Segmentation and Recognition of Eating Gestures from Wrist Motion using Deep Learning

Yadnyesh Y. Luktuke
Department of Electrical & Computer Engineering
Clemson University
Clemson, SC, U.S.A.
yluktuk@g.clemson.edu

Adam Hoover
Department of Electrical & Computer Engineering
Clemson University
Clemson, SC, U.S.A.
ahoover@clemson.edu

***Abstract*—This paper describes a novel approach of segmenting and classifying eating gestures from wrist motion using a deep learning neural network. It is inspired by the approach of fully-convolutional neural networks in the task of image segmentation. Our idea is to segment 1D gestures the same way 2D image regions are segmented, by treating each inertial measurement unit (IMU) datum like a pixel. The novelty of our approach lies in training a neural network to recognize data points that describe an eating gesture just like it would be trained to recognize pixels describing an image region. The data for this research is known as the Clemson Cafeteria Dataset. It was collected from 276 participants that ate an unscripted meal at the Harcombe Dining Hall at Clemson University. Each meal consisted of 1 - 4 courses, and 488 such recordings were used for the experiments described in this paper. Sensor readings consist of measurements taken by an accelerometer (x, y, z) and a gyroscope (yaw, pitch, roll). A total of 51,614 unique gestures associated with different activities commonly seen during a meal were identified by 18 trained raters. Our neural network classifier recognized an average of 79.7% of 'bite' and 84.7% of 'drink' gestures correctly per meal. Overall 77.7% of all gestures were recognized correctly on average per meal. This indicates that a deep learning model can successfully be used to segment eating gestures from a time series recording of IMU data using a technique similar to pixel segmentation within an image.**

***Index Terms*—Deep learning, eating gestures, energy intake, IMU sensors, segmentation.**

I. Introduction

This paper describes a novel approach of using a deep learning classifier for segmenting eating gestures from wrist-motion. These are associated with upper limb motion of short duration during activities such as moving food from the plate towards the mouth, taking a drink from a cup, stirring soup and cutting food into bite sized pieces. Tracking such activity can prove useful in estimating calorie intake in humans.

This research is motivated by the rise of obesity. In the United States of America, from 1999-2000 to 2017-2018 the prevalence of obesity increased from 30.5% to 42.4% [8] and at least 20% of the adult population in each state considers themselves overweight or obese [9]. Worldwide it affects people of all ages [7] and can often lean to serious conditions such as certain types of cancer, cardiovascular diseases, diabetes and even premature death [6], [8]. As per the World Health Organization (WHO) monitoring energy intake and expenditure can promote people to take healthier life choices and thus manage obesity [7]. However the former has received relatively less interest, often being limited to self-reporting and 24-hour recalls such as those described in [10] and [11]. These are tedious and time-consuming, and often lead to non-compliance over long periods of time [2].

Different sensing modalities have been studied to measure consumption [19] including acoustic sensors that detect chewing or swallowing sounds within the ear canal or around the throat region [20], [21], [22], [23], camera sensors that estimate the 3D volume of food [2] or serve as retrospective memory aids in 24-hour recalls [24], [25], [26] and smart eyeglasses that track activity in the temporalis muscle (associated with mastication) using an electromyography (EMG) sensor alone [27] or one integrated with an accelerometer [28], [29].

In comparison to these inertial measurement unit (IMU) sensors that are fitted in wrist worn devices such as smartwatches offer a convenient, reliable and comfortable way of monitoring eating activity as discussed in [2] and [5]. Our group has been studying the recognition of eating activities using IMU sensors for 10 years. The original algorithm known as the 'Bite Counter' [12], [13], [14] detects specific patterns of wrist motion associated with the intake of a single bite of food using a set of heuristics and thresholds. More recent work has focused on using hidden Markov models (HMM) to classify wrist activity into a fixed number of categories using inter-gesture sequential dependencies [17]. This method achieves 96.5% accuracy at detecting eating gestures from a data set of 25 meals eaten by different subjects. This was extended to three main variations of HMM for studying the effect of contextual variables such as age, gender and ethnicity in [2] and [18].

