Monitoring Brushing behaviors using Toothbrush Embedded Motion-Sensors

Mahmoud Essalat, Oscar Hernan Madrid Padilla, Vivek Shetty, Gregory Pottie

***Abstract*—Dental disease is largely preventable and closely linked to poor toothbrushing behaviors. Motion-sensors, such as accelerometers, gyroscopes, and magnetometers, allow for monitoring of toothbrushing behaviors. Researchers have attempted to infer tooth surface coverage using sensors attached to the toothbrush handle or embedded in smartwatches. However, the inferences may be deficient because the datasets were collected under structured toothbrushing assumptions performed in controlled laboratory settings and not the free-form and irregular brushing patterns observed in real-world settings. To address the aforementioned problem, we collected a dataset of 187 brushing sessions, including free-form brushing. We present, to our knowledge, the first motion-sensor dataset obtained during free-form brushing. Using our experiences, we discuss the challenges of studying toothbrushing behaviors in naturalistic settings. We introduce a meticulous labeling method to ensure the high accuracy in our data annotation process. We also propose a three-stage method (i.e. pre-processing, brush transition time detection, and time-series classification) to detect the teeth surfaces brushed during a session. Our findings are two-fold: (a) the classification of teeth surfaces during free-form toothbrushing is more challenging than during brushing in controlled settings; (b) high classification accuracy can be achieved using random train-test split of the data (i.e. k-fold cross-validation); however, generalization beyond the participants in the training set poses difficulties. Beyond publishing the first dataset of free-form toothbrushing, we validate our findings by applying our proposed method to our provided dataset, as well as the datasets of toothbrushing in controlled settings.**

***Index Terms*—classification, motion-sensors, change point detection, brushing behavior, smart toothbrush**

GIVEN the ubiquity of dental disease and its consequences, dental care providers emphasize preventive oral self-care that can maintain oral health in the home setting [2]. A particular focus is the removal of dental plaque, a sticky film of bacteria and food debris that covers the teeth, and causes periodontal (gum) disease and caries [3]. If not removed regularly and adequately, plaque produces acids that break down tooth enamel and irritate the gums, leading to tooth and gum disease and eventual loss of teeth [4], [5].

Brushing one's teeth regularly is the primary means of removing plaque from the tooth surfaces; the brushing technique needs to be consistent and thorough to ensure that plaque is removed from all tooth surfaces [6], [7]. Yet, oral self-care behaviors are rarely performed proficiently as recommended by care providers; individuals frequently skip sessions entirely or devote most of the time to brushing the outer teeth surfaces rather than the inner surfaces [8]. Thus, monitoring the frequency and quality of toothbrushing in the home setting is important for providing corrective feedback on actual brushing behaviors and eventually, decreasing the incidence and impact of dental disease.

The hyper-connectivity presented by the now ubiquitous mobile devices and a growing confluence with behavioral science have stimulated the use of digital approaches to reinforce healthy behaviors and self-care (digital health interventions or DHIs) on a large scale [9]. In many applications, low-cost sensors allow unobtrusive monitoring to provide a realistic picture of a patient's behavior in real-world settings [10], [11].

Researchers have proposed a wide variety of models to monitor activities of daily living (ADL) that include hand washing [12], flossing [13], walking, eating or drinking, cellphone use, and functional transfers, such as standing up from a chair, and lying down on bed [14], [15], [16]. These activity recognition systems can be used to deliver timely interventions for behavioral changes such as smoking cessation [17], [18], [19] or control of overeating [20]. However, these monitoring systems are specific to the given application; methods specific for a certain ADL are not generalizable to other areas of activity recognition. This issue is manifest in the large diversity of approaches for different activity recognition.

Monitoring toothbrushing behaviors differs significantly from other activity recognitions that depend on macro-motion patterns that are large and easily observable (e.g. walking, standing, and running). Toothbrushing activities, such as brushing adjoining tooth surfaces, comprise of micro-motions and the subtle differences in acceleration patterns captured by toothbrush sensors pose significant challenges for accurately inferring toothbrushing patterns [21].

Several researchers have attempted to study toothbrushing behaviors by solving the brushing region detection problem, namely, to infer the dental regions being brushed during a brushing session. Some have used motion-sensors that are embedded in the smartwatches worn by the users. In 2016, Huang et al. [22] proposed a cloud-based system for detecting the brushing region using a wrist-watch embedded with accelerometer, gyroscope and magnetometer. They used

The research was funded by National Institute of Dental and Craniofacial Research grant numbers R01DE025244 and 1UG3DE028723. O.H.M.P. was partially funded by NSF grant DMS-2015489. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH or NSF.

M.E. (email: mahmoudessalat@ucla.edu) and G.P. (pottie@ee.ucla.edu) are with the Department of Electrical and Computer Engineering, O.H.M.P (oscar.madrid@stat.ucla.edu) is with the Department of Statistics, and V.S. (vshetty@ucla.edu) is with the School of Dentistry; all at the University of California, Los Angeles (UCLA).

Our provided dataset as well as the code to generate the results are publicly available at the GitHub repository in [1].

a manual toothbrush and modified it by attaching a couple of small magnets to its handle to help with the detection of the rotations of the toothbrush. Their 3-week study involving 12 participants, was conducted in two phases: In the first phase, they trained the participants to perform the structured Bass brushing technique [23] for a specific period of time. In the second phase, they asked the participants to brush using the learned Bass brushing technique. They proposed a machine learning algorithm based on the Naive Bayes classifier. They used both time and frequency domain hand-tailored features from the accelerometer and magnetometer signals to achieve an accuracy of 85.6% in detecting 15 brushing regions. They later showed that by using a two-layer shallow neural network, they could achieve an accuracy of 91.2% [24]. However, their work is not generalizable to real-world brushing scenarios because the structured Bass brushing technique performed in controlled settings does not replicate the free-form and irregular brushing patterns performed in naturalistic settings (e.g. at home). Their algorithm selects only brushing sessions that are performed using the Bass brushing technique; it cannot detect brushing regions for non-conforming brushing sessions. Also, they used the probability matrix of transitions between consecutive brushing regions for each of the 12 participants. This uses each individual's brushing habits with manual toothbrushes and may not be generalizable to a larger population given that they did not report the leave-one-subject-out cross-validation accuracy. Furthermore, since the participants in their study used manual toothbrushes only, it is unclear if their proposed method can be extrapolated to electronic brushes.

Luo et al. [21] used both manual and electronic toothbrushes in a month-long study involving 10 participants wearing wrist-worn accelerometers. They used low-pass filtering to eliminate the differences between the accelerometer signals during manual and electronic toothbrushing. Using an attention-based LSTM network [25], they achieved an accuracy of 97.3% in detecting 15 brushing regions. Here again, the data was collected after teaching the participants to brush according to the structured Bass brushing technique. Also, their results are based on a random train-test split of all the data collected and do not show the generalization power of their model beyond the participants in their study.

Others have tackled the brushing region detection problem by attaching motion-sensors to the brush handle. Lee et al. [26] conducted a study with 15 participants using manual toothbrushes equipped with motion-sensors (including accelerometer, gyroscope and magnetometer) to detect 16 brushing regions. Participants brushed their teeth for three minutes in a prescribed sequence, brushing each dental region for five seconds before transitioning to the next region. They proposed a two-stage algorithm wherein the first stage used the rolling angle of the brush to classify the brush orientation into four categories; up, down, left, and right. In the second stage, they used a K-means classifier based on accelerometer, magnetometer and Euler Angles (EAs) values. The authors reported difficulties in brushing region detection when it came to detecting right versus left brushing regions in their initial studies [26], [27]. However, in a later study [28], they claimed that using the heading angle of the mirror used by the participants helped them to solve the challenge. Despite a reported classification accuracy of 97.1% in detecting 15 brushing regions, the studies [26], [27], [28] had some limitations: first, the dataset of three minutes of brushing per participant is limited; second, they assumed knowledge of the heading angle of the mirror in front of the participants; third, participants were asked to keep their head and body fixed while brushing; and fourth, participants were prescribed the Bass brushing technique and had to follow a predefined, structured brushing sequence. All these limit the applicability of their methods to real-life settings.

