2016 IEEE First Conference on Connected Health: Applications, Systems and Engineering Technologies

Recognizing Eating Gestures Using Context Dependent Hidden Markov Models

Yiru Shen
Department of Electrical and
Computer Engineering
Clemson University
Clemson, SC
yirus@clemson.edu

Eric Muth
Department of Psychology
Clemson University
Clemson, SC
muth@clemson.edu

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

***Abstract*—This paper considers the problems of recognizing eating gestures by tracking wrist motion. Hidden Markov models (HMMs) were developed to capture variations in motion patterns of subgroups of participants. Specifically, we examined if foreknowledge of the gender, age, and utensil used for eating could improve recognition accuracy. Improvement in accuracy was measured by comparing to a baseline HMM that was trained on all participants. Data was collected for 276 participants eating a single meal within a cafeteria setting. A total of 44,873 gestures were manually labeled using video synchronized with the wrist motion tracking device. Results show that gender HMMs performed slightly better than the baseline, indicating that there is not much difference in wrist motion patterns during eating between females and males. Age HMMs provided a 4.3% increase in accuracy and utensil HMMs provided a 6.2% increase in accuracy. The results suggest that contextual variables can be used for improving gesture recognition.**

I. Introduction

Mobile health (mHealth) technologies can help people monitor body status and track behaviours to empower self-management of health conditions [1]. The problem considered in this work is the monitoring of energy intake, a measure of the amount of food and drink ingested [2]. The motivating health problem is obesity. In the U.S., 17% of children and more than 30% of adults are considered obese [1]. Self-monitoring of energy intake has been found to significantly correlate with weight loss [3]. Additional evidence shows that for obese adults, self-monitoring of energy intake improves the outcome for maintaining weight loss [4]. Conventional methods for self-monitoring include manual entry of self-reported intake into food diaries and 24-hour recalls [5], [6]. However, these methods are prone to under-reporting and under-estimation and are tedius to use resulting in non-compliance over the long term.

Wearable sensors offer the opportunity for automatic monitoring of energy intake [7], [8]. Several positions on the human body can be instrumented to detect activities associated with eating [7]. The ear can be instrumented with a microphone to detect sounds associated with chewing [9]–[13]. The jaw can be instrumented with a strain sensor to detect jaw motions associated with chewing [14]–[16]. The throat can be instrumented with a microphone to detect sounds associated with swallowing [8], [15], [17]–[19]. The arms can be instrumented with motion sensors to detect patterns of limb motion associated with activities during eating, such as the use of cutlery and hand-to-mouth gestures [20]–[22]. In a simpler configuration, the wrist alone can be instrumented with motion sensors to detect patterns of wrist motion associated with eating [9], [23]–[26].

Our research group has been investigating methods based upon wrist motion tracking [24], [25]. In [24] we describe a pattern of motion indicative of hand-to-mouth gestures and an algorithm to detect and count their occurrences, which we call bite counting. In [25] we describe methods using hidden Markov models (HMMs) to detect five different types of gestures (food bite, drink bite, utensiling, rest, and other), and an algorithm that improves their recognition through gesture-to-gesture sequential modeling. In this work, we investigate whether contextual variables can improve the recognition of these gesture types. Specifically, we consider if foreknowledge of the gender of the subject, the age of the subject, and the type of utensil being used can improve recognition accuracy. We investigate this hypothesis by building HMMs trained for each of these contextual variables, and comparing their recognition accuracy against a baseline HMM having no foreknowledge.