Another group of researchers detected food intake from IMU data recorded using commercial smartwatches as reported in [30]. They modeled eating activity as a combination of five specific wrist micromovements or micro gestures that were detected using a deep learning neural network. Using convolutional layers to learn the probability distribution of each micromovement, which is then fed into long short-term memory (LSTM) layers their model detects sequences containing food intake cycles. It achieves the highest F1 detection score of 0.913 in a leave-one-out crossvalidation approach, when compared to other state-of-the-art methods

978-1-7281-6251-5/20/$31.00 ©2020 IEEE

including the one in [13]. This is motivating since it suggests that a deep learning classifier can be used to detect eating activity from an IMU time-series recording.

In [2], [17] and [30] the authors considered IMU recordings that were manually segmented. In addition the data used in [30] only contains recordings of people eating using forks and knives. Other utensils and eating with the hands are not considered at all, neither is activity such as drinking which often occurs along with eating during a meal. On the other hand, the data used in this research [1] contains recordings of multiple gesture types including drink, and also contains recordings of people eating with a variety of utensils and their hands as well. The neural network proposed as part of this research builds on the works reported in [2], [17] and [30], but extends to data that is unsegmented and contains recordings for different gesture types using a variety of utensils. It is similar to the approach for image segmentation as seen in [3] and [4]. In these 2D pixels are spatially grouped and labeled as belonging to a particular region [3], [31], [32]. In a similar manner we treat a single datum recorded using IMU sensors as a part of an eating gesture and detect multiple consecutive instances that form a complete gesture. Our approach is novel since it trains a neural network to simultaneously detect periods of specific wrist motion, and classify these as eating gestures accordingly. The rest of this paper describes this idea in greater detail.

II. Methods

A. Data

The data used for this research is part of the Clemson Cafeteria Dataset [1]. It was collected from 276 participants that each ate a single meal consisting of 1-4 courses at the Harcombe Dining Hall at Clemson University. A total of 380 different food and beverage types were considered including stir fried vegetables, shoestring French fries, pasta, water and soda. Four different utensils were considered, viz. forks, spoons, chopsticks and hands. The group consisted of 131 male and 140 female participants, of which 114 were in the age group 18 - 23, 76 in 24 - 30, 27 in 31 - 40, 33 in 41 - 50 and 21 were between 51 - 75 years old.

Wrist motion was recorded using a custom device that measured wrist-acceleration using an accelerometer (x, y, z) and wrist-rotation using a gyroscope (yaw, pitch, roll) at 15 Hz. In previous works [2], [17] our group had identified five unique gesture categories including ‘bite’, ‘drink’, ‘utensiling’, ‘rest’ and ‘other’. A total of 18 trained raters observed a synchronized video feed and the measured signals simultaneously to identify gestures in each recording as shown in Figure 1. In this figure, the green line indicates the current position of the recording while gestures are identified using a color code; red for ‘bite’, aqua for ‘drink’, ‘orange’ for utensiling, ‘black’ for rest and gray for ‘other’. A total of 51,614 unique gestures, each of unequal duration were identified by the raters. The category ‘unlabeled’ (shown in plain white) was identified in this research to mark all instances of time not having a unique label in the ground truth. This is necessary to train

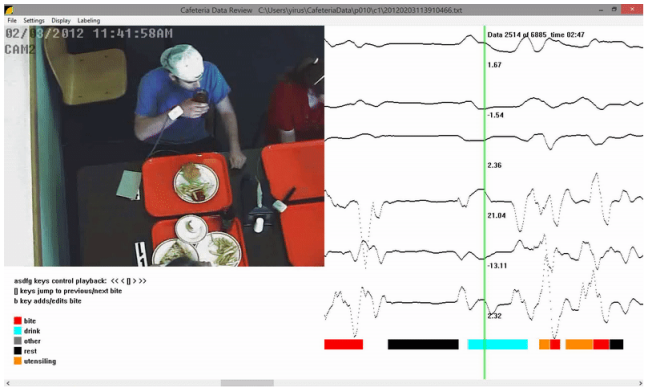

Fig. 1: Custom tool for observing video feed and recorded signals simultaneously. From top to bottom signals are: accelerometer (x, y, z) and gyroscope (yaw, pitch, roll). Gesture labels are red: bite, aqua: drink, gray: other, black: rest, orange: utensiling.

a neural network for segmenting eating gestures from within each recording of IMU data. For a detailed description of the labeling process, including the complete gesture definitions the interested reader is referred to [2] and [5].