Recently, Hussain et al. [29], [30] published a dataset collected from accelerometer and gyroscope sensors attached to the brush handle during toothbrushing. They used 17 participants in their study, each brushing their teeth for one to five brushing sessions for a total of 64 brushing sessions. Three participants used electronic toothbrushes while the rest used manual brushes. They instructed the participants to brush each dental region for about seven seconds according to the Bass brushing technique and tackle the teeth surfaces in a prescribed order. They calculated pitch and roll angles using a complementary filter [31] and used a random forest classifier to detect the brushing regions.

To address the problem of monitoring naturalistic brushing, we:

1) Provide the first publicly available dataset of 9-axis motion-sensor (including accelerometer, gyroscope, and magnetometer) attached to the handle of the toothbrush. This dataset is unique because it includes both manual and electronic toothbrushes and involves free-form brushing sessions. Also, we include two different facing direction during brushing in our dataset; a factor not considered in the previous studies or bypassed by modifying the brush as in [24].
2) We mention the challenges associated with labeling the brushing dataset and propose a relabeling procedure to enhance the quality of dataset annotation.
3) Propose a three-stage framework (i.e., pre-processing, brush transition times estimation, and time-series classification) to tackle the toothbrushing region detection problem. By applying our algorithm, both to our dataset as well as the dataset collected under the structured brushing assumption [29], we show that region detection during free-form brushing is more challenging than detection in controlled settings .
4) Show that a high classification accuracy can be achieved using a random train-test split (i.e. k-fold cross validation), but generalization beyond the participants in the training set (i.e. leave-one-subject-out cross-validation) poses difficulties. Our observation should guide other researchers in reporting the results of their brushing detection studies.
5) Compare different classification methods based on feature engineering and deep learning techniques. We apply Transformer Encoder model to our dataset and show that additional customization can be done for different types of brushes (i.e. manual and electronic) and the attributes of the user (i.e. left-handed and right-handed)

by introducing feature embedding technique used in natural language processing applications (NLP). We show that our proposed classifier can achieve a higher accuracy compared to the method proposed in [30] under all evaluation settings.

The rest of this paper is structured as follows: 1) Section I defines the brushing regions and describes ways to estimate orientation of an object using motion-sensors. We will explain our experimental setup for data collection in Section III. Section II discusses the challenges of our problem and why machine learning is a reasonable approach to tackle the brushing region detection problem. Section V discusses our proposed method. We start with data pre-processing in Subsection V-A, continue with proposing a method for estimating the times that the brush transitions between different regions of the mouth in Subsection V-B, and finally in Subsection V-C, we introduce the machine learning methods that we utilized in our work. We present the results in Section VI, discuss and propose potential directions for future work, in Section VII, and provide conclusions in Section VIII.

I. Toothbrushing Regions

Summarized by Figure 1, we use the standard convention of dividing the dental arches into 16 regions [26].

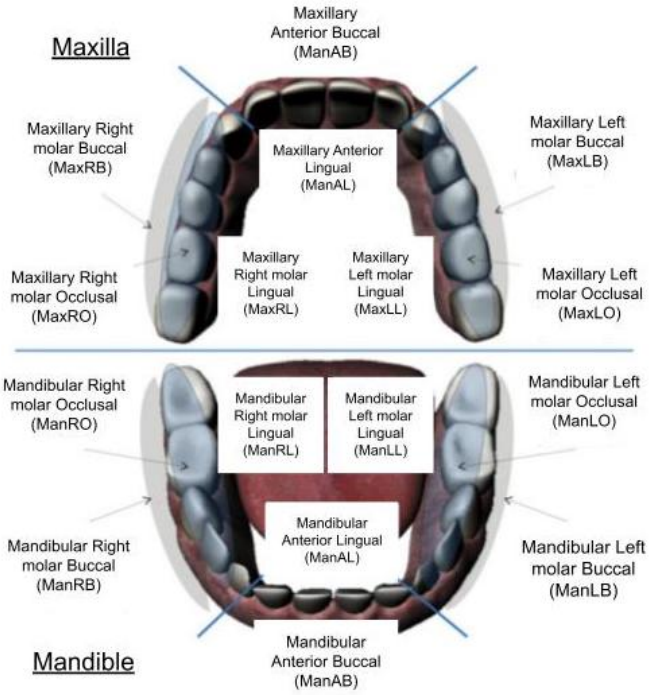

Fig. 1. Defined 16 brushing regions of mouth

II. Challenges of Brushing Region Detection

Detecting the regions that are being brushed would be simple if we could accurately track the position of the toothbrush head in the mouth. However, motion-sensors are not able to do that because of the following limitations:

1) Hardware limitations: Considering the micro-motions and rapid transitions of the toothbrush head during brushing, accurate estimation of toothbrush head position needs very accurate sensors and a sampling rate that greatly exceeds the capabilities of the existing motion-sensors. Moreover, transmitting data –typically via Bluetooth– at a high rate, is not possible due to communications limitations.
2) Size of mouth: For position estimation purposes, the mouth is a small cavity. Further dividing it into 16 brushing regions makes position tracking very difficult. In particular, to estimate the position of the toothbrush in mouth, we should do a double discrete integration of the gravity compensated accelerometer vector defined as follows:

$$A_{GC} = R_B^W \times A - \begin{bmatrix} 0 \\ 0 \\ g \end{bmatrix}. \quad (1)$$

Here, g is the gravitational acceleration which is approximately 9.8 m/s^2, A is the three-dimensional acceleration vector measured by the accelerometer, and R_B^W is the rotation matrix representing the orientation of the brush with respect to the world reference frame. The world reference frame is a coordinate frame with its X-axis in the direction of the north magnetic pole, its Z-axis in the opposite direction of the gravity, and its Y-axis in the direction of the outer product $Z \times X$. The double integration of A_{GC} can amplify any measurement errors in the accelerometer values. Also, the errors that exist in estimating the rotation matrix R_B^W can introduce too much error in location estimation. This makes location estimation using motion-sensors not a suitable approach in our application.

Therefore, most studies just focus on detecting the regions that are being brushed using the orientation of the brush head.

There are numerous methods for orientation estimation, from Kalman based methods [32], [33], [34], to other complementary filters [35], [36]. Kalman based methods are complex, require a lot of parameter hand-tuning, and they usually need a very high sampling rate on the order of kHz [37]. In 2010, Madgwick [37] proposed a method for orientation estimation that could be used for both IMU (i.e. accelerometer plus gyroscope) and MARG (IMU plus magnetometer) sensors. Their method is computationally light and is suitable for low sampling rate scenarios (on the order of a few dozen hertz). We have used this filter in our study.

However, there are some challenges in using orientation of the toothbrush head to detect the brushing regions:

1) Similar brush-head orientation during brushing different regions: Several brushing regions manifest as similar brush orientations. An example can be seen in Figure 2 between ManLO and ManRO regions. In Figure 3, we show EAs of 16 brushing regions that were brushed with manual toothbrush during a brushing session. Many regions show overlapping EAs and convey similar orientations.
2) Intra-subject variation of brushing styles: Even within a particular subject, there are considerable variations in the

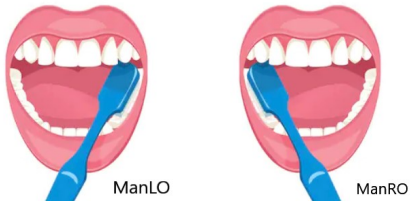

Fig. 2. Two brushing regions which manifest similar brush orientations

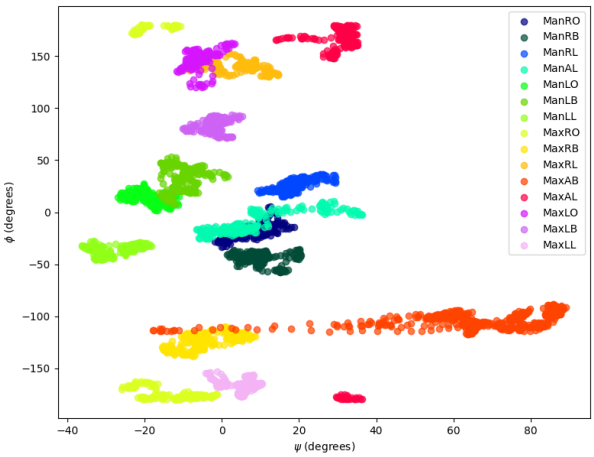

Fig. 3. Euler Angles of the 16 brushing regions of participant #1 during brushing session #1

orientation of the brush-head between multiple brushing sessions. Figure 4 shows EAs of the brush obtained from the same subject in Figure 3 during another brushing session with a manual toothbrush. Another factor that can make this problem even more challenging is the variation in brushing styles over time.