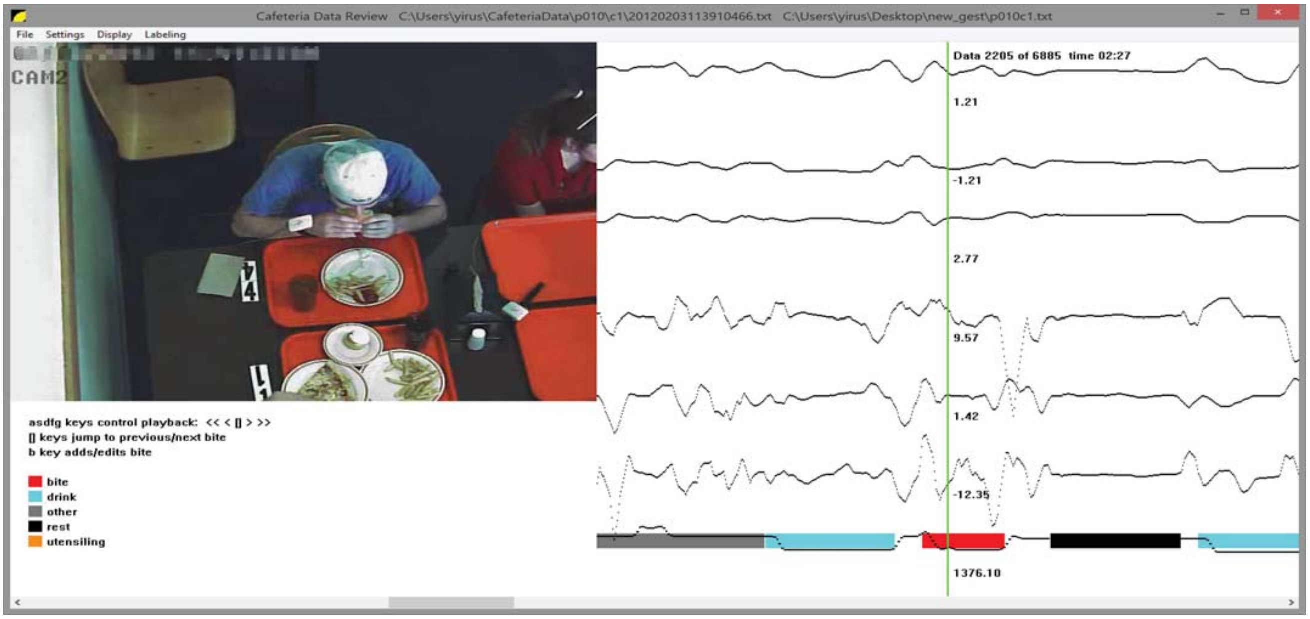
II. Methods

A. Data

Data was collected for 276 subjects eating a single meal in the Harcombe Dining Hall at Clemson University [27]. Participants were free to choose any foods and beverages available. In total, 380 different food and beverage types were chosen, for example stir fry vegetables, pasta, shoestring French fries, salad bar, water, soda, etc. Four different utensils were used: fork, spoon, chopsticks and hand. Tables I-II list the gender and age distributions of participants that were tested as contextual variables in this work.

Digital cameras were mounted on the ceiling to video record participants while they ate. A custom device was designed to record the wrist motion of subjects at 15 Hz during eating, using accelerometers to measure acceleration of x, y and z axis (AccX, AccY, AccZ), and gyroscopes to measure rotational velocity around yaw, pitch and roll [28]. In previous work our group defined a set of five gesture types (rest, utensiling,

978-1-5090-0943-5/16 $25.00 © 2016 IEEE
DOI 10.1109/CHASE.2016.9


Cafeteria Data Review C:\Users\yirus\CafeteriaData\p010\c1\20120203113910466.txt C:\Users\yirus\Desktop\new_gest\p010c1.txt
File Settings Display Labeling
CAM2
1.21
-1.21
2.77
9.57
1.42
-12.35
1376.10

Data 2205 of 6885 time 02:27

asdfg keys control playback: << < [] > >>
[] keys jump to previous/next bite
b key adds/edits bite

🟥 bite
🟦 drink
⬛ other
⬛ rest
🟧 utensiling

Figure 1. A custom program for gesture labeling.

bite, drink, and other) based on subject intent that could be visually determined by observing the video [29]. A custom tool was developed for labeling periods of time as gestures. Figure 1 shows a screenshot of this tool. Top to bottom shows the 6 axes of motion (AccX, AccY, AccZ, yaw, pitch and roll) with a seventh line at the bottom indicating tray weight as measured from a table embedded scale (not used in this work). Boxes overlaid over this seventh line indicate periods of time labeled as gestures (for example, red = bite). Unlabeled segments with duration longer than 4 seconds are considered as type other, unlabeled segments shorter than 4 seconds are considered transitions between gestures and are ignored [29].

Gender	#Participants
Male	96
Female	119

Table I: Gender distribution of participants.

Age	#Participants
18-30	152
31-40	18
41-50	26
51-75	19

Table II: Age distribution of participants.

Prior to processing, each of the 6 axes of data is smoothed using a Gaussian weighted window as shown in equation 1, where R_t and S_t are the raw data and smoothed data at time t, respectively. We use N = 15 and σ^2 = 10 for best performance [28].