B. Deep learning neural network

The neural network architecture developed as part of this research is shown in Figure 2. It consists of three convolutional blocks that form the encoder stage of the network, and three deconvolutional blocks as part of the decoder stage. It is inspired by the U-Net models used for image segmentation in [3] and [4]. In these models the convolutional blocks learn a set of filters for transforming the input into its feature-space representation. This stage extracts the high resolution information from the input. At the end of each encoder block the output is downsampled using max-pooling to retain only the strongest responses to the learned filters in that block. In contrast, each decoder block combines the output of its preceding decoder block with that from an appropriate encoder block and increases its resolution through a process known as deconvolution [15], [16]. The resolution is steadily increased till it is the same as that of the input. This structure helps a U-Net model achieve high contextual accuracy through the information extracted in the encoding phase and high localization accuracy through the decoding phase [3]. For a detailed description of each encoder and decoder block, the reader is referred to the author’s earlier work [5].

In order to train the neural network on input sequences of arbitrary length, each recording was separated into multiple consecutive and overlapping segments known as sliding windows. Each window was of 30 seconds duration, corresponding to 450 samples due to the sampling rate of 15 Hz. In each iteration the window was shifted by 1 second or 15 samples. This process continued till the last sliding window fit within the recording without zero-padding the input sequence. Further each sliding window was reshaped from size 450×6 to an array of size $1 \times 450 \times 6$ as shown in Figure 3. This means

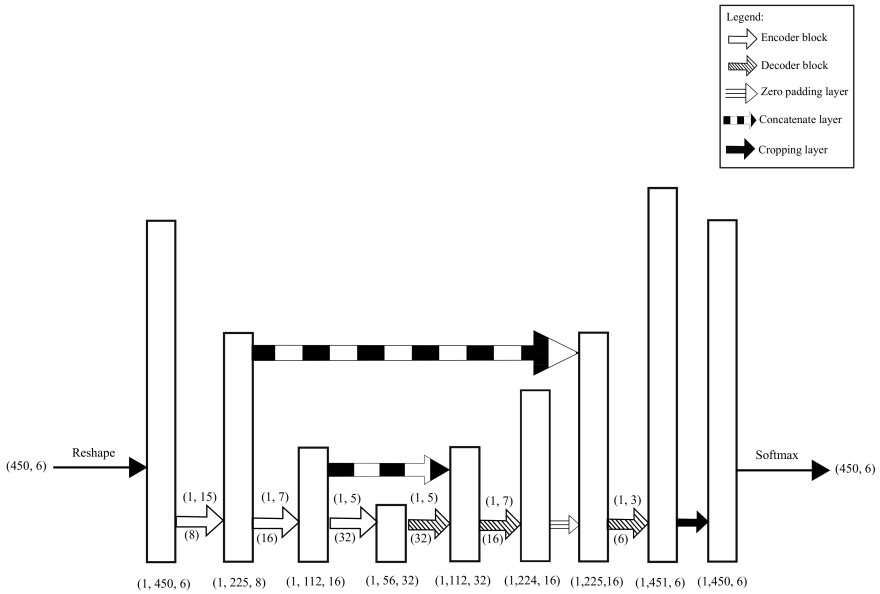

Fig. 2: Encoder-decoder architecture: Arrow blocks show filter size (top) and layer depth (bottom) in that particular block. Other numbers indicate the size of the output after previous operation. Concatenate layers merge connected arrays.