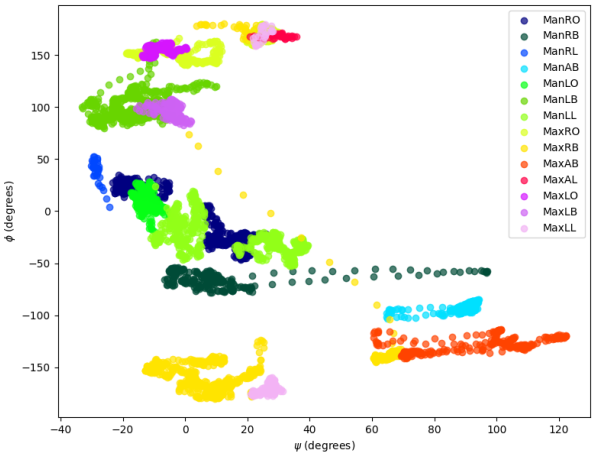

Fig. 4. Euler Angles of the 16 brushing regions of participant #1 during brushing session #27

3) Inter-subject variation in brushing styles: Individuals differ in their brushing styles. In Figure 5, we show the EAs of the brush orientation during a brushing session of a different individual than Figures 4 and 3. Clearly, the brush orientations are different among different individuals brushing the same regions.

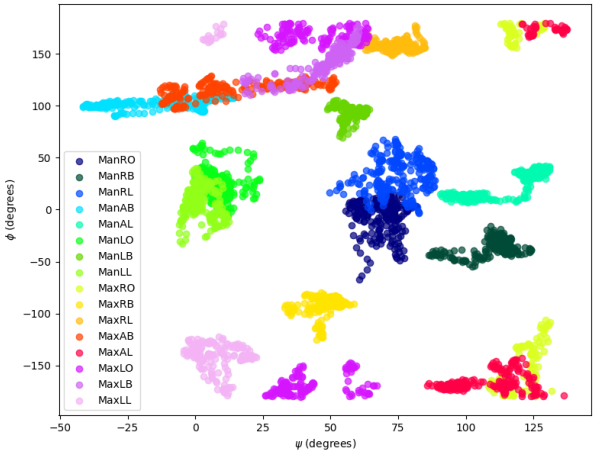

Fig. 5. Euler Angles of the 16 brushing regions of participant #2 during brushing session #2

4) Head movements: To accurately detect the region being brushed, the toothbrush-head orientation with respect to the user's head orientation is needed. Hence, a user's head movements/rotations, typical for a free-form brushing session, can generate artifacts in detecting the brushing regions.

Because of these challenges, determining the boundaries of EAs associated with different brushing regions is not straight forward. Thus, machine learning techniques can be a suitable approach for this application. Also, our premise is that there should exist some time or frequency domain features that can be extracted from the brushing data on the population level as well as the personal level to classify the brushing regions. In fact, some studies have found promising results. For instance, by collecting motion-sensor data from during brushing, the authors in [38] found that the frequency of brushing is higher on the left side of the mouth compared to the right side among right-handed individuals. Also, [39] showed that some statistical features are different in brushing ManLL/MaxLL regions compared to ManAB/MaxAB regions.

III. Data Collection and Experimental Setup

Despite numerous studies on monitoring toothbrushing behavior using motion-sensor data, there is no publicly available dataset that includes free-form brushing. Hence, we designed a study that collected brushing data of 12 participants (2 left-handed and 10 right-handed) over 187 brushing sessions. All participants provided written informed consent and the study protocol was reviewed and approved by the Institutional Review Board of the University of California, Los Angeles (IRB#18-000874).

Our experimental setup is shown in Figure 6. Each brushing session included a random choice of the following factors:

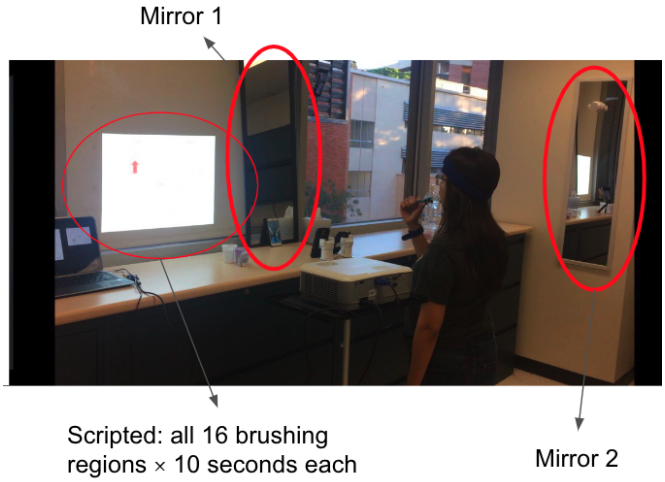

Fig. 6. Our experimental setup

1) Brush type: Because of the differences associated with brushing styles, we included both manual and electric toothbrushes in our study. Brushing with an electric toothbrush entails slower and smoother brushing strokes whereas brushing with a manual toothbrush entails faster and agitated brushing strokes. Also, transitions of the electric brush head from one region to another involves more gliding motions across the teeth surfaces, compared to erratic shuffles of a manual brush head. Previous studies mainly focused on manual toothbrushes. [22], [24], [28], [26]
2) Direction: Our study participants were brushing before two mirrors positioned at 90 degrees angle (heading angle) to each other. Considering the effect of direction on magnetometer [22], our dataset examines robustness to face direction during brushing. A factor that happens naturally because of the users having different face directions in different locations during brushing. Lee et al. [26] did not experiment with different face directions during brushing. Huang et al. [22] proposed a method that was robust to different face directions at the expense of modifying the toothbrush by attaching a couple of strong magnets to the brush handle.
3) Brushing Method: During free-form toothbrushing, individuals often switch frequently from one region to another. These frequent transitions of the brush-head between different regions makes the process of labeling the regions brushed during a session, challenging and time-consuming. Moreover, some brushing sessions can be too short for machine learning purposes which favor having a large amount of labeled data. Hence, in our data collection process, we provided two toothbrushing methods:
    a) Scripted: in which the participants followed a random sequence of brushing regions (without repetition). We created a slide deck in which each slide contained a figure of the mouth with an arrow indicating the region to be brushed at a given time. This sequence was projected on a screen adjacent to the mirror that participants used for brushing. Each slide lasted for 10 seconds. The last 3 seconds of each slide displayed a small pictogram of the next brushing region in order to prepare the participants for the switch to the next brushing region.
    b) Free-form: in which the participants were allowed to brush freely as they do normally. The datasets from previous studies were collected under several constraints, including restrained head and body movements or scripted brushing sequences [21], [22], [26], [29]. Including freestyle brushing in our dataset was essential for developing a brushing region detection algorithm appropriate for real-world brushing scenarios.

We used two 9-axis MARG sensors to collect motion-sense data in each brushing session. One sensor was attached to each participant's toothbrush handle the other one was embedded in a wristband that the participants wore during toothbrushing. All motion-sensors (accelerometer, gyroscope, and magnetometer) collected data at a 25Hz sampling rate. The data was transmitted to a nearby phone via Bluetooth and then stored. We attached the sensors to the same location on all the brushes. The coordinates of the sensors on the brushes as well as the wristband are shown in Figure 7. The mounting orientation of the sensors on the brush and wrist served to avoid the gimbal lock problem [40].

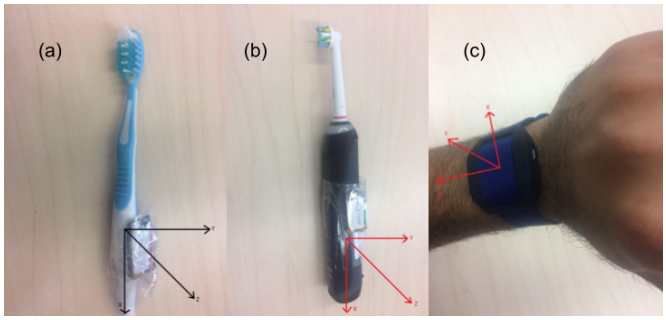

Fig. 7. Sensors used in our study. (a) manual toothbrush, (b) electronic toothbrush, and (c) wristband. The sensors' coordinates are shown in red.

