mTeeth: Identifying Brushing Teeth Surfaces Using Wrist-Worn Inertial Sensors

SAYMA AKTHER, University of Memphis
NAZIR SALEHEEN, University of Memphis
MITHUN SAHA, University of Memphis
VIVEK SHETTY, University of California, Los Angeles
SANTOSH KUMAR, University of Memphis

Ensuring that all the teeth surfaces are adequately covered during daily brushing can reduce the risk of several oral diseases. In this paper, we propose the *mTeeth* model to detect teeth surfaces being brushed with a manual toothbrush in the natural free-living environment using wrist-worn inertial sensors. To unambiguously label sensor data corresponding to different surfaces and capture all transitions that last only milliseconds, we present a lightweight method to detect the micro-event of *brushing strokes* that cleanly demarcates transitions among brushing surfaces. Using features extracted from brushing strokes, we propose a Bayesian Ensemble method that leverages the natural hierarchy among teeth surfaces and patterns of transition among them. For training and testing, we enrich a publicly-available wrist-worn inertial sensor dataset collected from the natural environment with time-synchronized precise labels of brushing surface timings and moments of transition. We annotate 10,230 instances of brushing on different surfaces from 114 episodes and evaluate the impact of wide between-person and within-person between-episode variability on machine learning model's performance for brushing surface detection.

CCS Concepts: • **Human-centered computing** → *Ubiquitous and mobile computing design and evaluation methods*;

Additional Key Words and Phrases: mHealth, brushing detection, flossing detection, hand-to-mouth gestures

ACM Reference Format:
Sayma Akther, Nazir Saleheen, Mithun Saha, Vivek Shetty, and Santosh Kumar. 2021. mTeeth: Identifying Brushing Teeth Surfaces Using Wrist-Worn Inertial Sensors . *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 2, Article 53 (June 2021), 25 pages. https://doi.org/10.1145/3463494

1 INTRODUCTION

Dental disease (caries and gum disease) is very prevalent globally, affecting 53 million people in USA alone. A primary reason for continued prevalence of dental diseases despite regular brushing is that people may not be brushing each tooth surface adequately, missing some surfaces completely, while spending disproportionate time on other surfaces. When saliva combines with particles from food and drinks we consume, a colorless, sticky biofilm containing bacteria known as *dental plaque* forms on our teeth. Unmindful or poor brushing habits allow plaque to accumulate over time, leading to gum disease, tooth decay (and cavities), and tooth loss. Beyond the pain and suffering, oral health problems affect the ability to eat and swallow, speak and socialize.

Authors' addresses: Sayma Akther, sakther@memphis.edu, University of Memphis, Memphis, Tennessee, 38152; Nazir Saleheen, nsleheen@memphis.edu, University of Memphis, Memphis, Tennessee, 38152; Mithun Saha, msaha1@memphis.edu, University of Memphis, Memphis, Tennessee, 38152; Vivek Shetty, vshetty@ucla.edu, University of California, Los Angeles, Los Angeles, California, 90095; Santosh Kumar, skumar4@memphis.edu, University of Memphis, Memphis, Tennessee, 38152.

2474-9567/2021/6-ART53 $15.00
https://doi.org/10.1145/3463494

Importantly, because the mouth is the main portal for entry to the body, poor oral health can contribute to a range of conditions and diseases including respiratory diseases, endocarditis, cardiovascular diseases, and pregnancy and birth complications. What makes matters even worse is the accompanying steep cost of dental and health care, that many without insurance struggle to bear. However, the good news is that people can still prevent much of the complications arising from poor brushing habits through technology-driven awareness.

Smart toothbrushes [3] equipped with Bluetooth connectivity, gyroscopes, and accelerometers are beginning to address some key aspects of oral hygiene. They use beeps, vibrations, and visualizations on smartphones to reinforce a recommended routine of spending 30 seconds on each quadrant — upper right, upper left, lower right, and lower left — for adequate brushing, a key component of proper dental care [7, 26]. Identification and evaluation of toothbrushing activities coupled with a feedback system to encourage proper brushing has been a focus of several works on understanding and improving human oral health behavior. They include assistive technologies to promote brushing habits among children through playful experiences [9, 21]; supporting users in learning a complex brushing technique with realtime feedback [11, 16, 17]; encouraging regular toothbrushing using virtual aquarium or mirror [28, 29]; enabling self-examination and creating awareness about common oral health conditions [25]; creating plaque awareness [36]; and helping handicapped people without arms to brush their teeth correctly [4]. But, these works use either smart toothbrushes, electric toothbrushes, or toothbrushes fitted with sensors [11, 12, 15, 28]. As such toothbrushes are still used by a small minority, these advances do not benefit most people who still use manual toothbrushes.

Wrist-worn inertial sensors in smartwatches and activity trackers are increasingly being used to detect various activities like eating [33], smoking [31, 32], drinking [13], and hand washing [24]. A recent work [5] presented the mORAL model to detect the start and end times of brushing and flossing activities. Although this work enables monitoring of brushing and flossing events for a large population of users still brushing with a regular toothbrush, the capability of monitoring which surfaces are not adequately being brushed is still lacking.

In this paper, we present a new *mTeeth* model to detect which tooth surface is being brushed using a regular uninstrumented toothbrush from data collected by inertial sensors in wrist-worn activity trackers and smartwatches. We successfully address several challenges in detecting brushing on specific tooth surfaces, which receive only a few seconds of brushing before a user transitions to another surface.

First, we enhance the utility of a publicly available wrist-worn inertial sensor dataset collected from daily life of participants by annotating it with fine-grained labels of which surface is being brushed on and moments of transition. We develop a hierarchical categorization of teeth surfaces in nine types that is suited to detection by sensors. We analyze the labeled data to quantify between-person variability in brushing patterns, within-person between-episode variability, and within-episode between-surface variability in brushing duration.

Second, we find that there are time synchronization errors of several seconds between sensor data and associated video, even though both are collected on the same smartphone. As transition among teeth surfaces last only milliseconds, we propose an algorithm to tightly synchronize the two data sources that does not have any explicit anchor event. We find that this improves the F1 score for surface classification by 13%.

Third, we observe that time spent on a brushing surface can be as low as a few milliseconds and as high as few tens of seconds. This prevents unambiguous label assignment in fixed-length window-based approach to data segmentation. We observe that an anchor micro-event called *brushing stroke* occurs during all surface transitions. We propose a computationally lightweight method to identify brushing strokes using wrist-worn inertial sensors.

Fourth, we identify and compute several features from each brushing stroke. To leverage the hierarchical organization of teeth surfaces and sequence of transitions among them, we select and train a Dynamic Bayesian Ensemble model. We train and test on one week of brushing data from 19 participants to analyze the impact of wide between-person and within-person variability on the performance of machine learning models for dynamic brushing surface identification using wrist-worn inertial sensors.

2 RELATED WORKS AND KEY CONTRIBUTIONS

Our proposed mTeeth model assumes that the start and end of a brushing event can be identified from wrist-worn inertial sensors. Development of such models has progressed from detecting brushing from hand gestures in the context of detecting a vast amount of activities of daily living (ADL) [8, 14] in scripted settings to recently proposed models for automatically detecting toothbrushing in the wild using wrist-worn inertial sensors [5, 27]. In the following, we discuss prior works that aim to detect the specific teeth surface being brushed on.

2.1 Toothbrushing Surface Detection from Smart or Instrumented Toothbrushes

In [22] and [23], the authors designed a smart toothbrush fitted with a 3-axis accelerometer and magnetometers to trace which group of teeth the user was brushing at a particular moment. This work divided the teeth into several brushing regions before developing a k-means clustering-based model to detect them [22] and determine if brushing in each of those areas was done appropriately or not. Their smart toothbrush based approach achieved an overall accuracy of 97.1% for a total of 15 brushing regions. Smart toothbrush based solutions are now commercially available that guide users on the correct way of brushing. For example, [3] includes a brushing head capable of giving real-time feedback to the user based on brushing pressure. A paired smartphone provides visual display to determine which surface is being brushed.

To provide an alternative to smart toothbrushes, [15] proposed a smartwatch based recognition system to evaluate the brushing quality. They attached magnets to a normal toothbrush to build an arm motion model with inertial data collected from wrist-worn sensors for real-time detection of brushing gestures. The system was able to detect brushing surfaces with an average precision of 85.6% by dividing the teeth set into 16 different surfaces following the Bass technique. In [28], a 3D colored ball was attached at the tail of a toothbrush to estimate which dental side was being brushed by analyzing the spatial position and orientation of the ball. As these methods rely on sensors in a toothbrush, they are not applicable to detecting brushing surfaces with regular toothbrushes.