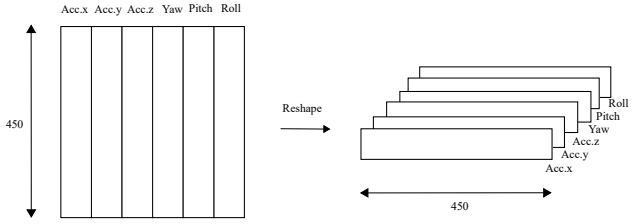

Fig. 3: Reshaping input for compatibility with the neural network.

that each datum within an IMU time-series recording is treated like a pixel within an image. This is the true novelty of this research. By treating each datum like a pixel within an image, a gesture is then similar to a region within an image that needs to be segmented as demonstrated in [3] and [4]. The neural network thus learns to classify and group all such segments corresponding to the eating gestures from each recording. All data was normalized using min-max normalization for an entire meal before being used to train the neural network.

C. Training the neural network

The neural network was trained to minimize the categorical cross-entropy loss between the model prediction for each datum and its ground truth label. Each ground truth label was converted into its one-hot representation as explained in [5]. The model was trained using a 5-fold cross validation, with the number of epochs fixed at 200 during each training fold. This number was empirically determined after observing that the model performance improved steadily after 100 epochs but showed no improvement beyond 200 epochs. The training process including the one-hot representation is described in greater detail in [5].

D. Model output and evaluation

Due to the softmax activation used in the final decoder block the neural network output for each datum is a probability distribution over the class labels [4]. The class having the highest value is then retained as the final prediction for each datum. For data points that occur in more than one sliding window the max-voting strategy was used to retain the label that occurred most frequently as explained in [5]. For data points that have ambiguous labels the label of the preceding datum was used as the final model output instead. Such data points were observed at the boundary of two gestures owing to the differences in temporal relationships between such data points and their neighbors in consecutive sliding windows.

Once the final model output was generated the gestures were compared with the ground truth labels using a method similar to the one used to measure inter-rater reliability in [2]. In [2] inter-rater reliability was used to resolve discrepancies between two ground truth labels made by two different human raters. Instead in this research we assume that the ground truth labels are correct, and only the classifier output needed to be evaluated in order to assess its accuracy. Model accuracy was assessed in three stages, first by counting the agreement between indices of the ground truth and the model output, then by measuring total agreement between gestures in the ground truth and model output for the entire recording and finally by evaluating the agreement between different gesture types. The category ‘unlabeled’ was not considered during the evaluation phase, nor were the gestures that occurred outside the start and end of the ground truth gestures.

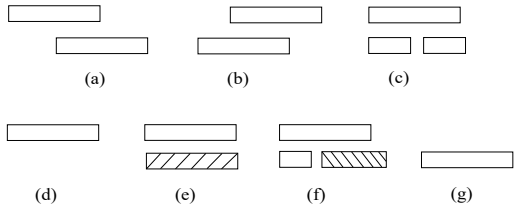

Fig. 4: Comparing ground truth (top) and classifier output (bottom). Top row, (a), (b) and (c) represent agreement between gestures. In the bottom row we see, from left to right: missed (d), mislabeled (e), mangled (f) and false positive (g).

Agreement among indices was evaluated using a percentage of matching between the model output and the ground truth. A high percentage of agreement between indices indicates that the model output closely resembles the ground truth label for a particular recording. In the later stages agreement between all gestures in a meal was evaluated using the different types of overlap between model output and ground truth as shown in Figure 4. As seen in parts (a), (b) and (c) the model output has more than 50% overlap with the ground truth gestures along with the correct label and is hence considered as agreement between gestures. On the other hand in parts (d), (e), (f) and (g) the model misses the ground truth gesture, mislabels its output, identifies at least one wrong gesture category and falsely identifies a gesture where none exists in the ground truth file. Thus these overlaps are treated as erroneous model output, termed as missed, mislabeled, mangled and false positive respectively. It should be clear to the reader that classifier outputs from (d) to (g) are undesired as they indicate that the model was unable to identify that particular ground truth gesture.