In order to obtain the ground truth labels of the dental regions being brushed during a session, we used two cameras to record the participants' faces from the bottom and top angle.

In order to sync the motion-sensor signals with the videos, we did a pre-processing method described in Subsection V-A1. This paper focuses on brushing region detection using the motion-sensor attached to the toothbrush; the wristband sensor data will be studied in our future work.

IV. Labeling

Labeling toothbrushing regions can be a challenging task due to the fast and frequent transitions of the toothbrush, obscured brushing regions in the video, and brushing of

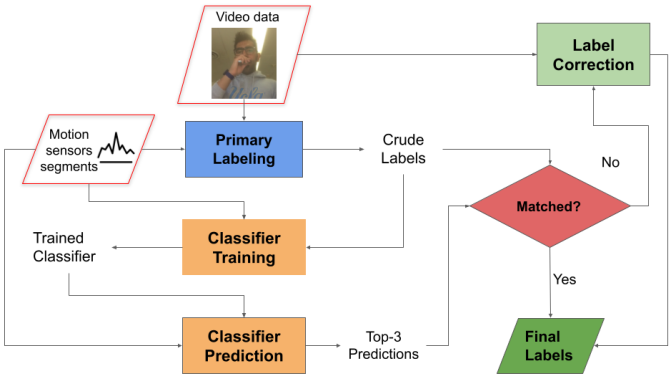

Fig. 8. Relabeling Procedure. Flowchart of our proposed method to enhance the annotation accuracy.

neighboring dental regions together. Due to these challenge, the resulting labels may contain noise and cannot be entirely relied upon. Therefore, we implemented a two-step procedure to enhance the labeling as illustrated in Figure 8, which involved:

1) In the first step, we diligently endeavored to label our dataset using the video data and subsequently applied our proposed algorithm, as described in Section V, to train a classifier based on these initial "crude labels". To accomplish the primary labeling step, after synchronizing the videos with the motion-sensor signals, we labeled the starting and ending samples for each brushing region within the brushing session, excluding the periods when the brush was transitioning between brushing regions or when the user was not actively brushing any region.
2) In the second step, we employed our trained classifier to predict the brushing regions and identified regions that did not rank among the top three predictions made by the classifier. Subsequently, we rectified these regions using the brushing video.

Through this method, we were able to correct roughly 10% of the "crude labels" produced in the first labeling step.

V. Proposed Method

In this section, we propose a three-stage method for the brushing region detection problem which consists of: 1) pre-processing; 2) brush transition time estimation: to find the times that the toothbrush transitions between the brushing regions; and 3) time-series classification: to classify the samples in between consecutive brush transition times into brushing regions.

A. Pre-processing

We performed three pre-processing steps:

1) Interpolation: Bluetooth packets have non-uniform delays during transmission. To deal with this issue, we fit a regression line to the timestamps of the received data packets on the phone. This will adjust for any delays during transmission and results in timestamps with uniform intervals.

Also, the motion-sense signals have missing values due to Bluetooth transmission having missing packets. To overcome this issue, we interpolated the motion-sense signals using spline interpolation. This will yield signals with a true 25Hz sampling rate. To do this interpolation, we dedicated a counter as part of the Bluetooth packets to keep track of the packet number.

2) Low-Pass filtering: We low-pass filtered accelerometer and magentometer signals before feeding them to the orientation filter to estimate the EAs. The reason for doing so is that the orientation filters use accelerometer and magnetometer signals as a feedback signal with low dynamic response to compensate for the erroneous high frequency components in the estimated orientation generated by integrating the gyroscope values. Specifically, the accelerometer and magnetometer measurements are affected by the brushing strokes which appear as high frequency components of their spectrum. Therefore, low-pass filtering only keeps the low-pass components of accelerometer and magnetometer signals that relate to the orientation of the brush. We used a Butterworth low-pass filter of order 5 with a cutoff frequency of 2Hz.

3) Magnetometer Calibration: Magnetometer measurements are affected by the geomagnetic field as well as the ferromagnetic sources that are present in the vicinity of the magnetometer. These ferromagnetic sources, which cause distortion in the magnetometer measurements, can be divided into two categories:

1) Hard-iron distortions: hard-iron sources are the ferromagnetic components that are present on the printed circuit board (PCB) of the sensor [39]. These components are fixed and will reside with the magnetometer after it is manufactured. Hard-Iron sources create a constant bias in the magnetometer measurements and can be modeled by a constant vector added to the magnetometer values.
2) Soft-iron distortions: soft-iron sources are the magnetic components that are present in the environment that the magnetometer is located. Soft-iron sources, include electrical appliances, metal furniture and metal structures within a building's construction [37] which can affect the magnetometer measurements not only directly but also by inducing a temporary magnetic field into normally unmagnetized ferromagnetic components, such as steel shields and batteries that exist on the PCB [39]. The effect of the soft-iron sources can be modeled by a 3*3 matrix multiplied by the magnetometer values.

Hence, magnetometer measurements can be modeled as follows:

$$B_{measured} = W_{soft-iron} \times B_{sensor} + B_{hard-iron}. \quad (2)$$

Where $B_{measured}$ is the distorted magnetometer measurement and B_{sensor} is the ideal magnetometer measurement in the absence of distortions. Of note, any distortion factor that can be modeled as matrix multiplication, such as accelerometer and magnetometer axes misalignment or non-orthogonality of magnetometer axes, can be included in $W_{soft-iron}$. In a similar way, $B_{hard-iron}$ includes any non-zero field offset added during factory calibration of the magnetometer .

During magnetometer calibration, the purpose is to find $B_{hard-iron}$ and $W_{soft-iron}$ so that the magnetic distortions can be compensated in order for the B_{sensor} to be estimated.

It can be easily shown as in [39] that the distorted magnetometer measurements ($B_{measured}$) form the locus of an ellipsoid when rotated enough under arbitrary orientations. To fit an ellipsoid to the magnetometer measurements, we used the method proposed in [41] based on least squares fitting.

By fitting an ellipsoid to the distorted magnetometer measurements ($B_{measured}$), we can find $B_{hard-iron}$ and $W_{soft-iron}$. To find B_{sensor}, $W^{-1}_{soft-iron}$ is needed which can be found uniquely under the constraint of symmetric W matrix [39]. By calibration, B_{sensor} is determined which forms the locus of a zero-centered sphere.

B. Brush Transition Times Estimation

In the second stage of our proposed method, the transition times corresponding to the transitions of the brush between different brushing regions are estimated. As it can be seen in Figure 9, at the brush transition times, often a change in the values of one/multiple channels of accelerometer and magnetometer signals can be observed, which is because of the change in the brush orientation. Generally, detecting a change in signal is referred to as change point detection (CPD) in the literature [42]. Also, we call the signal samples that lie in between two consecutive change points, a segment.

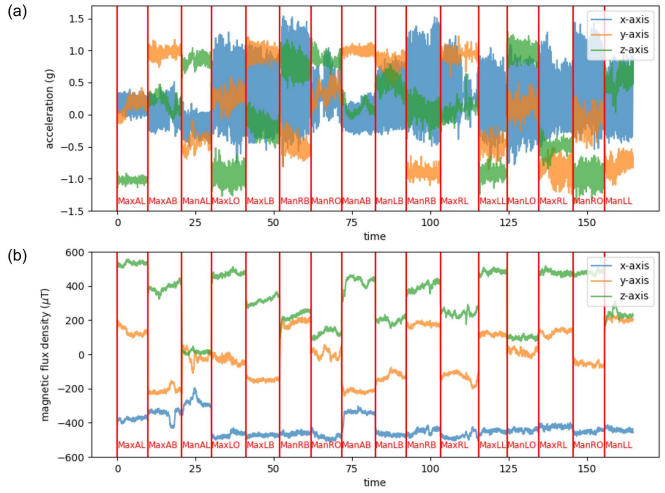

Fig. 9. Motion-sense signals during a brushing session. (a) accelerometer signals and (b) magnetometer signals. Times corresponding to the transition of the brush between consecutive brushing regions are shown with red vertical lines.

While there are many change point detection methods (see the review paper [43]), most of these methods - to the extent that we examined - either have high computational costs or are designed for specific applications. This limits their usability in our application. As an alternate, we propose a simple and computationally light window-based CPD method, which we call window-CPD.