2.2 Toothbrushing Surface Detection from Audio and Video

An initial work [18] evaluated brushing from acoustic signals captured by a smartphone placed next to the sink. It recorded audio signals from which 12-order Mel-Frequency Cepstral Coefficient (MFCC) features were extracted to train a Hidden Markov Model (HMM) for recognizing toothbrushing activities. It achieved a classification accuracy of 78.3%. Similarly, [30] proposed a tooth brushing monitoring system based on acoustic inputs. They deployed an asymmetrical sound-field detector which had a Bluetooth earphone and a throat microphone to capture acoustic inputs from the air and human body, respectively. The two different sources of inputs carried a rich set of characteristics from the environment and a living entity. To reduce computational complexity, a series of statistical inferences from time and frequency domains were extracted for training different models.

Some works use image analysis to detect teeth surfaces. A computer based web-cam is used in [21] to identify the position of the smart toothbrush. It has a visual feedback system, equipped with a physical avatar whose teeth are made of LEDs for tracking children's tooth-brushing activities in real time. Another work [28] detects toothbrush and the face of its user with the help of a smartphone's front camera. The smartphone's display works as a "virtual mirror" to locate a person's face with a toothbrush through a face tracker and replaces the captured image with that of an avatar. The avatar is able to completely mimic the user's gestures and expressions, and points out any wrong movement. As these works rely on some instrumentation of the environment to detect teeth surfaces, their methods are not directly applicable to address the technical challenges faced in detecting teeth surfaces being brushed from wrist-worn inertial sensors alone.

2.3 Toothbrushing Surface Detection from Wrist-Worn Inertial Sensors

Even though [15] used an instrumented toothbrush, they also trained a model to detect brushing surfaces using only wrist-worn inertial sensors, that was further improved by [27]. Sensor data is divided in 1 second segments

in [15] and 1.2 second segments in [27]. They used either a video or an observer to decide the surface labels of each data segment. For labels, [15] used 16 Bass technique surfaces, while [27] used 13 teeth surfaces, tongue brushing, and raised hand state. The number of episodes or the amount of data collected is not reported in either work. Also, details of how second-level precision was achieved in labeling either from video or observer are missing. Both trained their participants to brush using the Bass technique. Any brushing sequence not following the Bass technique are excluded from surface classification in [15]. Naive Bayes classifier together with a Hidden Markov Model is trained in [15], while an attention-based LSTM is used in [27]. A precision of 75.9% with wrist-sensor only model is reported in [15], while an accuracy of 97% is reported in [27], both using 10-fold cross-validation.

Our work differs from [15, 27] in several respects and presents an alternative approach to surface classification. First, the goal of both prior works was to achieve homogeneity in the brushing pattern of participants by training them. Our goal instead is to observe the natural brushing habits of participants and still aim to detect the surface being brushed and transitions among them, despite natural variability. Second, a fixed window of 1 or 1.2 seconds can include brushing on two different surfaces and the transition time, which creates ambiguity in which labels to assign to these windows. Leaving them unlabelled can exclude 30-40 seconds of data from a 120-second session, as an average of 30 transitions occur in a 90-second brushing session. Therefore, we use a new anchor micro-event (i.e., brushing strokes) that naturally occurs between all transitions among surfaces, and thus separates the data from different surfaces. Third, we find that the number of samples in our data segments (i.e., in a brushing stroke) consists of only 4-5 data points (at a sampling frequency of 16 Hz). They are sufficient to detect peaks and valleys, but are not suitable to train a deep learning or other models that identify complex features automatically. But, the Dynamic Bayesian Ensemble method we present is still able to achieve similar high accuracy (with median F1 scores of 94% to 100%) for distinguishing among in/out, left/center/right, and up/down surfaces.

2.4 Summary of Key Contributions

In summary, the presented work makes the following novel contributions over prior works.

(1) Unambiguous Labeling: Ours is the first work to use the micro-event of *brushing strokes* to assign clean labels to each sensor data segment. Prior works on brushing surface detection used fixed-length windows [15, 27] that may make unambiguous label assignment difficult.
(2) Brushing Stroke Detection: A method to detect brushing strokes using acoustics was presented in [15], with an average error rate of 10.3%. They posed the task of detecting brushing strokes from inertial sensors as an open problem. We successfully solve this open problem with less than 4.2% error.
(3) Tight Time-Synchronization We observe that as brushing strokes and transitions usually last < 300 milliseconds, the sensor data and video (that provides a way to obtain precise labels) needs to be tightly time synchronized. Even though prior works [15, 27] used video to obtain surface labels, ours is the first work to illustrate the label alignment challenge and presents an algorithm to achieve tight time synchronization.
(4) Within-Person Variability: Although between-person variability in brushing patterns have been reported previously in dentistry [10, 35], ours is the first work to report wide within-person between-episode variability, highlighting that the usual approach of single event observation from each participant may not suffice to analyze prevalent brushing patterns.
(5) Between-Person Model Generalizability: We quantify between-person variability in brushing patterns from video data and analyze the challenges it poses in achieving between-person generalizability of machine learning models for brushing surface detection.
(6) Challenges for Personalized Models: Although personalized models require person-specific training, they usually perform better than general models. We show that wide within-person between-episode variability impacts the performance of even personalized models for brushing surface detection.

(7) Brushing Duration Estimation: Ours is the first work to present estimation of the total duration of brushing on each surface in a brushing episode, and report a median absolute error of less than 5%.

3 DATA DESCRIPTION, LABELING, AND KEY FINDINGS FROM LABELED DATA

3.1 Dataset Selection

A wrist-worn inertial sensor data set consisting of labels of start/end of brushing and flossing episodes used in our mORAL [5] study is available publicly. This study recruited participants willing to brush at least twice — once with a manual toothbrush and once with a SmartBrush and floss at least once a day. Each participant wore a MotionSense wristband on each wrist during waking hours for seven days that included a 3-axis accelerometer and a 3-axis gyroscope sampled at 16 and 32 Hz, respectively. A study provided smartphone connected via Bluetooth technology continuously timestamped and logged incoming sensor data. Besides, participants used the phone's front camera to video record themselves (in their homes) during brushing, flossing and/or oral rinsing. The mORAL dataset currently consists of data from 30 participants (15 males, 15 females; mean age 28.5±10.6 years, 2 left handed) who have contributed 197 brushing episodes with a manual toothbrush.

In the public dataset, the start and end times of brushing episodes are annotated from self-recorded videos. But, the original annotations in the mORAL dataset [1] are insufficient for our modeling because it does not include any teeth surface annotations within a brushing episode. We used the original videos from this study to label precise times for when each teeth surface (i.e., groups of teeth portions) was being brushed, including marking of transitions among the surfaces. See Section 3.3 for details of surface definitions proposed.

3.2 Dataset Curation

Out of 197 brushing episodes, videos for some episodes were not usable for stroke-level annotation of surface transitions. First, some participants moved sideways, getting outside the camera range, during brushing. Second, some participants leaned forward to spit out the excess foam and did not revert to an upright posture. Third, some participants leaned the phone against the back wall or against the sidewall. Because the camera was tilted, it was pointing diagonally at the mouth, with their hand blocking a clear view of their mouth. Therefore, it was not possible to unambiguously determine from the video which surfaces participants were actually brushing.

For the above reasons, 83 brushing sessions had to be excluded from this modeling work. We annotated the remaining 114 episodes from 19 participants. For comparison, prior works on analyzing brushing patterns via video used 96 [35] and 101 brushing episodes [10] and prior works on the detection of brushing surfaces used data from 12 [15] and 10 participants [27].

3.3 Organizing and Naming of Teeth Surfaces for Labeling

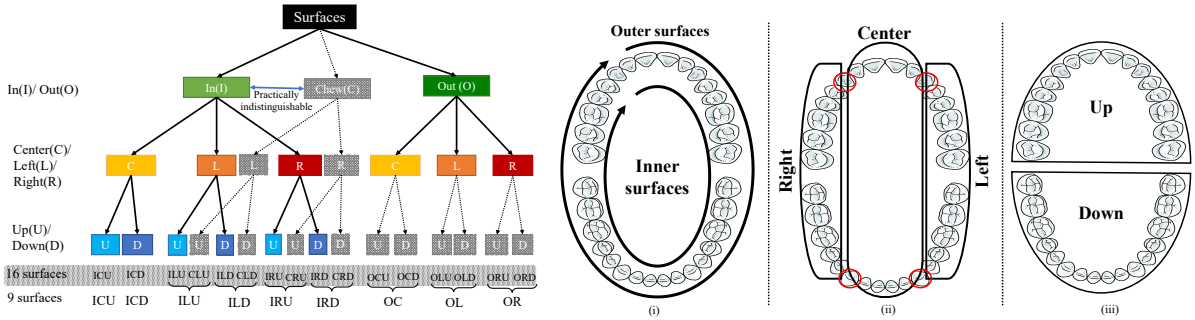
Various works [9, 12, 18–20, 34] organize teeth surfaces between 4 and 16 categories. For teaching brushing, the Bass technique [1, 6, 15, 22, 27] uses 16 surfaces. When observing brushing habits from self-recorded videos [10, 35], surfaces are grouped into fewer broad categories due to ambiguity and frequent transitions among teeth surfaces. We adopt a similar hierarchical organization to obtain nine categories that is suited to sensor-based detection.