III. Results

A. Correctly identifying indices

The model correctly matched 71.3% of indices on average per meal between the ground truth and the classifier output. The average standard deviation per meal of this matching was 10.8%, indicating that the model output closely resembles the ground truth at the index level on average per meal.

B. Correctly identifying all gestures from a meal

As it can be seen from Table I, the model correctly identified 77.7% of all gestures on average per meal which is sufficiently high for the chosen application. In addition the average standard deviation of correctly identified gestures per meal of 13.7% is reasonably low, while the average standard deviation of all incorrect mappings per meal is even lower. This indicates that the neural network is robust to differences in wrist motion for the large group of people considered and can be used to segment eating gestures from each recording.

C. Identifying eating related gestures from a meal

The model performs well at identifying eating-related gestures such as ‘bite’ and ‘drink’, and other gestures associated with eating such as ‘utensiling’ and ‘rest’ as seen in Table II. The average accuracy per meal for each of these gesture types is close to 80%, and the average standard deviation per meal is close to 20% in each case. However we also observe that the model was unable to identify any gestures from the ‘other’ category. This gesture was used to mark all activities that could not be categorized in the other four main categories, including periods of ambiguous behavior [2]. Hence it contains a lot of variation, especially for the large group of people considered. In addition it was observed that this gesture occurred very infrequently in only 123 meals out of 488 in the data set. Hence the data set is imbalanced with respect to the ‘other’ category of gestures. As machine learning models such as neural networks do not perform well on such type of data, we can expect that the model was unable to identify even a single gesture from this particular category.

TABLE I: Percentage of inter-gesture mapping per meal.

Metric	correct	missed	mislabeled	mangled	false positive
Avg.	77.7	11.2	5.9	6.2	16.6
St. dev.	13.7	8.4	5.8	4.9	11.1

TABLE II: Percentage of correctly identified gestures per meal by category.

Metric	bite	drink	utensiling	rest	other
Avg.	79.7	84.7	79.5	81.1	0
St. dev.	19.1	20.3	17.3	17.6	0

Figure 5 displays a typical result for one segment of the meal 215/c3. The image is plotted using CafeView, the custom tool designed to compare model output and ground truth for each file. For this meal the percentage of individual gestures recognized is 88.4%, 88.8%, 71.4% and 80% for ‘bite’, ‘drink’, ‘utensiling’ and ‘rest’ respectively, which indicates a high degree of matching with the ground truth. Note that this recording contained no gestures labeled as ‘other’ in the ground truth and classifier output as well, and hence this category is not mentioned or displayed.

On the other hand the model performed very poorly on

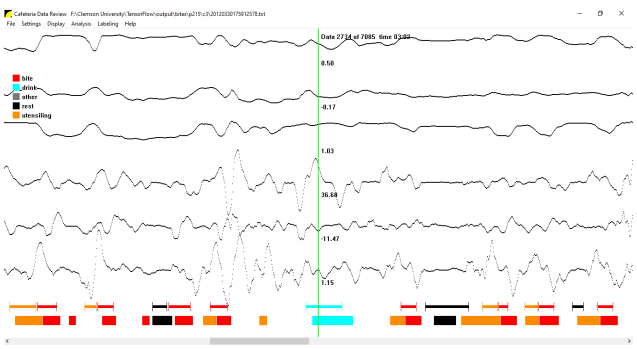

Fig. 5: An example typical result shown for the meal 215/c3 with ground truth (top) compared against model output (bottom). Gesture labels are red: bite, aqua: drink, gray: other, black: rest, orange: utensiling.

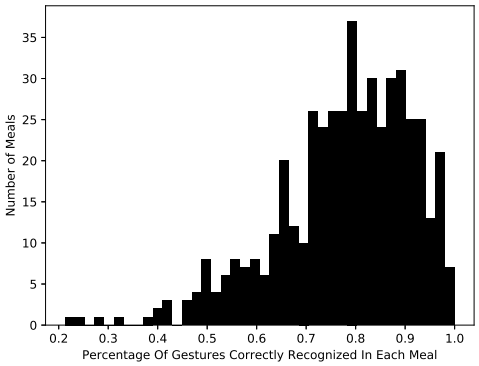

Fig. 6: Plotting histogram to identify meals with lower total gesture agreement.