In window-CPD, a sliding window of length L with a stride of P percent of the window length, runs over the motion-sense signals and detects a change in the signals' means between consecutive windows. Particularly, to detect a change in the signals, we use the following criterion:

$$\left\|\bar{S}_i - \bar{S}_{i-1}\right\|_2 > \|\sigma_{i-1}\|_2 \times \alpha + \beta. \tag{3}$$

This compares the Euclidean norm of the change in $\bar{S}$ (signals' temporal mean in a window) between the current and the previous window, to an affine function of σ_{i-1} (the Euclidean norm of the signals' temporal standard deviation in the previous window). Here, signal S represents the 6-channel signal resulted from the concatenation of accelerometer and magnetometer signals.

Including the standard deviation in the threshold causes the threshold to be adapted to the recent range of variations in the signals. From our observations, this adaptive threshold results in better change point estimations compared to having a constant value as the threshold.

We used $\alpha = 0.05$, $\beta = 0.2$, $L = 20$, and $P = 50$ heuristically in our study.

C. Time-series Classification

There are many machine learning algorithms to classify time-series data, ranging from classical methods such as feature engineering, in which features are hand-tailored, to more recent methods which use deep learning to automatically extract features. In this section, we explain the methods that we used to classify motion-sensor data and compare their performances in Section VI.

1) Feature Engineering Method: In this method, we extracted time and frequency domain features from the segments in our training set. Features extracted in our work were the features that were frequently used in the area of activity recognition using motion-sensors.

The features we extracted from a segment of motion-sense signals included mean and variance of the signals, coefficients of Auto-Regressive (AR) model of the signals, mean and variance of EAs, mean-frequencies of the motion-sense signals' spectra extracted using the Yule-Walker [44] spectrum estimation method (after filtering the signals with a band-pass Butterworth filter of order 5 and cutoff frequencies of 0.1Hz and 2Hz), cross correlation of the motion-sense signals, and mean and variance of the gravity compensated linear acceleration of the brush (A_{GC} in Equation 1).

Out of 120 features extracted in this stage, we selected the best 20 features using the Fisher score [45] based on finding the optimum projection matrix Φ to maximize the Rayleigh quotient defined as follows:

$$\mathcal{J}(\Phi) = \frac{\left|\Phi^T \hat{\Sigma}_b \Phi\right|}{\left|\Phi^T \hat{\Sigma}_w \Phi\right|}. \tag{4}$$

This is the ratio of between-class variance to within-class variance and hence $\hat{\Sigma}_b$ and $\hat{\Sigma}_w$ are called between-class and within-class scattering matrix, respectively. It can be shown that the solution to the above optimization problem can be obtained by solving the following generalized eigenvalue problem:

$$\hat{\Sigma}_b \Phi = \lambda \hat{\Sigma}_w \Phi. \tag{5}$$

After finding Φ and selecting the best features for classification, a classifier can be used to detect the region that is being brushed during the segment. We tried different classifiers such as Naive Bayes, Random Forest, Support Vector Machine (SVM). These classifiers performed similarly and hence we just showcase the result of the Naive Bayes classifier in Section VI.

2) Long Short-Term Memory (LSTM) Networks: LSTM networks [46] are specific types of Recurrent Neural Networks (RNNs) that can overcome the exploding and vanishing gradient problems happening during training of the Neural Networks. They are being used for classifying time-series data and are powerful at remembering the long term dependencies in time-series data. By using the LSTM networks, we intended to capture not only the features related to the specific orientations of the brush but also any motion patterns that exist during brushing a particular region.

We concatenated the accelerometer, magnetometer, and EAs to obtain 9-dimensional feature vectors and then fed them to our LSTM model. We used two layers of bidirectional LSTM networks, with 64 hidden units and a drop-out layer with a drop-out probability of 0.1 after each layer. At the end, we connected the LSTM output to a fully-connected layer and then calculated the Softmax probabilities of the brushing regions. Our model hyper-parameters were selected heuristically by a coarse-to-fine approach.

For training, we used multi-class cross entropy loss and Adam [47] optimizer with a learning rate of 1e-3. Our network was trained for 10 epochs with a batch-size of 1024. The number of epochs was chosen based on the training and evaluation accuracy to avoid overfitting.

We further describe our training process in Subsection V-C4.

3) Transformer Encoder: Transformers made a revolution in natural language processing (NLP). These models that were initially used in language translation tasks, consist of two parts: an encoder and a decoder. These models use the attention mechanism [48] to learn the long-term dependencies in a corpus and are shown to outperform the RNN models in almost all NLP tasks.

However both the encoder and decoder parts of the Transformers are needed for language translation tasks, the Transformer Encoder can be used by itself to learn a vector representation of the time-series data for classification tasks. This is accomplished by adding a classification token to the beginning of the time-series segment.

In this paper, we used the Transformer Encoder model to classify the segments of motion-sense signals during a brushing session. The architecture of our model is shown in Figure 10. Similar to the Subsection V-C2, we concatenated the accelerometer, magnetometer, and EAs as 9-dimensional feature vectors and then fed them into our Transformer Encoder model. We first used a fully-connected layer of size 32 to project our feature vectors into a feature space with higher dimensions. Then we added a positional embedding vector as well as brush-type and left-handedness embedding vectors of size 32 to the feature vectors. The addition of the latter embedding vectors will allow further personalization to be possible via Transformer Encoder models.

Then we used a stack of four attention layers each containing a feed-forward (fully-connected) layer with 256 hidden units. Each attention layer has two attention heads with a size of 128 for key, query and value vectors. We chose a drop-out probability of 0.1 for both the attention scores as well as the feed-forward layer inside each attention layer. We used Gelu activation function for the feed forward layers. Here again, our model hyper-parameters were selected heuristically by a coarse-to-fine approach.

Our training configuration was similar to the LSTM network discussed in Subsection V-C2, with the exception of using the BERT-Adam optimizer instead of the Adam optimizer.

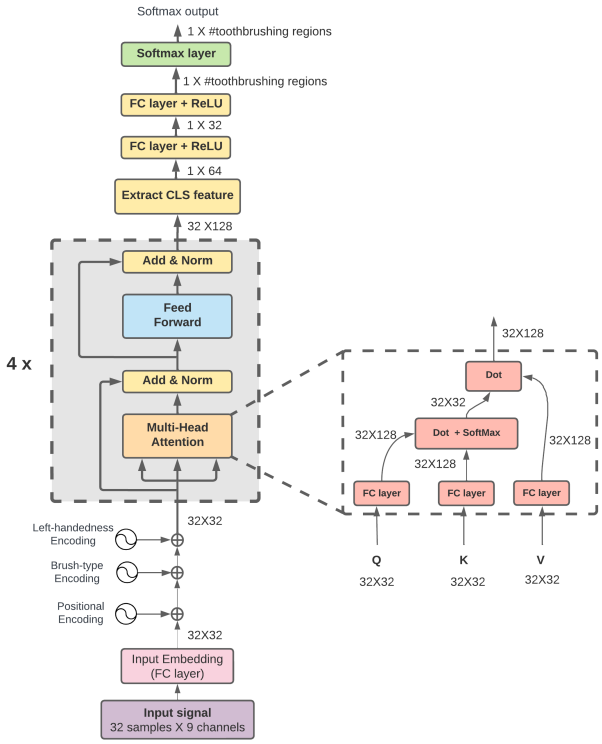

Fig. 10. Our Transformer Encoder model configuration

4) Majority voting technique: Since we know that each segment of the motion-sense data which lies in between two consecutive brush transition times corresponds to only one brushing region, any subsegment of each segment should also corresponds to the same brushing region. Therefore, we can split each segment into multiple subsegments using a sliding window of size l with a stride of p percent of the window length. At the training time, splitting our segments into multiple subsegments is useful due to the increase in the training set size which helps with reducing the generalization error of our classifiers. At the test time, to predict the class that each segment belongs to, we can split the segment into multiple subsegments and perform majority voting over the class predictions of the subsegments. For majority voting, we treated the subsegments as independent measurements. Therefore, we summed the log probabilities of the subsegments' distributions over the classes (i.e. brushing regions) to find the class with the highest probability as the final prediction.

We applied this technique to predict segments using both the LSTM and the Transformer Encoder models and observed a performance improvement of approximately 10% in accuracy compared to not using the technique. We set the values of $l = 32$ and $p = 0.25$ heuristically.