We organize teeth surfaces into three layers, as shown in Figure 1a.

Layer 1: In. The inner (tongue facing, i.e., lingual) surfaces of all the teeth and the occlusal (chewing) surfaces of the posterior teeth (premolars and molars) are labelled as the 'In' surface (see the inner arrow in Figure 1b(i)).

Layer 1: Out. The outer surfaces of the teeth (abutting lips and insides of cheeks, i.e., vestibular) are labelled as the 'Out' surface (see the outer arrow in Figure 1b(i)).

[1] https://mhealth.md2k.org/resources/datasets.html#mORAL



(a) Three layer teeth surface naming

(b) Three broader categories of brushing surfaces

Fig. 1. Teeth Surface Categorization: We reorganize the 16 teeth surfaces from the Bass technique into nine surfaces so that all surface transitions are detectable by wrist-worn inertial sensors. The grayed 7 surfaces are merged as shown.

Layer 2: Center. The 'Center' surface encompasses all the anterior incisor teeth (as shown in Figure 1b(ii)).

Layer 2: Left/Right. The posterior region incorporating the premolars/molars on the left and right sides are labelled as 'Left' or 'Right', respectively (see Figure 1b(ii)).

Layer 2: Undecidable. Finally, we place the canines (red) in the 'Undecidable' class. Depending on the brushing pattern, these teeth are dynamically assigned to one of the center/left/right surfaces rather than being apriori assigned all the time. In Figure 1b(ii), red colored teeth are considered as 'Undecidable'.

Layer 3: Up/Down. We define surfaces of teeth in the upper jaw as the 'Up' surface and surfaces from the lower jaw as the 'Down' surface (as shown in Figure 1b(iii)).

Of the 16 surfaces used in the Bass technique, we are unable to disambiguate brushing on chewing surface from inner surfaces due to frequent overlap, resulting in merging of 8 surfaces (chewing and inner) into 4 (inner) surfaces. Additionally, when brushing on the outer surfaces, we are unable to disambiguate brushing on upper and lower surfaces, due to frequent switching and overlap, resulting in merging of 6 (outer up and down) surfaces into 3 (outer) surfaces. Therefore, we end up with nine surface categories.

For naming of these nine surfaces, as we descend from Layer 1 to Layer 2 (in Figure 1a), and then to the leaf nodes in Layer 3, we concatenate the respective categories in each layer to derive the name for each leaf surface. For example, if we traverse the nodes in the order In->Center->Up from Layer 1 to Layer 3, we get the *In-Center-Up (ICU)* surface. Names of the other eight surfaces are: In-left-up (ILU), In-right-up (IRU), In-center-down (ICD), In-left-down (ILD), In-right-down (IRD), Out-center (OC), Out-left (OL), and Out-right (OR).

3.4 Determining the Timings of Teeth Surface Being Brushed On and Transitions from Video

We analyze the video recordings to annotate the start/end times of each teeth surface being brushed on. We precisely mark the transition among surfaces so that data labeled for a surface is not contaminated by any transition data, resulting in unambiguous and clean labels for model training and testing. This is arduous and time-consuming because the duration of brushing on any surface is quite short (less than 5 seconds) and transitions are rapid (lasting few hundred milliseconds) and frequent (tens of transitions in a brushing episode).

Since achieving precision at such granularity is harder for human eyes, we used ELAN [2], a freely available software for assistance in labeling the surface and transition times. We developed the following coding definitions for this labeling to correspond to hierarchical naming of surfaces. We annotated each of the three layers in our

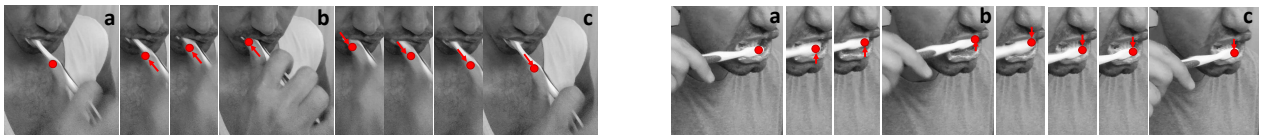


Fig. 2. Frame-by-frame annotation of left-right-left brushing stroke (in the left set of frames) and up-down-up brushing stroke (in the right set of frames). Frames (a) in both stroke types mark the start of the stroke, Frames (b) mark the end of half stroke, and Frames (c) mark the end of the full stroke.

naming hierarchy — Pass-I (Inner and Outer), Pass-II (Center, Left, and Right), and Pass-III (Up and Down). For each of these, we decide the start and end time as described below. Figure 2 shows two frame-by-frame examples.

Begin time: We assign the start time to the moment whenever a participant touches and starts to go back and forth or up and down with the brush in a periodic motion in any one of the In/Out/Center/Left/Right/Up/Down surfaces for the first time or every time after a transition from the previous surface.

End time: We assign end time to the moment whenever the participant stops the periodic back and forth or up and down motion with the brush at the current surface and begins to leave the surface by changing the motion.

To distinguish a surface from transition, we annotate a surface only if it receives at least three brushing strokes.

Switching interval: We use the following criteria for declaring a transition.

(1) When the participant, in a bid to move to the next surface, starts rotating the brush holding wrist to till the rotation stops, and the wrist is in a position from where it can start brushing the next surface.
(2) When the brush holding wrist enters the junction of any two surfaces to when it leaves.
(3) When the wrist holding the brush discontinues the periodic back and forth or up and down motion and slowly takes either a single forward or backward motion.
(4) When the wrist holding the brush suddenly stops brushing the current surface.

Two independent coders labeled all videos to annotate the start and end time of brushing on each surface. The duration of surface transitions is usually < 300 milliseconds, and our goal was to annotate the timings at the stroke-level precision. Therefore, instead of using 0.96 seconds [35], we consider annotations from two coders to match only if the discrepancy for any surface is less than a half-stroke, i.e., 150 milliseconds. We observe 342 discrepancies out of 10,230 surface annotations (3.34%). Discrepancies were resolved via joint viewing of the segment in doubt, and a consensus was reached regarding the labeling of the event in consideration.

4 KEY OBSERVATIONS FROM LABELED DATA

Brushing patterns from videos have been analyzed in dentistry during habitual brushing [35] and best-effort brushing [10] to assess shortcomings and to find ways to further improve brushing habits. As these works invited participants to the study site and recorded one episode from each participant, their observations largely focused on between-person variability. In contrast, we ask participants to video-record themselves in their homes without any explicit instructions, providing us repeated measurements from the same person in their natural environment. This data allows us to analyze within-person and within-episode variability during habitual brushing.

4.1 Between-Surface Variability in the Time Spent on Brushing Different Surfaces

Table 1 shows the mean and standard deviation duration of brushing on each of the nine surfaces. Figure 4 shows detailed distribution for each episode from each participant. Similar to [35], we find that the duration of brushing on left and right sides (across both inner and outer surfaces) are similar. But, we find that the total effective duration of brushing in our dataset is significantly lower at 92 seconds, compared with 155 seconds in [35] and 207 seconds in [10]. We make several new observations regarding between-surface variability.

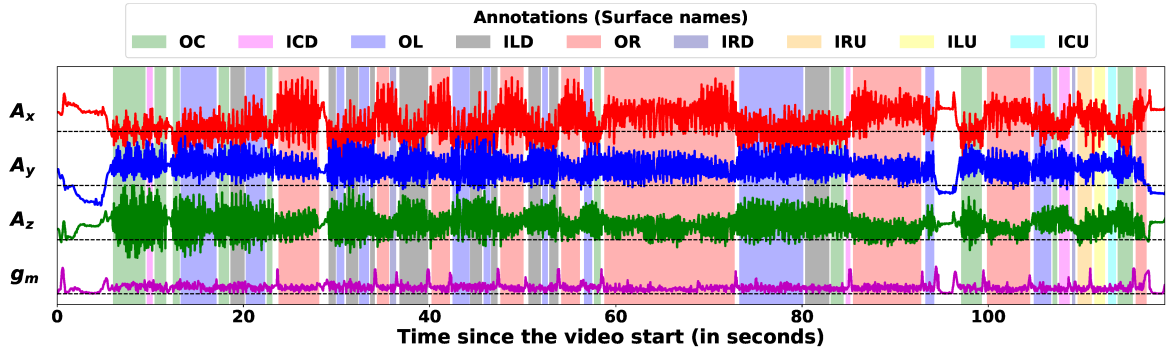


Fig. 3. Three axes of accelerometer and gyroscope magnitude during brushing at different annotated surfaces.