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

Due to the work-intensive nature of this labeling process, in previous work a subset of data from 25 participants was used [29]. At the time of this writing, a team of 22 raters has contributed to labeling the full data set and has so far completed 215 participants in approximately 500 man-hours of work.

B. Baseline HMMs

To evaluate the performance of our method, we implemented a baseline classifier which was designed in our previous research [29], [30]. This classifier models a gesture as a sequence of sub-gestures with each sub-gesture represented by a state within the HMM. For example, the action of taking a bite may consist of raising food towards the mouth, ingestion, and the return of the wrist to a rest position. This sequence of actions is modeled through a state sequence where each state models part of the motion pattern.

Features are calculated using a 0.5 second sliding window with 50% overlap for each sub-gesture. Figure 2 illustrates the process, where the shaded area is the 50% overlap between the first and second window positions. Within each window we calculate the following features: average, standard deviation and slope for each of the six motion axes (AccX, AccY, AccZ, yaw, pitch and roll). Equations 2-4 provide the formal definitions where w is the window size (for our data, 0.5 sec = 8 samples).

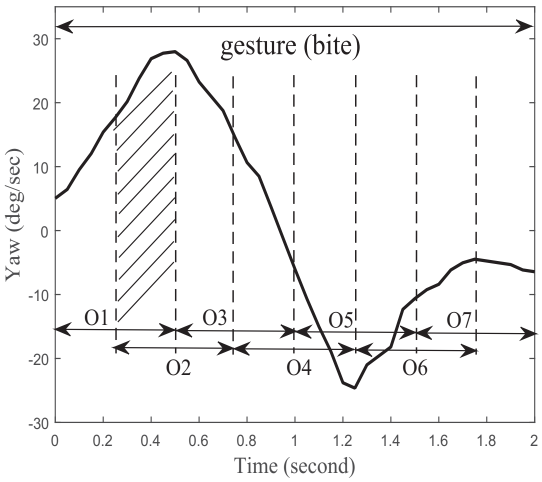


Figure 2. Example yaw motion of sub-gestures for bite.

$$\text{slope} = (x_{end} - x_{begin})/w \tag{2}$$

$$\text{mean} = \sum_{i=1}^{w} x_i/w \tag{3}$$

$$\text{std} = \frac{\sum\limits_{i=1}^{w}(x_i - \bar{x})^2}{w-1} \tag{4}$$

In total this provides 18 features comprising an observable feature vector $[o_1, o_2, o_3, ..., o_{18}]$. The number of feature vectors for a gesture depends upon its length in time; for example, the bite depicted in Figure 2 lasted 2 seconds and therefore provides a sequence of 7 observation vectors. Each vector is standardized as zero-mean and unit-variance as shown in Equation 5, where O is the original feature vector, $\bar{O}$ is the mean of the feature vector, and σ is its standard deviation.

$$O' = \frac{O - \bar{O}}{\sigma} \tag{5}$$

Emission probabilities were modeled by Gaussian mixture models (GMMs) as in Equation 6, where E is the emission probability distribution, O is the d-dimensional feature vector and M is the number of Gaussians in the model.

$$E = \sum_{i=1}^{M} c_i N(O; \mu_i, \Sigma_i) \quad \text{where} \quad \sum c_i = 1. \tag{6}$$

Each Gaussian N is defined by three parameters c_i, μ_i, and Σ_i representing the weight, mean vector and covariance matrix of the i^{th} Gaussian, respectively:

$$\begin{aligned} N(O; \mu_i, \Sigma_i) = & \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \\ & \exp(-\frac{1}{2}(O-\mu)^T \Sigma^{-1}(O-\mu)) \end{aligned} \tag{7}$$

The expectation maximization algorithm was used to calculate GMMs having diagonal matrices [31], [32]. HMMs were built for each gesture type (rest, utensiling, bite, drink and other) using an HMM toolbox [33]. The architecture for each HMM is left-to-right with skip as shown in Figure 3. Each HMM used 13 states with 5 Gaussians which was determined in previous work to be sufficient to capture the variation in motion patterns [29].

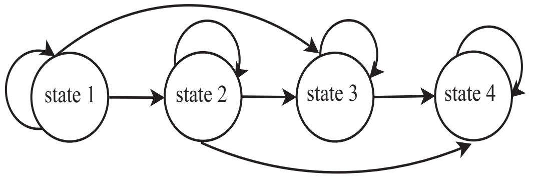


Figure 3. HMM architecture of left-to-right with skip.