VI. Results

In this section, we present the performance of our proposed algorithm using various segment classification algorithms described in Subsection V-C. First, we describe the evaluation metrics employed in our study, followed by a comparison of the performance of different classifiers based on these metrics.

A. Evaluation metrics

We report three cross-validation accuracy metrics defined as follows:

- One-subject-out accuracy: To calculate this metric, we train our model on brushing data from all participants except for one and evaluate its performance on the left-out participant. We repeat this process for each participant and then average the results. This metric evaluates the model's generalization power to new participants outside the training set and measures how well it can perform at the population level.
- One-session-out accuracy: For this cross-validation metric, we train our model on all brushing sessions of a participant except for one, which we use as the validation set. We repeat this process for all sessions of the participant and all participants, and then average the results. This method evaluates the personalization power of the models for each individual since the training set includes data from one participant.
- K-fold accuracy: For k-fold (we set k=5), cross-validation [49] all brushing sessions of all participants are used. We used this metric since it is conventionally used to evaluate the performance of the machine learning models.

B. Classification results

To evaluate the performance of our proposed algorithm, we report micro-F1 score [50] for sample-based and segment-based classification methods in Table 1. In the sample-based classification method, we classify each sample of motion-sense signals based on the method proposed in [30]. The Random Forest classifier is applied to the feature vectors generated by concatenation of accelerometer, magnetometer, and roll and pitch angles for each time-sample. The roll and pitch angles are estimated using the proposed complimentary filter in [30]. In contrast to classifying feature vectors of each time-sample in sample-based methods, in the segment-based classification methods, we classify the segments that lie in between the consecutive brush transition times. In segment-based methods, we ultimately assign the predicted class of each segment to all samples in the corresponding segment. Hence, the reported accuracy of segment-based methods is calculated based on the correct predictions of the samples. This makes the comparison of the performance of the segment-based and the sample-based methods valid.

For classification, since some regions are brushed together or the accurate location of the brush is not clear in the videos, we merged some regions which were often ambiguous during labeling following the convention proposed in [51]. Hence, we ended up with 9 brushing regions as follows:

$$R = \{ManRO/ManRL, ManLO/ManLL, MaxLO/MaxLL, MaxRO/MaxRL, MaxRB/ManRB, MaxLB/ManLB, MaxAB/ManAB, ManAL, MaxAL\}.$$

Table 1. F1 score of our proposed classification methods

			Our Dataset	Dataset in [29], [30]
Sample-Based (method in [29], [30])		K-fold	81.5	91.2
		Subject-out	52.4	61.5
		Session-out	60.7	67.3
Segment-Based	Feature Engineering	K-fold	80.3	89.5
		Subject-out	50.7	58.1
		Session-out	60.4	66.5
	LSTM	K-fold	89.5	93.3
		Subject-out	53.4	63.7
		Session-out	65.3	70.5
	Transformer Encoder	K-fold	95.3	97.2
		Subject-out	58.6	67.9
		Session-out	67.3	74.7

As it can be seen in Table 1, our proposed segment-based methods outperform the sample-based method proposed in [30] when applied on both their published dataset as well as our provided dataset. The difference in the performances of our model when applied on our dataset versus the dataset in [29], shows the challenging nature of brushing region detection when performed free-form versus under constraints (such as prescribed sequence of brushing, structured Bass brushing technique, etc.). While our models can achieve high k-fold cross-validation accuracy, one-subject-out classification accuracy is much more challenging. This large gap in the models' performances, manifest the vast variation in brushing styles and the other challenges discussed in Section II. Hence, it is important to assess models' performances based on one-subject-out evaluation metric. This should guide other researchers in the field when reporting their models' performance metrics.

Furthermore, it can be observed that one-session-out accuracy is partially higher than the one-subject-out accuracy. This may be because of the fact that the individuals' brushing habits might differ from each other and hence a personalized model would perform better than a general model for the whole population.

Overall, the Transformer Encoder model performed slightly better than the other segment-based models. The one-subject-out and one-session-out accuracy metrics for the Transformer Encoder model are shown in Figures 12 and 13, respectively. In the one-subject-out histogram, each bar represents the accuracy of the model when the data from that participant constituted the validation set. In the one-session-out histogram, each bar represents the accuracy result of the corresponding participant averaged over cases where each session of that

participant constituted the validation set and the rest of that participant's sessions constituted the training set. Figure 11 displays the confusion matrix of the Transformer Encoder model applied to our provided dataset, with the one-subject-out accuracy calculated for the case where participant #1 is left out. Our model performed well in distinguishing regions with different brush-head orientations (e.g., ManRO and MaxRO). However, it had some difficulties discerning regions with similar brush orientations, such as ManRO and ManLO regions.

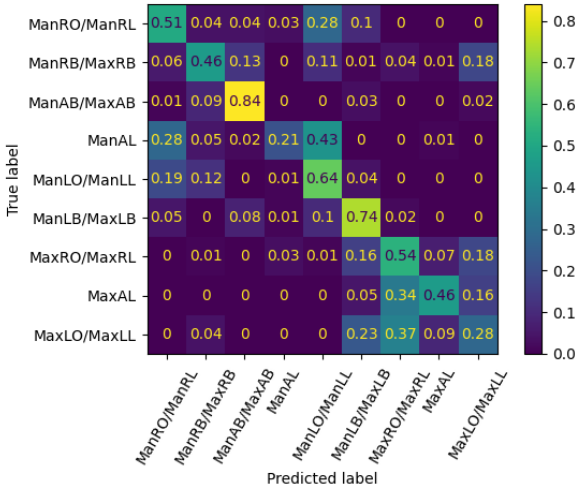

Fig. 11. Confusion matrix of the Transformer Encoder model applied to our provided dataset, with the one-subject-out accuracy calculated for the case where participant #1 is left out.

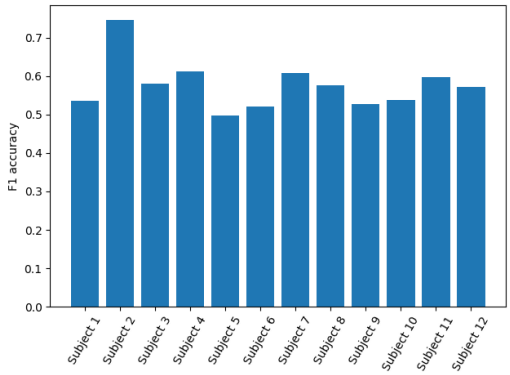

Fig. 12. F1 score of one-subject-out evaluation metric using the Transformer Encoder model

VII. Discussion and Future Work

A. Hyper-parameter Tuning

The outcomes, as presented in Section VI, were achieved via a hyperparameter tuning process. To optimize the hyperparameters of our classifiers, we employed a coarse-to-fine approach consisting of two steps:

1) Initially, a coarse grid of values was selected to evaluate the overall performance of the classifiers over a wide range of hyperparameter values.

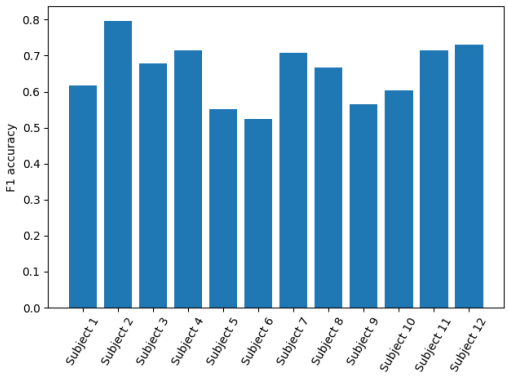

Fig. 13. F1 score of one-session-out evaluation metric using the Transformer Encoder model

2) Subsequently, in the second step, a finer grid of values was explored around the range of values that yielded the best results on the desired metric, particularly one-subject-out cross-validation accuracy.

In addition, in the case of neural network classifiers, we did not allow the time-consuming training process to run for multiple epochs. Instead, we terminated the training process after the first few epochs. This approach is in line with the concept of the critical learning period in deep learning [52], where the initial phase of training neural networks is shown to be critical in achieving the final performance of the network.