Table 1. Average duration (in seconds) of brushing on each of the nine surfaces

	Surfaces									Effective
	ICU	**ICD**	**IRU**	**ILU**	**ILD**	**IRD**	**OC**	**OL**	**OR**	brushing duration
Duration	2.50	3	5.8	5.9	9	9.2	13.3	20	23	91.7
Mean (± SD)	(±0.9)	(±1.1)	(±1.7)	(±1.4)	(±2.2)	(±1.9)	(±2.3)	(±2.5)	(±2.7)	(±3.4)

To test statistical significance, we take pairwise percentage difference in duration between two brushed surfaces, i.e., ratio of brushed surfaces for all the brushing episodes across all the participants. We want to find a value a such that mean of the percentage difference is significantly greater than the value a. Without loss of generality, we assume the mean of percentage differences is positive (otherwise we switch two duration lists). To find the value of a, we perform a left tailed t-test where the alternative hypothesis is mean $\mu < a$. We want to find the maximum a such that using a t-test we can reject the null hypothesis that the mean of the list is a, i.e., $H_0 : \mu = a$.

Observation 4.1. *Participants spend 40% more time brushing their outer (i.e., buccal or labial) teeth surfaces than their inner (i.e., lingual and occlusal) teeth surfaces across both upper and lower jaw. (p-value <0.008)*

Observation 4.2. *Participants spend 75% less time brushing their center (i.e., anterior) teeth surfaces as compared to their left or right surfaces (i.e., posterior). (p-value <0.009)*

Observation 4.3. *When brushing on inner (i.e., lingual) teeth surfaces, participants spend 75% more time in brushing down surfaces vs. up surfaces. (p-value <0.009)*

Observation 4.4. *The duration of the most brushed surface within an episode is 11 to 19 times the duration of the least brushed surface. But, the least- and most-brushed surfaces are not the same in all episodes. (p-value <0.0003)*

4.2 Between-Person Variability in the Brushing Time on Each Surface

Between-person variability in brushing patterns have been reported previously [10, 35]. As Figures 4 and 5 show, we also observe substantial within-person between-episode variability in amount of time spend on brushing surfaces. Our goal here is to quantify these differences to assess the feasibility of developing a common machine learning (ML) model that can work for all users.

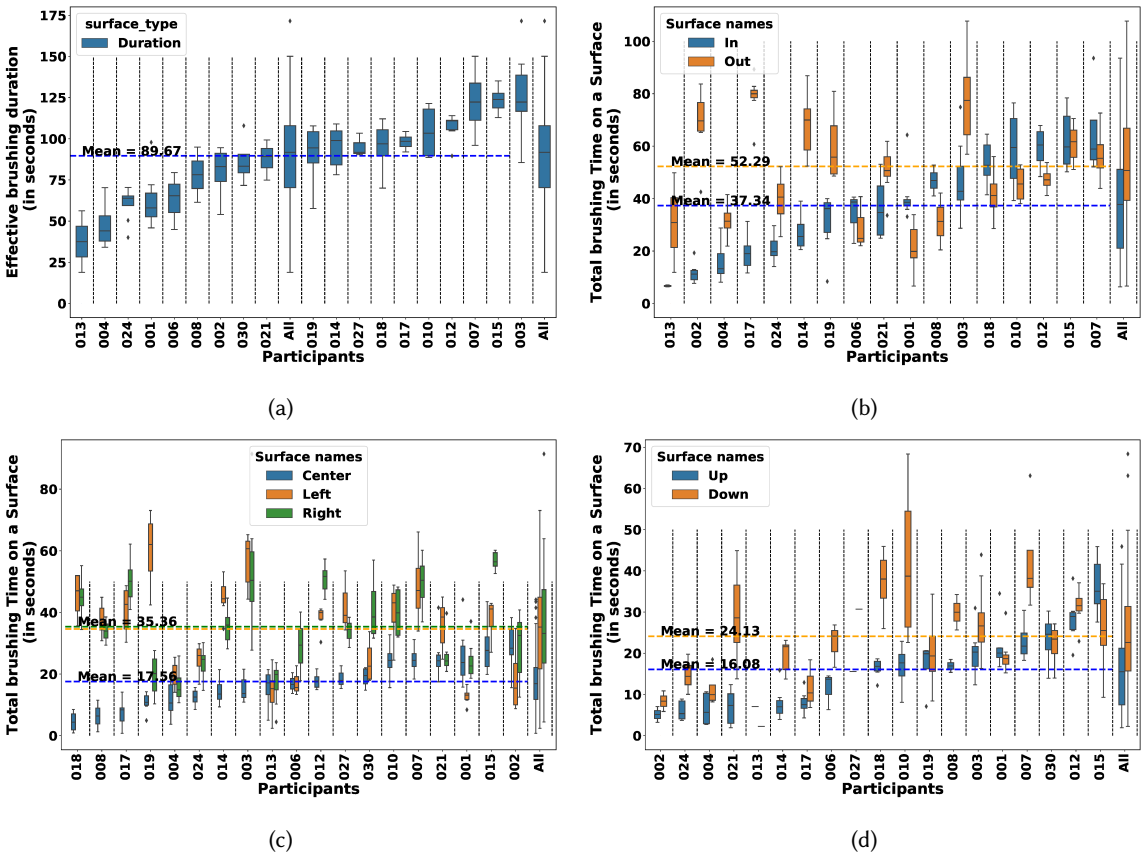


Fig. 4. Difference in total duration of brushing on different teeth surfaces. (a) Effective brushing duration(b) *In* and *Out* surfaces, (c) *Center*, *Left*, and *Right* surfaces, and (d) *Up* and *Down* surfaces.

4.2.1 Similarity of Persons with the Population Profile in Time Spent on Each Surface. First, we quantify how many participants have a brushing duration profile that matches the population average. For that, we represent each brushing episode as a duration vector of all the brushing surfaces, i.e., a nine value vector. This way, we form a list of vectors with one participant's data and combine list of vectors from the rest of the participants to create a population profile. From these two lists of vectors, to find whether a participant's data approximates the population profile, we perform the χ^2-test (Chi-squared test). We repeat this process for all the participants, and the resulting p-values are shown in Figure 6a. We see that only 2 out of 19 participants share profiles similar to the population one. Therefore, population-profile is not representative for most individuals.

4.2.2 Person-to-Person Similarity in Time Spent on Each Surface. Next, our goal is to see if there are clusters of participants sharing a similar profile amongst themselves. We test pairwise independence for all possible pairs of participants. We form two list of vectors from two participants following a similar method mentioned in Section 4.2.1 and perform the χ^2 hypothesis test to find if they are similar to each other. We repeat this process for all possible pairs of participants and present the test results as a heatmap in Figure 6b. Each cell (p_i, p_j) in the figure shows the p-value of the test for p_i and p_j. Only 4 out of 136 pairs show similarities in profile.