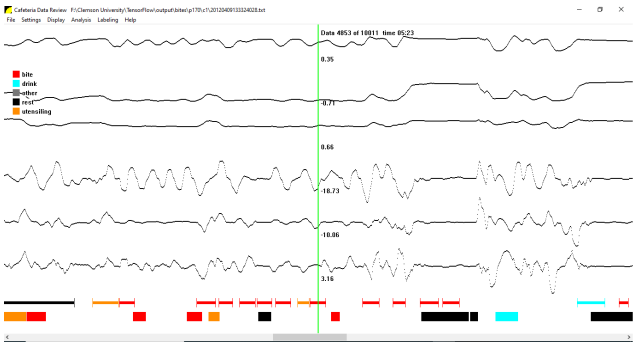

Fig. 7: A meal containing multiple gestures marked for the uninstrumented hand, corresponding to the recording 170/c1. Ground truth (top) and classifier output (bottom) are identified using labels are red: bite, aqua: drink, gray: other, black: rest, orange: utensiling.

some meals within the data set. On observing the distribution of correctly identified gestures per meal as a histogram in Figure 6, we see that the distribution is long tailed falling sharply beyond 3 standard deviations away from the mean. One such segment corresponding to the recording `170/c1` is shown in Figure 7. In this recording the percentage of 'bite' and 'drink' gestures correctly recognized was 18.5% and 50% respectively, while the percentage of total gestures correctly identified was 37.5%. On observing the third and fourth 'bite' gestures to the right of the green marker within this segment we see that the sensor did not record any activity for the duration of the gesture marked by the rater. It is very likely that the rater did in fact observe eating activity, but failed to notice that it occurred with the uninstrumented hand, which did not have the recording device mounted. Other meals on which the model performed poorly also contain multiple gestures accidentally marked for the uninstrumented hand by the rater. These were classified as 'rest' by the model owing to the relative inactivity as measured by the electronic sensors. This is thought to be the main reason for the mismatch between the classifier output and the ground truth gesture for these meals. However since the main goal of this research is to detect and classify all gestures associated with eating activity, it is important that the classifier identify such periods of wrist motion as well. It is known from the principle of symmetry in biology that motion in one arm/wrist/hand tends to cause related motions in the other arm/wrist/hand. Hence a strategy is discussed in section IV which can potentially help design a classifier that can detect such periods of associated motion in the instrumented hand.

IV. Conclusion

This research considers the problem of designing and implementing a deep learning model for simultaneously detecting and classifying periods of wrist motion into specific eating related categories or gestures. Wrist motion data such as the one considered in this research can be recorded using IMU sensors fitted inside a watch-like device such as a smartwatch or a fitness tracker. This works builds on HMM based classifiers, especially those designed previously in our research group [2], [17] and extends their application to data that is previously unsegmented. The deep learning model in particular is inspired by the success of similar approaches used for image segmentation, such as [3], [4], [31] and [32].

The deep learning model was able to successfully segment and classify 77.7% of all gestures on average per meal. It was also able to detect and identify individual gestures including 'bite','drink','utensiling' and 'rest' with an average accuracy of 79.7%, 84.7%, 79.5% and 81.1% respectively per meal. However it was unable to identify gestures belonging to the category 'other'. This is most likely because this category contains a lot of variation in the recorded signals, especially for the large group of people considered. It is also the most infrequently occurring category, as only 123 meals have ground truth gestures marked as 'other' and the presence of these gestures is limited in these meals as well. Thus the data set is very imbalanced with respect to this category, and hence the classifier performs poorly at detecting these gestures.

It was also observed that the neural network performed very poorly at identifying gestures correctly in some meals. These meals contain multiple periods of activity marked for the uninstrumented hand by the human rater as seen in Figure 7. This is thought to be the main reason for the poor performance of the neural network on such meals. It is however known from the principle of symmetry in biology that motion in one wrist tends to cause related motion in the other. Hence it is possible that a classifier can be designed that can detect motion in an instrumented hand that occurs relative to eating gestures by the other hand. This is discussed in section IV-A.