In general, we observed that for neural networks, regardless of the type of the three cross-validation metrics mentioned in Section VI, increasing model complexity enables attaining high training accuracy. However, this usually results in a decline in validation accuracy, commonly known as overfitting in machine learning. Below are the ranges of hyperparameters that we found to be effective for the classifiers we employed. Hyperparameter values falling outside these ranges may cause underfitting or overfitting for the classifiers we utilized:

- Transformer Encoder:
  - Hidden size for the key, query, and value: [128, 256]
  - number of stacked-attention layers: [3, 5]
  - number of hidden units in the attention layer: [128, 512]
  - number of attention heads: [2, 4]
  - dropout probabilities: [0.1, 0.4]
- LSTM:
  - hidden size: [64, 128]
  - dropout probability: [0.1, 0.4]
- Random-Forrest (for the method proposed in [30]):
  - number of trees: [70, 120]
  - trees' depths: [12, 18]

Regarding the optimization process, we discovered the subsequent range of values that prevent underfitting or overfitting:

- learning rate: [5e-4, 5e-3]
- number of epochs: [5, 10] for one-subject- and one-session- out cross-validation, and [20, 30] for k-fold cross-validation
- batch size: [256, 2048]

For the majority-voting technique:

- segment length: [0.4, 1.5] $\times$ sampling rate
- stride length: [0.1, 0.4] $\times$ sampling rate

Nevertheless, although we attempted to find the optimal values within the aforementioned ranges, additional experiments within these ranges may yield further optimization.

B. Labeling

Regarding the labeling process, we implemented a two-step algorithm that involves relabeling some of the inaccurate labels using the trained classifiers' predictions. However, the literature provides various methods to handle partially accurate labels, commonly referred to as noisy labels [53]. Some methods tackle this problem by modifying the loss function, such as Bootstrapping, Forward/Backward Loss Correction, and Mixup, while others rely on relabeling the data based on the trained classifiers' predictions. However, in our study, we proposed a relabeling approach that uses the classifiers' predictions in conjunction with the video recordings of the brushing sessions as the ground truth. Our approach is particularly suitable for our work due to the relatively small size of our dataset and the lack of brushing datasets in the field. Moreover, for regions that can be brushed together (e.g., ManAB, MaxAB), we could label them as ManAB/MaxAB, which could pave the way for multi-label classification [54].

C. Classification Accuracy

Our results signifies that generalization to specific manifolds of data (e.g. one-subject-out cross-validation), is a more challenging problem in cases such as toothbrushing region prediction which involves significant between- and with- subject variability. Considering the complexity and the variations included in our problem, we think that a bigger dataset (our speculation is at least 20 times bigger) is needed for the models to be able to generalize to the subjects outside our datasets' participants. In small data scenarios such as our case, we also hypothesize that other avenues of machine learning which are suitable for out-of-distribution genralization, such as transfer learning [55], can possibly be effective. This may need training models on other datasets that involve IMU sensors for detection of other activities other than toothbrushing.

VIII. Conclusion

In this paper, we provided a dataset for detecting the dental regions that are being brushed during a brushing session. Our dataset consists of toothbrushing motion-sense signals collected from accelerometer, gyroscope, and magnetometer sensors attached to the brush. We collected 187 brushing sessions from 12 participants. Unlike previous studies with constrained brushing conditions (e.g. structured Bass brushing technique, prescribed brushing transition sequence, fixed head/body position), our data collection includes free-form brushing that is representative of naturalistic settings. We discussed the challenges of labeling the brushing data and proposed a two-stage relabeling algorithm aimed at enhancing the accuracy of data annotation.

We discussed the difficulty of accurate position tracking of the toothbrush during brushing. We also proposed a three-stage algorithm (i.e. pre-processing, brush transition times detection, and time-series classification) to predict nine brushing regions. Our algorithm finds the times when the brush transitions between dental regions and classifies the segments that lie between consecutive brush transition times. We observed that free-style brushing is more challenging compared to the dataset in [29], which is collected under structured brushing assumptions. We also showed that classifiers can achieve high classification accuracy using a random train-test split (i.e. k-fold cross-validation); however, generalization to new toothbrush users (i.e. leave-one-subject-out cross-validation) poses significant challenges due to variations in brushing styles and other factors such as head movement during brushing. We believe that expanding the brushing dataset to include more participants and increasing the number of sessions for each participant can be effective in generating better models.

Acknowledgment

The authors wish to acknowledge the material support provided Oral-B/Procter & Gamble. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or Procter & Gamble. We also want to thank the students Isabel Roig Penso, Sumukh Kak, E Jun Chang, Anika Desar, Nicole Lin, Maia Le, and Alok Elashoff who assisted with the data collection/labeling.

References

[1] M. Essalat, "Toothbrushing-region-detection," 4 2023. [Online]. Available: https://github.com/ROBAS-UCLA/Toothbrushing-region-detection

[2] V. Shetty, D. Morrison, T. Belin, T. Hnat, and S. Kumar, "A scalable system for passively monitoring oral health behaviors using electronic toothbrushes in the home setting: Development and feasibility study," *JMIR mHealth and uHealth*, vol. 8, no. 6, p. e17347, 2020.

[3] A. M. Mark, "Options for dealing with tooth decay," *The Journal of the American Dental Association*, vol. 149, no. 10, pp. 927–928, 2018.

[4] P. D. Marsh, "Microbiology of dental plaque biofilms and their role in oral health and caries," *Dental Clinics*, vol. 54, no. 3, pp. 441–454, 2010.

[5] W. V. Giannobile, T. Beikler, J. S. Kinney, C. A. Ramseier, T. Morelli, and D. T. Wong, "Saliva as a diagnostic tool for periodontal disease: current state and future directions," *Periodontology 2000*, vol. 50, p. 52, 2009.

[6] T. J. O'Leary, "Oral hygiene agents and procedures," *Journal of Periodontology*, vol. 41, no. 11, pp. 625–629, 1970.

[7] E. B. Hancock, "Question set supports that gingivitis," *Ann Periodontol*, vol. 1, pp. 223–249, 1996.

[8] M. Essalat, D. Morrison, S. Kak, E. J. Chang, I. R. Penso, R. J. Kulchar, O. H. M. Padilla, and V. Shetty, "A naturalistic study of brushing patterns using powered toothbrushes," *Plos one*, vol. 17, no. 5, p. e0263638, 2022.

[9] S. Michie, L. Yardley, R. West, K. Patrick, and F. Greaves, "Developing and evaluating digital interventions to promote behavior change in health and health care: recommendations resulting from an international workshop," *Journal of medical Internet research*, vol. 19, no. 6, p. e232, 2017.

[10] ——, "Developing and evaluating digital interventions to promote behavior change in health and health care: recommendations resulting from an international workshop," *Journal of medical Internet research*, vol. 19, no. 6, p. e232, 2017.

[11] W. T. Riley, K. J. Serrano, W. Nilsen, and A. A. Atienza, "Mobile and wireless technologies in health behavior and the potential for intensively adaptive interventions," *Current opinion in psychology*, vol. 5, pp. 67–71, 2015.

[12] V. Galluzzi, T. Herman, and P. Polgreen, "Hand hygiene duration and technique recognition using wrist-worn sensors," in *Proceedings of the 14th International Conference on Information Processing in Sensor Networks*, 2015, pp. 106–117.

[13] S. Akther, N. Saleheen, S. A. Samiei, V. Shetty, E. Ertin, and S. Kumar, "moral: An mhealth model for inferring oral hygiene behaviors in-the-wild using wrist-worn inertial sensors," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 3, no. 1, pp. 1–25, 2019.

[14] B. Bruno, F. Mastrogiovanni, and A. Sgorbissa, "A public domain dataset for adl recognition using wrist-placed accelerometers," in *the 23rd IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 2014, pp. 738–743.

[15] B. Bruno, F. Mastrogiovanni, A. Sgorbissa, T. Vernazza, and R. Zaccaria, "Human motion modelling and recognition: A computational approach," in *2012 IEEE International Conference on Automation Science and Engineering (CASE)*. IEEE, 2012, pp. 156–161.

[16] ——, "Analysis of human behavior recognition algorithms based on acceleration data," in *2013 IEEE International Conference on Robotics and Automation*. IEEE, 2013, pp. 1602–1607.

[17] A. Parate, M.-C. Chiu, C. Chadowitz, D. Ganesan, and E. Kalogerakis, "Risq: Recognizing smoking gestures with inertial sensors on a wristband," in *Proceedings of the 12th annual international conference on Mobile systems, applications, and services*, 2014, pp. 149–161.