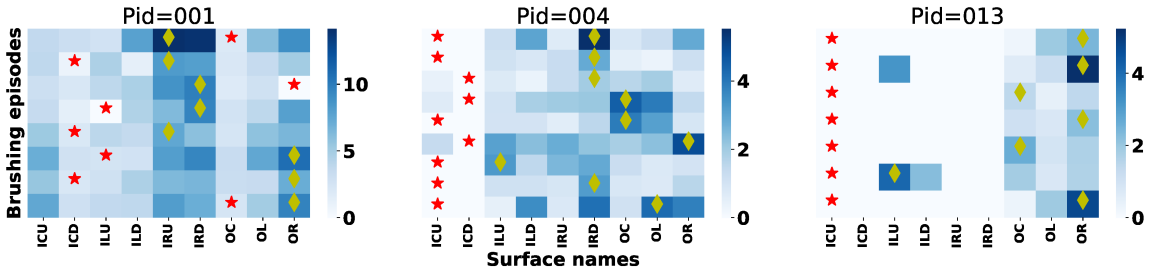
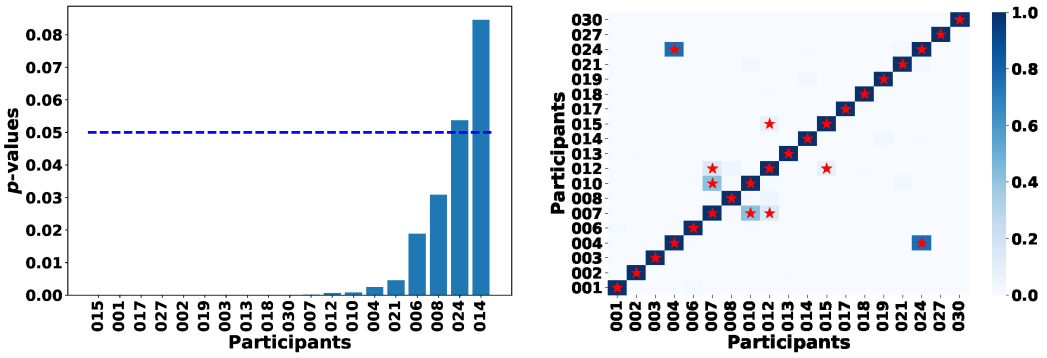


Fig. 5. Illustration of difference in the amount of time participants spend on each surface across different brushing episodes. Data from three representative participants are shown here. The least brushed surface and most-brushed surface within each episode (i.e., each row) are marked with a star and diamond, respectively. Numbers in legends represents surface brushing duration in seconds and darker colors represent longer duration.



(a) (b)

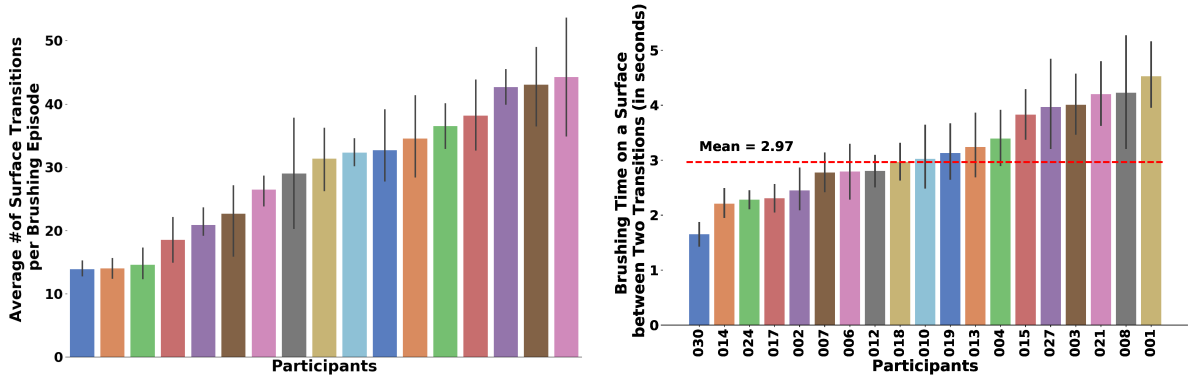
Fig. 6. Between-person similarity in the amount of time spent in each surface: (a) Person-to-population similarity, (b) Person-to-person similarity. Stars show statistical significance.

4.3 Within-Episode Patterns of Transitions Among Surfaces

Prior work [35] has observed preference among participants for frequent transitions among surfaces with an average of 45 transitions in brushing episodes lasting 155 seconds, on average. Figure 7a shows the distribution of average number of surface transitions and Figure 7b shows the average time spent brushing a surface between transitions in our dataset. We observe significant variability in both the frequency of transitions and the time between transitions both between-person and within-person.

5 OVERVIEW OF THE MTEETH MODEL

Figure 8 presents an overview of all the steps in the *mTeeth* model. The input to the model are the inertial sensor data (accelerometer and gyroscope) and the start/end of brushing episodes from a brushing detection model such as mORAL [5]. We first define an *anchor event* (detectable from sensor data) that can be used to segment the time series data cleanly so each segment can receive the unambiguous label of one surface (in Section 6). Subsequently, we develop a method to tightly synchronize sensor data with video so that labels of surface transitions correspond



(a) Number of surface transitions per brushing episode.

(b) Brushing time on a surface between two transitions.

Fig. 7. Surface transition frequency and staying time on a surface between transitions.

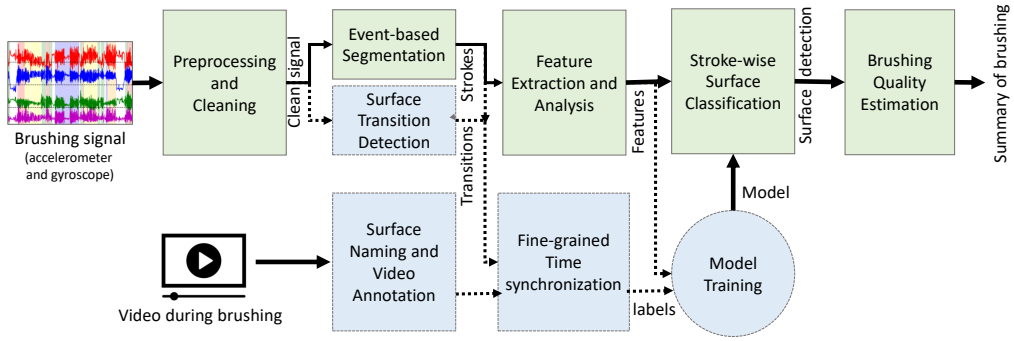
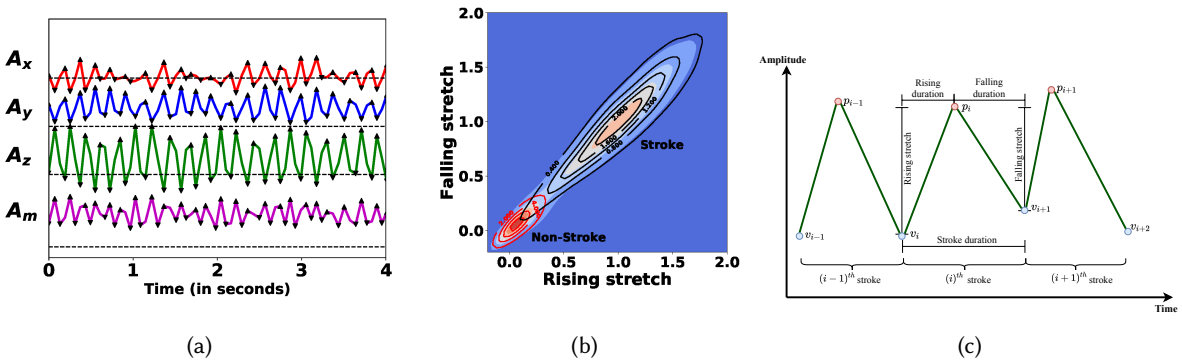


Fig. 8. The pipeline of data processing stages for training and testing the *mTeeth* model.

to the sensor data segments at millisecond precision (in Section 7). Then, we identify and compute event-based features and select distinctive features for each surface (in Section 8). Finally, we train a Dynamic Bayesian Ensemble model to assign each data segment to the most likely surface (in Section 9). This generates a sequence of brushing surfaces in each brushing episodes, with its duration and the number of strokes in it.

6 DEFINING AND DETECTING ANCHOR EVENTS FOR TIME-SERIES SEGMENTATION

There is wide between-person variability in how people brush, including the pattern of back and forth or up and down motion of the brush, time spent in each surface during brushing, and transition sequence among surfaces. In our labeled data, we observe that the time spent on a brushing surface varies from a few milliseconds to as long as 10 seconds. This poses a significant challenge for finding an optimal window of sensor data that can be treated as a single unit of assessment from which features can be extracted to train a machine learning model. The traditional approach of sliding or fixed time-based windowing is unlikely to work. If we choose a window size of few milliseconds to deal with the short duration in some surfaces, we may end up with insufficient data to find distinguishable feature(s). If the window size is too large, several transitions occurring within it may go



(a) (b) (c)

Fig. 9. (a) Brushing Strokes. (b) 2D Gaussian Kernel density function of rising stretch and falling stretch. (c) Stroke-wise computed features from the accelerometer signal.

undetected, resulting in missed transitions and mixing of surfaces in one window, creating both ambiguity and mismatch in label assignment. Therefore, we seek a dynamic-length event-based approach to data segmentation.

For an event-based approach to succeed, we need anchor events such that the likelihood of not detecting this event is low, the event should be efficiently detectable, and the event should cleanly isolate data segment belonging to different surfaces. When developing a model to detect brushing [5], flossing, eating [33], drinking [13], or smoking [31, 32], a hand-to-mouth gesture works as an anchor event. The events of hand reaching the mouth and hand coming back from the mouth isolates segments of sensor data that can be treated as a candidate for each of these hand-to-mouth gesture events and can be tested by the respective machine learning models. But, the hand-to-mouth gesture only occurs at the start and end of a brushing event and hence can't be used to segment the sensor data within a brushing event to distinguish among various teeth surfaces.