A. Future work

One way to improve the accuracy of the neural network at identifying gestures in the instrumented hand is to consider a deeper neural network consisting of more convolutional and deconvolutional blocks. It is expected that a larger number of blocks will generate better feature mappings corresponding to

wrist-micromovements that occur during eating and non-eating gestures [30]. This might help to capture subtle movements in the instrumented hand that occur relative to specific motion in the uninstrumented hand when the subject eats with the wrong hand. These may also prove useful to improve the accuracy of the classifier for the 'other' category of gestures thus improving the overall gesture recognition accuracy of the model.

Another way to improve the accuracy of the classifier is by expanding the set of ground truth labels considered. This can be done by including the hand with which the gesture occurred, which would improve the contextual meaning of the ground truth much like the work done in [2]. This is expected to improve the accuracy of the classifier at detecting gestures in both hands as well.

Finally considering sliding windows having different lengths of time can also be considered to improve the accuracy of the model. Longer sequences than the one considered in this research would mean more temporal relationships would be seen by the neural network, thereby improving the feature-mappings generated. However these would require larger memory for training the neural network in practice and longer training times as well. Hence the trade-off between the length of the sliding window and the accuracy of the neural ntwork needs to be carefully studied.

REFERENCES

[1] A. Hoover, "Data description: Clemson Cafeteria Dataset," Online, *URL: http://cecas.clemson.edu/ ahoover/cafeteria/*, 2020.

[2] Y. Shen, "Using contextual information to improve hidden Markov model recognition of wrist motions during eating activities," PhD. Thesis, Clemson University, December 2018.

[3] O. Ronneberger, P. Fischer and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2015, pp. 234–241.

[4] TensorFlow, "TensorFlow core - Tutorials: Image segmentation", Online, *URL: https://www.tensorflow.org/tutorials/images/segmentation*, 2020.

[5] Y.Y. Luktuke, "Segmentation and recognition of eating gestures from wrist motion using deep learning," MS. Thesis, Clemson University, May 2020.

[6] World Health Organization, "WHO - Health topics: Obesity", Online, *URL: https://www.who.int/topics/obesity/en*, 2020.

[7] World Health Organization, "WHO Newsroom, Fact sheets - Detail: Obesity and overweight", Online, *URL: https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight*, 2020.

[8] Centers for Disease Control and Prevention, "Adult Obesity Facts", Online, *URL: https://www.cdc.gov/obesity/data/adult.html*, 2020.

[9] Centers for Disease Control and Prevention, "Adult Obesity Prevalence Maps", Online, *URL: https://www.cdc.gov/obesity/data/prevalence-maps.html*, 2019.

[10] 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, Nature Publishing Group, vol. 30, no. 3, 2006, pp. 574–577.

[11] D.A. Schoeller, "Limitations in the assessment of dietary energy intake by self-report," Metabolism, Elsevier, vol. 44, 1995, pp. 18–22.

[12] Y. Dong, A. Hoover and E. Muth, "A device for detecting and counting bites of food taken by a person during eating," 2009 IEEE International Conference on Bioinformatics and Biomedicine, 2009, pp. 265–268.

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

[14] Y. Dong, J. Scisco, M. Wilson, E. Muth and A. Hoover, "Detecting periods of eating during free-living by tracking wrist motion," IEEE Journal of Biomedical and Health Informatics, vol. 18, no. 4, 2013, pp. 1253–1260.

[15] M.D. Zeiler, D. Krishnan, G.W. Taylor and R. Fergus,"Deconvolutional networks," 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2010, pp. 2528–2535.

[16] P. Louis Pröve, "An introduction to different Types of convolutions in deep learning," Online *URL: https://towardsdatascience.com/types-of-convolutions-in-deep-learning-717013397f4d*, 2017.