During the recognition process, each HMM outputs a log probability of a sequence of observations. The recognition problem can be viewed as evaluating how well each HMM matches the observable sequence of a gesture, where a higher probability indicates a better match. Therefore, the HMM which provides the maximum log probability determines the classification of the gesture. Figure 4 demonstrates the process. where the observable sequence O is a series of feature vectors in one single gesture, and the highest log probability output by the five HMMs determines the classification.

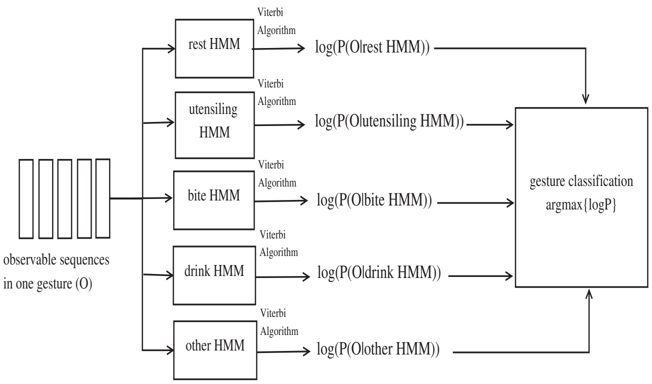


Figure 4. Baseline HMM gesture recognition.

C. Context Dependent HMMs

Three contextual variables were studied in this work to determine if they provide increased recognition accuracy compared to the baseline classifier. Each was tested independently.

1) Gender HMMs: HMMs were trained independently for females and males using the same steps described for the baseline classifier. This yielded 10 total HMMs (2 genders × 5 gesture types). During the recognition process, it is assumed that the gender of the participant is known a priori and thus can be used to determine which set of HMMs to use to recognize gestures. Figure 5 shows the process. After selecting gender, the remainder of the process works as outlined in Figure 4.

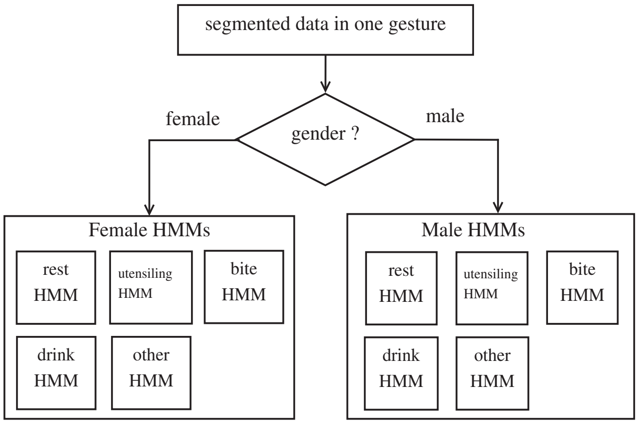


Figure 5. Gender HMMs gesture recognition.

2) Age HMMs: HMMs were trained independently for each age category listed in table II. This yielded 20 total HMMS (4 age groups × 5 gesture types). During the recognition process, it is assumed that the age of the participant is known a priori and thus can be used to determine which set of HMMs to use to recognize gestures. This selection process is similar to the one outlined in Figure 5. After selecting age group, the remainder of the process works as outlined in Figure 4.

3) Utensil HMMs: Four utensil types were available in our data set: fork, spoon, hands and chopsticks. However, utensil use is not necessarily unique throughout an entire meal. For example, a participant may use a fork for some bites and hands for other bites. Therefore, we defined a fifth category as mixed utensil use. If no single utensil type was used for more than 65% of bite gestures by a participant, then their utensil type was considered mixed. Table III lists the totals.

Utensil	#Participants
fork	103
spoon	36
hands	85
chopsticks	4
mixed	90

Table III: Utensil distribution of participants.

This yielded 25 total HMMS (5 utensil types × 5 gesture types). During the recognition process, it is assumed that the utensil type of the participant is known a priori and thus can be used to determine which set of HMMs to use to recognize gestures. This selection process is similar to the one outlined in Figure 5. After selecting utensil type, the remainder of the process works as outlined in Figure 4.

4) Evaluation Metric: The accuracy of each classifier was evaluated as the total percentage of gestures that were labeled correctly in all data. Accuracy by gesture type is also reported.