For our purpose, we need to find an anchor event that clearly demarcates when brushing surface changes, this event occurs during most surface transitions, and the event is efficiently detectable from sensor data. As our goal is to find the start and end times of brushing on each surface, the transition from one surface to another initially appears to be an obvious choice for an anchor event. But, the transition itself is so short-lived that it is improbable to detect some of the transitions from sensor data. Moreover, some of these transitions are difficult even to annotate from the video. Thus, accurate detection of all transitions is quite challenging, and failure to do so results in mixing of data from two or more surfaces. Therefore, transitions do not qualify as the anchor events.

We observe that there is one activity that is both potentially discernible in sensor signals and is common across all surface transitions. During brushing, people generally perform back-and-forth or up-and-down periodic motion with the brush. We select this movement activity, called a *brushing stroke,* as our anchor event. From five types of strokes known (e.g., circular) [35], we observe brushing strokes follow either a up-down-up or back-forth-back motion. Since these are the two primary periodic movements a person makes during brushing, if and when a surface is brushed with at least one brushing stroke, the chances of its trace remaining in the sensor data is high, significantly reducing the possibility of a missed stroke. This, in turn, limits the surface identification error. Moreover, no brushing stroke takes place between transitions, preventing the mixing of any two surfaces in any segment of sensor data selected for assessment by a machine learning model for surface identification.

6.1 Brushing Stroke Detection

To efficiently detect brushing stroke from sensor data, we identify the signature of periodic up-and-down or back-and-forth movement in the wrist-worn accelerometer signal in the form of a peak-valley pair.

Figure 9a shows plots of three axes accelerometer signal and its magnitude during one such surface brushing. In the signal time series, we define peak as the point in each cycle where the signal is at its maximum, whereas a valley as where the signal is at its minimum. We mark those peaks and valleys of a signal with black up-pointing and down-pointing triangles, respectively. Let $\mathcal{P} = \{p_1, p_2, ..., p_n\}$ be the peaks, and $\mathcal{V} = \{v_1, v_2, ..., v_n\}$ be the valleys. Once we carefully detect all these peaks and valleys using a peak-valley detection algorithm, we define brushing stroke as a cycle of valley-peak-valley combination, i.e., an i^{th} brushing stroke is $S_i = \langle v_i, p_i, v_{i+1} \rangle$.

Now, in Figure 9a, we observe series of peaks and valleys in all the four signals, but most of them are temporally unaligned across the signals. Since we get four sequences of peak-valley cycles or therefore strokes, out of these four signals, we need to select the one that will represent the start and end times of the brushing strokes optimally. We note that even though the magnitude contains information from all the three axes, it is not a suitable choice due to lack of synchronized alignment across the three axes.

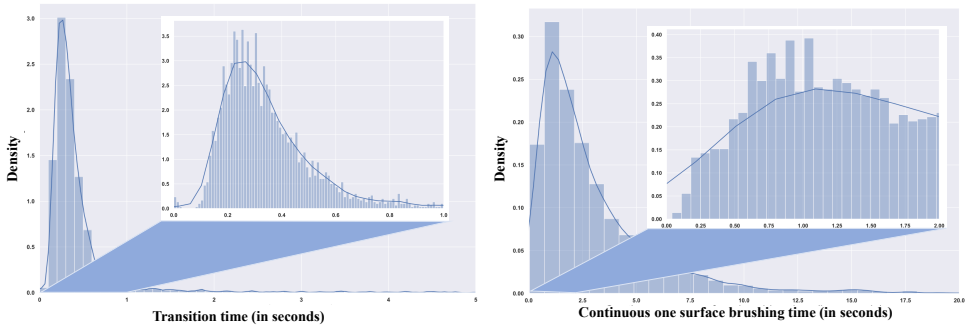
We first define the stretch of a stroke as the difference between the amplitude of its peak and valley. If the stretch of a stroke along any axis is low, that corresponds to having least wrist movement along that axis at that moment. Conversely, if the stretch of a stroke along an axis is high, that signifies a likely wrist movement along that axis during a brushing stroke, making it the dominant axis for this stroke. Hence, we select a brushing stroke along a particular axis that has the maximum stretch. To brush different surfaces, orientation of the wrist changes, so does the movement of the toothbrush along with it. Following the movement, the acceleration of the wrist along a particular axis changes the most. In addition, we observe that the dominant axis remains unchanged throughout brushing on a single surface. When the user switches to the next brushing surface, depending on the type of surface, the dominant axis may either change or continue to be the same.

To distinguish brushing from other activities (e.g., walking) which also involves periodic wrist movement, we define two thresholds, $\mathcal{T}_{dur}$ and $\mathcal{T}_{stretch}$ such that for any $\langle v_i, p_i, v_{i+1} \rangle$ peak-valley cycle to be a brushing stroke, the time difference between v_{i+1} and v_i can be at most $\mathcal{T}_{dur}$ and the stretch needs to be at least $\mathcal{T}_{stretch}$. The average duration of a stroke is $230(\pm 60)$ milliseconds, and the average stretch is $0.57(\pm 0.35)g$. We remove all peak-valley cycles that are two standard deviations away from the mean stroke duration and the mean stroke stretch. These thresholds retain all the brushing strokes in our data, i.e., achieve 100% recall.

7 FINE-GRAINED TIME SYNCHRONIZATION BETWEEN VIDEO AND SENSOR DATA

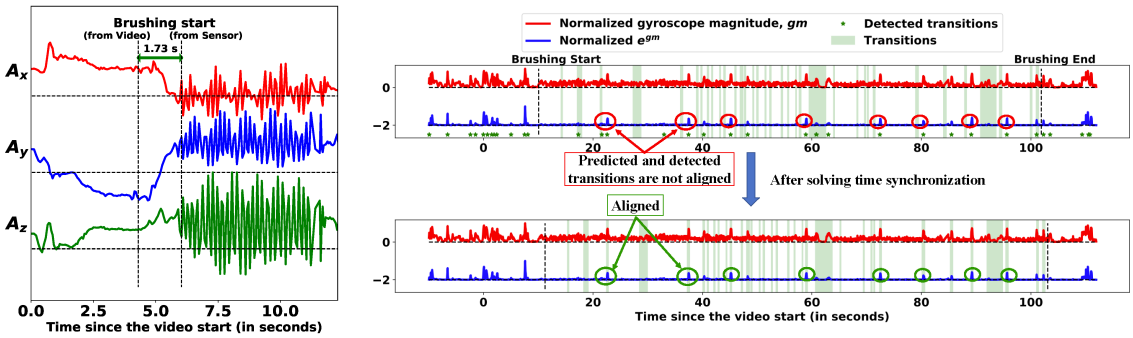
A key premise for our categorization of human teeth into nine surfaces is that the pattern of brushing on each of these surfaces is likely to be sufficiently unique making the corresponding sensor data distinguishable. But, to enable successful modeling for recognizing each of the nine surfaces from sensor data, a brushing episode should include at least a few seconds of brushing on each surface interspersed with milliseconds of surface switching times (Figure 10a). However, in reality (see Figure 10b) some users spend only milliseconds on a surface before switching to another surface. Thus, for accurate estimation of brushing surfaces from such short spans of time, precise time synchronization between the sensor and video data becomes critical. More specifically, we need to synchronize the start of a brushing session extracted from the video data with that of brushing events automatically detected from inertial sensors [5] at millisecond-level precision.

Even though it was assumed in [5] that the mORAL dataset has tight time synchronization between video and sensor data, we find that the video and sensor data have a time synchronization error of several seconds (see Figure 11a for an example). This may be because even though the sensor data from wrist-worn devices were streamed in real-time to the same phone recording the video, time lapse between the sensor data being received on the phone and assignment of a timestamp to them may be of the order of seconds. For the task of detecting



(a) Distribution of transition time (b) Distribution of brushing duration on a surface

Fig. 10. The fine-grain time synchronization becomes the critical requirement because of the quick transition between two surfaces and short staying time in each surface.



(a) (b)

Fig. 11. a) Time difference between brushing start from video and from sensor to show the time synchronization problem. b) Our approach to solve time synchronization. The top figure shows annotation before time synchronization where green start are the detected transitions from the sensor. The bottom figure shows annotation after time synchronization.

the start and end of a brushing session that lasts 2 minutes, few seconds of time lapse may be tolerable. But, for our purposes where the brushing duration on a surface and transition times are only few milliseconds long, time synchronization errors of seconds can render the modeling process extremely challenging. If the lag between video and sensor data is not adjusted properly, part of sensor data which is actually a surface may be mistaken for a transition and vice-versa, or data from different surfaces may get mixed. The performance of a machine learning model will suffer as the quality of these labels drive the accuracy of the model.