[18] N. Saleheen, A. A. Ali, S. M. Hossain, H. Sarker, S. Chatterjee, B. Marlin, E. Ertin, M. Al'Absi, and S. Kumar, "puffmarker: a multi-sensor approach for pinpointing the timing of first lapse in smoking cessation," in *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 2015, pp. 999–1010.

[19] A. Parate, M.-C. Chiu, C. Chadowitz, D. Ganesan, and E. Kalogerakis, "Risq: Recognizing smoking gestures with inertial sensors on a wristband," in *Proceedings of the 12th annual international conference on Mobile systems, applications, and services*, 2014, pp. 149–161.

[20] E. Thomaz, I. Essa, and G. D. Abowd, "A practical approach for recognizing eating moments with wrist-mounted inertial sensing," in *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 2015, pp. 1029–1040.

[21] C. Luo, X. Feng, J. Chen, J. Li, W. Xu, W. Li, L. Zhang, Z. Tari, and A. Y. Zomaya, "Brush like a dentist: Accurate monitoring of toothbrushing via wrist-worn gesture sensing," in *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*. IEEE, 2019, pp. 1234–1242.

[22] H. Huang and S. Lin, "Toothbrushing monitoring using wrist watch," in *Proceedings of the 14th ACM Conference on Embedded Network Sensor Systems CD-ROM*, 2016, pp. 202–215.

[23] J. Gibson and A. B. Wade, "Plaque removal by the bass and roll brushing techniques." *Journal of periodontology*, vol. 48, no. 8, pp. 456–459, 1977.

[24] H. Huang and S. Lin, "Toothbrushing recognition using neural networks," in *2017 IEEE/ACM Second International Conference on Internet-of-Things Design and Implementation (IoTDI)*. IEEE, 2017, pp. 309–310.

[25] Y. Wang, M. Huang, X. Zhu, and L. Zhao, "Attention-based lstm for aspect-level sentiment classification," in *Proceedings of the 2016 conference on empirical methods in natural language processing*, 2016, pp. 606–615.

[26] Y.-J. Lee, P.-J. Lee, K.-S. Kim, W. Park, K.-D. Kim, D. Hwang, and J.-W. Lee, "Toothbrushing region detection using three-axis accelerometer and magnetic sensor," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 3, pp. 872–881, 2011.

[27] J.-W. Lee, K.-H. Lee, K.-S. Kim, D.-J. Kim, and K. Kim, "Development of smart toothbrush monitoring system for ubiquitous healthcare," in *2006 International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2006, pp. 6422–6425.

[28] K.-H. Lee, J.-W. Lee, K.-S. Kim, D.-J. Kim, K. Kim, H.-K. Yang, K. Jeong, and B. Lee, "Tooth brushing pattern classification using three-axis accelerometer and magnetic sensor for smart toothbrush," in *2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2007, pp. 4211–4214.

[29] Z. Hussain, D. Waterworth, M. Aldeer, W. E. Zhang, and Q. Z. Sheng, "Toothbrushing data and analysis of its potential use in human activity recognition applications: dataset," in *Proceedings of the Third Workshop on Data: Acquisition To Analysis*, 2020, pp. 31–34.

[30] Z. Hussain, D. Waterworth, M. Aldeer, W. E. Zhang, Q. Z. Sheng, and J. Ortiz, "Do you brush your teeth properly? an off-body sensor-based approach for toothbrushing monitoring," in *2021 IEEE International Conference on Digital Health (ICDH)*. IEEE, 2021, pp. 59–69.

[31] M. Pedley, "Tilt sensing using a three-axis accelerometer," *Freescale semiconductor application note*, vol. 1, pp. 2012–2013, 2013.

[32] E. Foxlin, "Inertial head-tracker sensor fusion by a complementary separate-bias kalman filter," in *Proceedings of the IEEE 1996 Virtual Reality Annual International Symposium*. IEEE, 1996, pp. 185–194.

[33] H. J. Luinge, P. H. Veltink, and C. T. Baten, "Estimation of orientation with gyroscopes and accelerometers," in *Proceedings of the First Joint BMES/EMBS Conference. 1999 IEEE Engineering in Medicine and Biology 21st Annual Conference and the 1999 Annual Fall Meeting of the Biomedical Engineering Society (Cat. N*, vol. 2. IEEE, 1999, pp. 844–vol.

[34] J. L. Marins, X. Yun, E. R. Bachmann, R. B. McGhee, and M. J. Zyda, "An extended kalman filter for quaternion-based orientation estimation using marg sensors," in *Proceedings 2001 IEEE/RSJ International Conference on Intelligent Robots and Systems. Expanding the Societal Role of Robotics in the the Next Millennium (Cat. No. 01CH37180)*, vol. 4. IEEE, 2001, pp. 2003–2011.

[35] M. Euston, P. Coote, R. Mahony, J. Kim, and T. Hamel, "A complementary filter for attitude estimation of a fixed-wing uav," in *2008 IEEE/RSJ international conference on intelligent robots and systems*. IEEE, 2008, pp. 340–345.

[36] R. Mahony, T. Hamel, and J.-M. Pflimlin, "Nonlinear complementary filters on the special orthogonal group," *IEEE Transactions on automatic control*, vol. 53, no. 5, pp. 1203–1218, 2008.

[37] S. Madgwick, "An efficient orientation filter for inertial and inertial/magnetic sensor arrays," *Report x-io and University of Bristol (UK)*, vol. 25, pp. 113–118, 2010.

[38] E. Inada, I. Saitoh, Y. Yu, D. Tomiyama, D. Murakami, Y. Takemoto, K. Morizono, T. Iwasaki, Y. Iwase, and Y. Yamasaki, "Quantitative evaluation of toothbrush and arm-joint motion during tooth brushing," *Clinical oral investigations*, vol. 19, no. 6, pp. 1451–1462, 2015.

[39] T. Ozyagcilar, "Calibrating an ecompass in the presence of hard and soft-iron interference," *Freescale Semiconductor Ltd*, pp. 1–17, 2012.

[40] J. Diebel, "Representing attitude: Euler angles, unit quaternions, and rotation vectors," *Matrix*, vol. 58, no. 15-16, pp. 1–35, 2006.

[41] S. Bektas, "Orthogonal distance from an ellipsoid," *Boletim de Ciências Geodésicas*, vol. 20, no. 4, pp. 970–983, 2014.

[42] R. P. Adams and D. J. MacKay, "Bayesian online changepoint detection," *arXiv preprint arXiv:0710.3742*, 2007.

[43] R. Killick, K. Haynes, I. Eckley, P. Fearnhead, and J. Lee, "Package 'changepoint'," *R package version 0.4.-2011.-http://cran. rproject. org/web/packages/changepoint/index. html*, 2016.

[44] Y. Chan and R. Langford, "Spectral estimation via the high-order yule-walker equations," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 30, no. 5, pp. 689–698, 1982.

[45] Q. Gu, Z. Li, and J. Han, "Generalized fisher score for feature selection," *arXiv preprint arXiv:1202.3725*, 2012.

[46] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[47] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[48] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[49] Y. Bengio and Y. Grandvalet, "No unbiased estimator of the variance of k-fold cross-validation," *Journal of machine learning research*, vol. 5, no. Sep, pp. 1089–1105, 2004.

[50] C. Goutte and E. Gaussier, "A probabilistic interpretation of precision, recall and f-score, with implication for evaluation," in *Advances in Information Retrieval: 27th European Conference on IR Research, ECIR 2005, Santiago de Compostela, Spain, March 21-23, 2005. Proceedings 27*. Springer, 2005, pp. 345–359.

[51] S. Akther, N. Saleheen, M. Saha, V. Shetty, and S. Kumar, "mteeth: Identifying brushing teeth surfaces using wrist-worn inertial sensors," *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, vol. 5, no. 2, pp. 1–25, 2021.

[52] A. Achille, M. Rovere, and S. Soatto, "Critical learning periods in deep neural networks," *arXiv preprint arXiv:1711.08856*, 2017.

[53] H. Song, M. Kim, D. Park, Y. Shin, and J.-G. Lee, "Learning from noisy labels with deep neural networks: A survey," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[54] G. Tsoumakas and I. Katakis, "Multi-label classification: An overview," *International Journal of Data Warehousing and Mining (IJDWM)*, vol. 3, no. 3, pp. 1–13, 2007.

[55] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *Journal of Big data*, vol. 3, no. 1, pp. 1–40, 2016.