[17] R.I. Ramos-Garcia, E.R. Muth, J.N. Gowdy and A.W. Hoover, "Improving the recognition of eating gestures using intergesture sequential dependencies," IEEE Journal of Biomedical and Health Informatics, vol. 19, no. 3, 2014, pp. 825–831.

[18] Y. Shen, J. Salley, E. Muth and A. Hoover, "Assessing the accuracy of a wrist motion tracking method for counting bites across demographic and food variables," IEEE Journal of Biomedical and Health Informatics, vol. 21, no. 3, 2016, pp. 599–606.

[19] T. Prioleau, E. Moore II and M. Ghovanloo, "Unobtrusive and wearable systems for automatic dietary monitoring," IEEE Transactions on Biomedical Engineering, vol. 64, no. 9, 2017, pp. 2075–2089.

[20] S.Päßler and W. Fischer, "Food intake monitoring: Automated chew event detection in chewing sounds," IEEE Journal of Biomedical and Health Informatics, vol. 18, no. 1, 2013, pp. 278–289.

[21] S. Päßler, W. Fischer and I. Kraljevski, "Adaptation of models for food intake sound recognition using maximum a posteriori estimation algorithm," 2012 Ninth International Conference on Wearable and Implantable Body Sensor Networks, IEEE, 2012, pp. 148–153.

[22] Y. Gao, N. Zhang, H. Wang, X. Ding, X. Ye, G. Chen and Y. Cao, "iHear food: eating detection using commodity bluetooth headsets," 2016 IEEE First International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE), 2016, pp. 163–172.

[23] O. Amft, M. Stäger, P. Lukowicz and T. Tröster, "Analysis of chewing sounds for dietary monitoring," International Conference on Ubiquitous Computing, Springer, 2005, pp. 56–72.

[24] S. Hodges, L. Williams, E. Berry, S. Izadi, J. Srinivasan, A. Butler, G. Smith, N. Kapur and K. Wood, "SenseCam: A retrospective memory aid," International Conference on Ubiquitous Computing, Springer, 2006, pp. 177–193.

[25] L. Gemming, A. Doherty, J. Utter, E. Shields and C.N. Mhurchu, "The use of a wearable camera to capture and categorise the environmental and social context of self-identified eating episodes," Appetite, Elsevier, 2015, vol. 92, pp. 118–125.

[26] M. Sun, L.E. Burke, Z. Mao, Y. Chen, H. Chen, Y. Bai, Y. Li, C. Li and W. Jia, "eButton: a wearable computer for health monitoring and personal assistance," Proceedings of the 51st Annual Design Automation Conference, 2014, pp. 1–6.

[27] Q. Huang, W. Wang and Q. Zhang, "Your glasses know your diet: Dietary monitoring using electromyography sensors," IEEE Internet of Things Journal, 2017, vol. 4, no. 3, pp. 705–712.

[28] R. Zhang and O. Amft, "Bite glasses: measuring chewing using emg and bone vibration in smart eyeglasses," Proceedings of the 2016 ACM International Symposium on Wearable Computers, 2016, pp. 50–52.

[29] R. Zhang and O. Amft, "Monitoring chewing and eating in free-living using smart eyeglasses," IEEE Journal of Biomedical and Health Informatics, 2017, vol. 22, no. 1, pp. 23–32.

[30] K. Kyritsis, C. Diou and A. Delopoulos, "Modeling wrist micromovements to measure in-meal eating behavior from inertial sensor data," IEEE Journal of Biomedical and Health Informatics, 2019, vol. 23, no. 6, pp. 2325–2334.

[31] V. Badrinarayanan, A. Kendall and R. Cipolla "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, vol. 39, no. 12, pp. 2481–2495.

[32] G. Wang, W. Li, M.A. Zuluaga, R. Pratt, P.A. Patel, M. Aertsen, T. Doel, A.L. David, J. Deprest, S. Ourselin and others, "Interactive medical image segmentation using deep learning with image-specific fine tuning," IEEE Transactions on Medical Imaging, 2018, vol. 37, no. 7, pp. 1562–1573.