic Press/Elsevier, 2 edition, 2008.

[24] F. E. Thompson, A. F. Subar, C. M. Loria, J. L. Reedy, and T. Baranowski. Need for technological innovation in dietary assessment. *Journal of the American Dietetic Association*, 110(1):48–51, 2010.