7.1 Time Synchronization Problem

We start by defining the time synchronization problem. Let the start time of the i^{th} brushing event, based on video time be t_i^v. From the time at which sensor data is captured by the wrist device to when it reaches the smartphone and receives a timestamp, there is a time lag. We need to find the offset o_i such that when added to

t_i^v, it corresponds to the starting time of the i^{th} brushing event from sensor data t_i^s. Therefore, $o_i = t_i^s - t_i^v$. Since each packet of sensor data contains both accelerometer and gyroscope data, offsets are the same for both.

A brushing event is composed of multiple bushing strokes. Using the brushing stroke detection method discussed in Section 6.1, we can extract the start times of all brushing strokes and use the first stroke from both video and sensor data to synchronize. Since this method is based on accurately locating the first stroke in both video and sensor data, there are at least two cases when this method may fail. First, several participants start brushing before starting the video, missing the first stroke in video. Second, when participants put toothpaste on the brush head, even one up-and-down or back-and-forth movement may create a false first brushing stroke pattern in the sensor data. Therefore, we next propose a more robust method for time synchronization.

7.2 Multi-point Synchronization Approach

We observe that during some transitions from one brushing surface to another, e.g., from left to right, the wrist rotation is significantly higher than that when brushing on any surface. The gyroscope can detect hand rotation, and the contrast in magnitude between brushing and some of the surface switching are clearly identifiable from the gyroscope signal in Figure 11b. Note that for several transitions, the amount of rotation is negligible. But, if we can detect some surface transitions from the sensors, we can map these detected transitions with annotated transitions and find the offset for the synchronization. We build upon this idea to solve the time synchronization problem. Our algorithm consists of three main steps described below.

(Step 1) Rotation-based Transition Detection: During brushing, the wrist moves linearly back and forth or up and down, which are captured by accelerometers. When transitioning from one surface to another, if the wrist holding the brush changes the direction of movement, a rotational change is seen in gyroscope.

To find rotation-based transitions, we first compute the gyroscope magnitude from the 3-axes gyroscope data. Then, we normalize the gyroscope magnitude. To amplify the differences between rotation during brushing and rotation during transition, we take the exponential of each value of the normalized gyroscope magnitude sample. We find a threshold such that if the gyroscope value is higher than the threshold, we consider it as the beginning of a transition. To find the threshold, we first apply the Gaussian Mixture Model (GMM) to find two clusters: one for the lower values (during brushing or stationary) of the signal and the second for higher values (during transition). All the points in Cluster 2 are considered as transitions and time of those points are stored in $\mathcal{T}$.

(Step 2) Candidate Offset Detection: Let $\mathcal{T}_\mathcal{G}$ be the transitions from the video annotation. We seek to maximize the matching between the detected transitions from sensor and video. Therefore, we compute all the possible offset values to identify candidate offsets as $O = \{(t - t_g)\}_{t \in \mathcal{T}, t_g \in \mathcal{T}_\mathcal{G}}$.

(Step 3) Selecting the Best Offset: To find an offset that maximizes the number of matching, we find total matchings for each candidate offset value. We align the timing of the detected transitions by adding a candidate offset to the timestamp of each transition. We then find the closest distance from the marked transitions from video (i.e, ground-truth). If this distance is $< \epsilon$, we consider it a match. Then, we compute the number of detected transitions with a match. Finally, we select the offset that maximizes the number of matching to align the timestamp between video and the sensor data.

8 STROKE-WISE FEATURE EXTRACTION AND SELECTION

After identifying brushing strokes as events within a brushing episode to segment the sensor data stream, we identify and compute several features from sensor data comprising each brushing stroke. We identify those features that are expected to vary during brushing of different surfaces, contributing to successful differentiation among each of them from the sensor data using a trained machine learning model.

8.1 Accelerometer Features

In Figure 9c, we define brushing stroke i as a tuple of three points in a signal. Let $(time(v_i), value(v_i))$ be the (timestamp, amplitude) of the valley v_i and $(time(p_i), value(p_i))$ be the (timestamp, amplitude) of the peak p_i of the i^{th} brushing event. We identify 8 distinct features that are computed from the accelerometer 3-axes signal.

- **Peak Amplitude:** Peak amplitude corresponds to the amplitude value ($value(p_i)$) in each stroke duration, where the signal is at its maximum.
- **Valley Amplitude:** Valley amplitude corresponds to the amplitude value ($value(v_i)$) in each stroke duration, where the signal is at its minimum.
- **Rising Stretch:** Rising stretch is defined as the difference in amplitude of the peak ($value(p_i)$) and the valley immediately appearing before it ($value(v_i)$) of the i^{th} brushing cycle/stroke duration (see Figure 9c).
- **Falling Stretch:** Falling stretch is defined as the difference in amplitude of the peak ($value(p_i)$) and the valley immediately following it ($value(v_{i+1})$) of the i^{th} stroke duration (see Figure 9c).
- **Rise-Fall Ratio:** Rise-Fall ratio is defined as the ratio of rising stretch to the falling stretch.
- **Rising Duration:** Rising duration corresponds to the time elapsed from a valley of a stroke duration, to the subsequent peak (see Figure 9c).
- **Falling Duration:** Falling duration corresponds to the time duration between a peak and the subsequent valley in a stroke duration (see Figure 9c).
- **Stroke Duration:** Stroke duration is the sum of rising and falling duration.

We compute the above eight time-domain features for each of the 3-axis and magnitude signal of the accelerometer for a brushing stroke, resulting in a set of 32 features. In addition to these features, we compute **Correlation** measure that expresses the extent to which two variables are linearly related. As a result, three more correlation features among X, Y and Z axes are added, namely **corrXY**, **corrYZ**, and **corrZX**. In total, we have a set of 35 features computed for each brushing event or stroke.

8.2 Gyroscope Features

In addition to the accelerometer features, we also compute several features from gyroscope data. Since the gyroscope captures the amount of rotation in each axis, which is used to capture the surface switching/transition, we compute several statistical features, such as **mean** and **standard deviation**, to obtain the transition and the amount of rotation within each stroke. In total, we compute six features from three axes.

8.3 Orientation Features

The wrist's orientation with respect to gravity during brushing varies from surface to surface because of the position of the surface and angle of the wrist with the elbow. Recall that a brushing stroke consists of one forward movement (from the valley to peak in the signal) and one backward movement (from peak to next valley). During these two movements, the wrist has linear acceleration, but at the peak, the wrist gets stable, i.e., no linear acceleration, and prepares to move in the other direction. To capture the wrist's orientation, we compute **roll**, **pitch**, and **yaw** when the wrist is at the peak, i.e., at a stable state.

9 MODEL SELECTION AND TRAINING

During routine dental care, people generally initiate brushing sequence with the outer surface, followed by the inner surfaces. We observe a similar pattern among the study participants where they start and more importantly, cover all the portions of the outer surface first before moving onto the inner surface. To capture the natural layered hierarchy that is also captured in our organization of teeth surfaces (i.e., in/out, left/right/center, and up/down) as well as sequence of transition from one surface to the next, we select a hierarchical model that allows leveraging of any sequence patterns. We train a Hierarchical Bayesian Network for our model training.

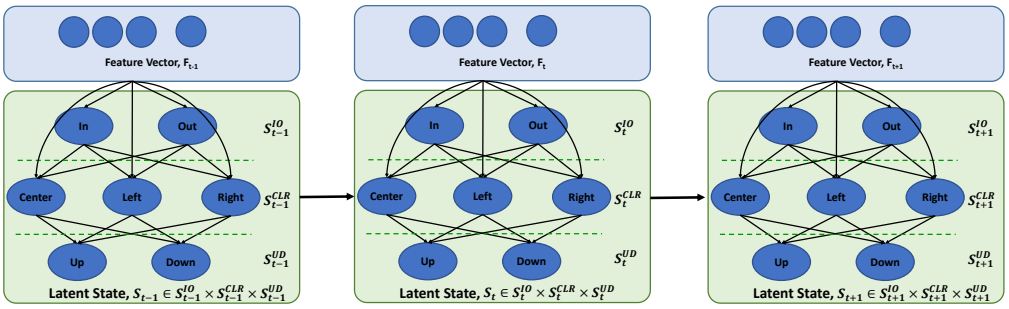


Fig. 12. Bayesian network with state transition

9.1 Bayesian Ensemble Method

A Bayesian network is a type of probabilistic graphical model that uses Bayesian inference for probability computations. Bayesian network aims to model conditional dependence, and therefore causation, by representing conditional dependence through edges in a directed graph.