III. Results

The total data set consists of 44,873 manually labeled gestures. All classifiers were trained and tested using leave-one-out cross validation; for each participant, the HMMs were trained on the data for all other participants and then tested on the one left out. Due to the Monte Carlo nature of HMM training, each classifier was run 10 times and the average is reported. An example of a classifier output is shown in Figure 6. The upper row of gestures are ground truth and the bottom row are the classification results from the utensil HMMs. In this figure, it can be seen that the utensil HMMs misclassified one gesture of other as utensiling. This is likely because both utensiling and other gestures cover a wide range of possible motion patterns.

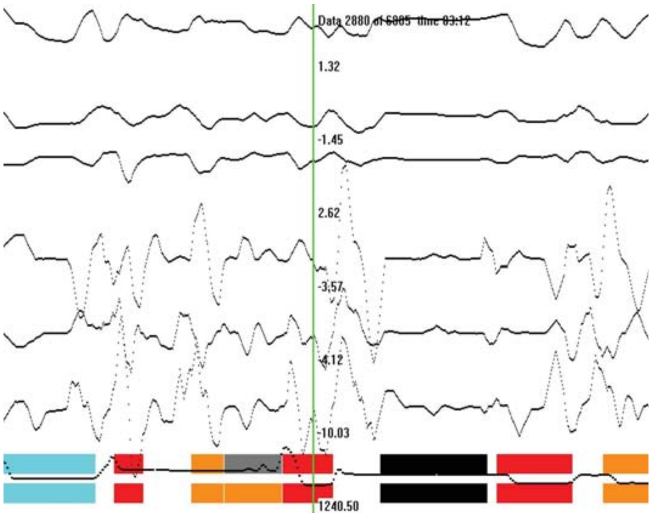


Figure 6. Example output and evaluation. The top row of boxes is ground truth labels of gestures and the bottom is classifier labels of gestures. Aqua = drink, red = bite, orange = utensiling, grey = other and black = rest.

Table IV presents the overall accuracy of the different classifiers. The accuracy of the baseline HMM is 77.8%. Our gender HMMs show slightly higher accuracy than baseline HMMs. We conclude that there is not much difference in wrist motion patterns during eating between females and males. Our age HMMs provided 4.3% accuracy higher than the baseline. This is further explored in table V, where it can be seen that accuracy is much lower for the age group 18-30 compared to all the older age groups. This suggests a larger variability in motion patterns in younger people. Finally, our utensil HMMs provided 6.2% higher accuracy than the baseline. This indicates that eating gestures with the same utensil share more similar gesture patterns across participants. Table VI explores this further, listing the accuracy for each utensil type. It can be seen that gestures while eating with chopsticks are recognized with very high accuracy, while hand and mixed utensil eating provides the greatest challenge to recognition.

Table VII shows the recognition result of each gesture by the different classifiers. The improvement provided by each

Classifier (HMMs)	Accuracy (%)
Baseline, generic HMM	77.8
Context, gender included	78.7
Context, age included	82.1
Context, utensil included	84.0

Table IV: Overall recognition accuracy.

Age group	Accuracy (%)
$[18, 30]$	**79.0**
$[31, 40]$	90.3
$[41, 50]$	89.3
$[51, 75]$	88.7

Table V: Recognition accuracy of age HMMs.

Utensil	Accuracy (%)
Fork	83.5
Spoon	87.5
Hand	82.2
Chopsticks	99.2
Mix-utensils	82.2

Table VI: Recognition accuracy of utensil HMMs.

Classifier (HMMs)	Rest (%)	Utensiling (%)	Bite (%)	Drink (%)	Other (%)
Baseline	80.0	79.9	81.4	94.3	54.5
Gender	80.7	81.2	83.0	94.5	54.9
Age	82.5	83.3	87.3	95.6	77.2
Utensil	82.9	85.2	87.8	98.8	68.1

Table VII: Recognition accuracy for five gestures.

contextual variable for each gesture type ranged from 0.2%-22.7%.

IV. Conclusion

This paper considers the problem of recognizing eating gestures by tracking wrist motion. We developed hidden Markov models to capture variations in motion patterns of subgroups of participants. Specifically, we examined if foreknowledge of the gender, age, and utensil used for eating could improve recognition accuracy. Improvement in accuracy was measured by comparing to a baseline HMM that was trained on all participants. Results show that gender HMMs performed slightly better than baseline, indicating that there is not much difference in wrist motion patterns during eating between females and males. Our age HMMs provided 4.3% accuracy higher than the baseline. The youngest group that ranges between 18 and 30 was found to have the lowest accuracy among four age groups, suggesting a larger variability in motion patterns in younger people. Our utensil HMMs provided 6.2% accuracy higher than the baseline. This suggests that eating gestures with the same utensil share some similarity across participants.

In previous work, our group demonstrated that the recognition of individual gestures could be improved through gesture-to-gesture sequential modeling [29]. For example, a common sequence is utensil-bite-rest. In the future we plan to combine the contextual approach described in this paper with sequential modeling. We also plan to explore additional contextual variables, such as groups of common food types.

Acknowledgment

We gratefully acknowledge the support of the NIH via grant 1R01HL118181-01A1.

References

[1] S. Kumar, W. Nilsen, A. Abernethy, A. Atienza, K. Patrick and others, "Mobile health technology evaluation: The mhealth evidence workshop," *American Journal of Preventive Medicine*, vol. 45, pp. 228–236, 2013.

[2] C. Ogden, M. Carroll, B. Kit, and K. Flegal, "Prevalence of obesity in the united states, 2009-201," *The Journal of the American Medical Association*, vol. 311, no. 8, pp. 806–814, 2014.

[3] L. Burke, J. Wang, and M. Sevick, "Self-Monitoring in Weight Loss: A Systematic Review of the Literature," *Journal of the American Dietetic Association*, vol. 111, pp. 92–102, 2011.

[4] A. Lang and E. Froelicher, "Management of Overweight and Obesity in Adults: Behavioral Intervention for Long-Term Weight Loss and Maintenance," *European Journal of Cardiovascular Nursing*, vol. 5, no. 2, pp. 102–114, 2006.

[5] D. Schoeller, "Limitations in the assessment of dietary energy intake by self-report," *Metabolism*, vol. 44, Supplement 2, pp. 18–22, Feb. 1995.

[6] J. Hins, F. Series, N. Almeras, and A. Tremblay, "Relationship between severity of nocturnal desaturation and adaptive thermogenesis: preliminary data of apneic patients tested in a whole-body indirect calorimetry chamber," *International Journal of Obesity*, vol. 30, pp. 574–577, 2006.

[7] O. Amft and G. Trster, "On-body sensing solutions for automatic dietary monitoring," *IEEE Pervasive Computing*, vol. 8, no. 2, pp. 62–70, 2009.

[8] E. Sazonov and S. Schuckers, "The energetics of obesity: A review: Monitoring energy intake and energy expenditure in humans," *IEEE Engineering in Medicine and Biology Magazine*, vol. 29, no. 1, pp. 31–35, 2010.

[9] O. Amft, "A wearable earpad sensor for chewing monitoring," in *2010 IEEE Sensors*, Kona, HI, 2010, pp. 222–227.

[10] J. Liu, E. Johns, L. Atallah, C. Pettitt, B. Lo, G. Frost, and G. Yang, "An intelligent food intake monitoring system using wearable sensors," in *2012 Ninth International Conference on Wearable and Implantable Body Sensor Networks*, May 2012, pp. 154–160.

[11] J. Nishimura and T. Kuroda, "Eating habits monitoring using wireless wearable in-ear microphone," in *3rd International Symposium on Wireless Pervasive Computing*, May 2008, pp. 130–132.

[12] S. Pler, M. Wolff, and W. Fischer, "Food intake monitoring: an acoustical approach to automated food intake activity detection and classification of consumed food," *Physiological Measurement*, vol. 33, no. 6, pp. 1073–1093, 2012.

[13] M. Shuzo, S. Komori, T. Takashima, G. Lopez, S. Tatsut and others, "Wearable eating habit sensing system using internal body sound," *Journal of Advanced Mechanical Design, Systems, and Manufacturing*, vol. 4, pp. 158–166, 2010.

[14] J. Fontana and E. Sazonov, "A robust classification scheme for detection of food intake through non-invasive monitoring of chewing," in *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2012, pp. 4891–4894.

[15] E. Sazonov, S. Schuckers, P. Lopez-Meyer, O. Makeyev, N. Sazonova, E. Melanson, and M. Neuman, "Non-invasive monitoring of chewing and swallowing for objective quantification of ingestive behavior," *Physiological Measurement*, vol. 29, no. 5, p. 525, 2008.

[16] E. Sazonov and J. Fontana, "A sensor system for automatic detection of food intake through non-invasive monitoring of chewing," *IEEE Sensors Journal*, vol. 12, no. 5, pp. 1340–1348, 2012.

[17] P. Lopez-Meyer, O. Makeyev, S. Schuckers, E. Melanson, M. Neuman, and E. Sazonov, "Detection of food intake from swallowing sequences by supervised and unsupervised methods," *Annals of Biomedical Engineering*, vol. 38, no. 8, pp. 2766–2774, 2010.

[18] E. Sazonov, O. Makeyev, S. Schuckers, P. Lopez-Meyer, E. Melanson, and M. Neuman, “Automatic detection of swallowing events by acoustical means for applications of monitoring of ingestive behavior,” *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 3, pp. 626–633, 2010.

[19] W. Walker and D. Bhatia, “Towards automated ingestion detection: Swallow sounds,” in *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Sep. 2011, pp. 7075–7078.

[20] O. Amft, H. Junker, and G. Trster, “Detection of eating and drinking arm gestures using inertial body-worn sensors,” in *IEEE Proceedings of the Ninth International Symposium on Wearable Computers*, Oct. 2005, pp. 160–163.

[21] O. Amft and G. Trster, “Recognition of dietary activity events using on-body sensors,” *Artificial Intelligence in Medicine*, vol. 42, no. 2, pp. 121–136, 2008.

[22] H. Junker, O. Amft, P. Lukowicz, and G. Troster, “Gesture spotting with body-worn inertial sensors to detect user activities,” *Pattern Recognition*, vol. 41, no. 6, pp. 2010–2024, 2008.

[23] Y. Dong, A. Hoover, and E. Muth, “A device for detecting and counting bites of food taken by a person during eating,” in *IEEE Intl Conf on Bioinformatics and Biomedicine*, Washington, Nov. 2009, pp. 265–268.

[24] Y. Dong, A. Hoover, E. Muth, and J. Scisco, “A new method for measuring meal intake in humans via automated wrist motion tracking,” *Applied Psychophysiology and Biofeedback*, vol. 37, no. 3, pp. 205–215, 2012.

[25] R. Ramos, E. Muth, J. Gowdy and A. Hoover, “Improving the recognition of eating gestures using inter-gesture sequential dependencies,” *IEEE Journal of Biomedical and Health Informatics*, vol. 19, no. 3, pp. 825–831, May 2015.

[26] E. Thomaz, I. Essa, and G. Abowd, “A practical approach for recognizing eating moments with wrist-mounted inertial sensing,” in *the proc. of Ubicomp*, 2015.

[27] Z. Huang, “An assessment of the accuracy of an automated bite counting method in a cafeteria setting,” Master’s thesis, Clemson University, Clemson, Aug. 2013.

[28] Y. Dong, “Tracking Wrist Motion to Detect and Measure the Eating Intake of Free-living Humans,” Ph.D. dissertation, Univ. of Clemson, Clemson, May 2012. [Online]. Available: http://www.ces.clemson.edu/ ahoover/theses/dong-diss.pdf

[29] R. Ramos, E. Muth, J. Gowdy, and A. Hoover, “Improving the recognition of eating gestures using intergesture sequential dependencies,” *IEEE Journal of Biomedical and Health Informatics*, vol. 19, pp. 825–831, 2015.

[30] R. Ramos, “Using Hidden Markov Models to Segment and Classify Wrist Motions Related to Eating Activities,” Ph.D. dissertation, Univ. of Clemson, Clemson, May 2014. [Online]. Available: http://www.ces.clemson.edu/ ahoover/theses/ramos-diss.pdf

[31] L. Rabiner and M. Hill, “A tutorial on hidden markov models and selected applications in speech recognition,” *Proceedings of the IEEE*, vol. 77, pp. 257–286, 1989.

[32] G. Xuan, W. Zhang, and P. Chai, “Em algorithms of gaussian mixture model and hidden markov model,” in *International Conference on Imgage Processing*, Thessaloniki, Oct. 2001.

[33] K. Murphy. Hidden markov model (hmm) toolbox for matlab. [Online]. Available: http://www.cs.ubc.ca/ murphyk/Software/HMM/hmm.html