Architecture of the Bayesian network for our brushing surface detection problem is shown in Figure 12. We organize the nine surfaces in the three surface layers as $S^{IO} = \{I, O\}, S^{CLR} = \{C, L, R\}, S^{UD} = \{U, D, *\}$, with $*$ denoting the ambiguity between up and down for the outer surfaces. Since any surface label is a combination of nodes from the tree layers, the class label set for the nine brushing surfaces is, $s \in \mathcal{S} \subset S^{IO} \times S^{CLR} \times S^{UD}$.

Now, for a given feature vector $f \in F$ of a brushing stroke, the model computes the likelihood of each surface label $s \in S$ as the predicted class using conditional probability $Pr[S = s|F = f]$. The model then outputs the class with the maximum probability as the final prediction for the brushed surface, i.e., $s = argmax_{s \in S} Pr[S = s|F = f]$.

Inference: For any feature vector f of brushing stroke, to compute probability of any surface $s \equiv (x, y, z)$, where $s \in S$, $x \in S^{IO}$, $y \in S^{CLR}$, and $z \in S^{UD}$, we use the following joint probability distribution function,

$$\begin{aligned} Pr[S = s|F = f] &= Pr[S^{IO} = x, S^{CLR} = y, S^{UD} = z|F = f] \\ &= \big(Pr[S^{IO} = x|F = f] \times Pr[S^{CLR} = y|F = f, S^{IO} = x] \\ &\quad \times Pr[S^{UD} = z|F = f, S^{IO} = x, S^{CLR} = y]\big) \end{aligned} \tag{1}$$

For example, probability of surface label $S = ICU$ is computed as,

$$\begin{aligned} Pr[S = ICU|F = f] = \big(&Pr[S^{IO} = I|F = f] \\ &\times Pr[S^{CLR} = C|F = f, S^{IO} = I] \\ &\times Pr[S^{UD} = U|F = f, S^{IO} = I, S^{CLR} = C]\big) \end{aligned} \tag{2}$$

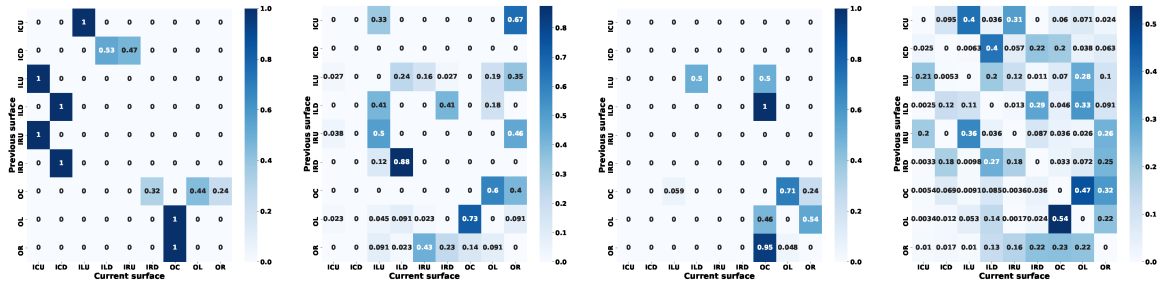
To compute these conditional probabilities, we learn a machine learning classifier–Random-forest model in each layer (a brief description of all the models is listed in Table 2). We then ensemble the outputs of these machine learning models using Equation 1 to produce the final output of the model.

9.2 Dynamic Bayesian Ensemble (DBE) Method

Despite the wide variability in the brushing duration on each surface, we also observe stable patterns in surface transitions [15] for most of the participants, as shown in Figure 13. Dynamic Bayesian Ensemble (DBE) method uses the transitions to update the probabilities when it computes the probability of a surface that is different from the previously detected surface. Let T^* be the transition probability matrix, where each $T^*_{i,j}$ is the transition

Table 2. All the models that generate all required conditional probabilities

Models	Tasks (Generates probabilities of surfaces)	Outputs
$M_{IO}(f)$	'in' and 'out' from feature vector f of a stroke	$\langle Pr[S^{IO} = I\|f], Pr[S^{IO} = O\|f]\rangle$
$M_{CLR\|I}(f)$	'center', 'left', and 'right' from f given $S^{IO} = I$	$\langle Pr[S^{CLR} = C\|I, f], Pr[S^{CLR} = L\|I, f]$, $Pr[S^{CLR} = R\|I, f]\rangle$
$M_{CLR\|O}(f)$	'center', 'left', and 'right' from f given $S^{IO} = O$	$\langle Pr[S^{CLR} = C\|O, f], Pr[S^{CLR} = L\|O, f]$, $Pr[S^{CLR} = R\|O, f]\rangle$
$M_{UD\|I,C}(f)$	'up' and 'down' from f given $S^{IO} = I$ and $S^{CLR} = C$	$\langle Pr[S^{UD} = U\|C, I, f], Pr[S^{UD} = D\|C, I, f]\rangle$
$M_{UD\|I,L}(f)$	'up' and 'down' from f given $S^{IO} = I$ and $S^{CLR} = L$	$\langle Pr[S^{UD} = U\|L, I, f], Pr[S^{UD} = D\|L, I, f]\rangle$
$M_{UD\|I,R}(f)$	'up' and 'down' from f given $S^{IO} = I$ and $S^{CLR} = R$	$\langle Pr[S^{UD} = U\|R, I, f], Pr[S^{UD} = D\|R, I, f]\rangle$



(a) Participant '001' (b) Participant '004' (c) Participant '013' (d) All participants

Fig. 13. Consistency in state-to-state transitions

probability from surface i to surface j, where $i, j \in S$. We use the $*$ in a symbol to denote that the states can be over all nine surfaces or only over the groups of surfaces. We end up with four transition matrices, one for all nine surfaces and one each for the three layers. Note that we only consider transition probability when the current surface is changed, i.e., $T_{i,i} = 0$. Therefore, the updated probabilities are computed as follows,

$$Pr'[S^*_t = x|F_t = f_t, S^*_{t-1}] = \begin{cases} Pr[S^*_t = x|F_t = f_t] & , \text{if } S^*_{t-1} == x \\ \alpha * Pr[S^*_t = x|F_t = f_t] + (1-\alpha) * T^*_{y,x} & , \text{else if } S^*_{t-1} == y, \forall y \neq x \end{cases}$$

Here, f_t is the feature vector of t^{th} brushing stroke, α is the parameter of the weighted average of two values, and $Pr[S^*_t = x|F_t = f_t]$ is computed using Equation 1. We use Pr' to denote the updated probability.

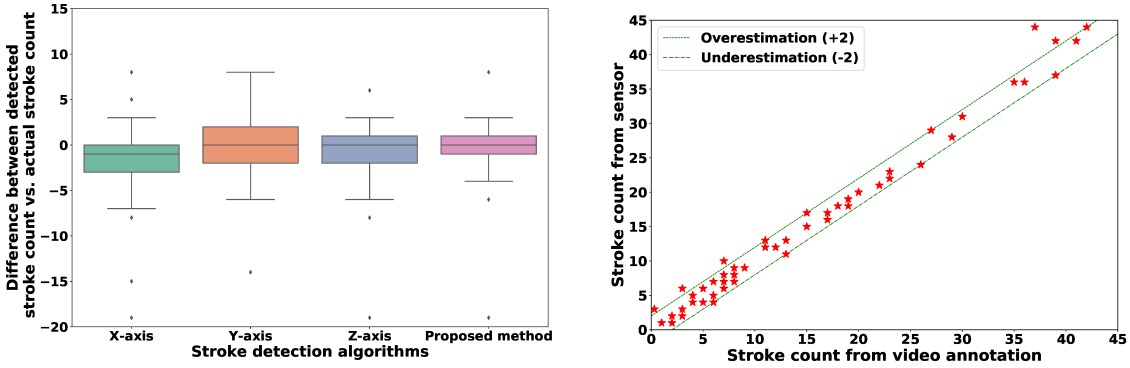
Inference: The selected class of t^{th} brushing stroke is given by

$$s_t = \arg\max_{s \in S} Pr'[S_t = s|F_t = f_t, S_{t-1} = s_{t-1}],$$

where $\langle f_1, f_2, ..., f_m\rangle$ denotes a sequence of features. The model produces the surface sequence, i.e., $\langle s_1, s_2, ..., s_m\rangle$.

10 MODEL EVALUATION

The dataset we use confirms the wide between-person variability reported in dentistry [10, 35]. Additionally, it reveals substantial within-person between-episode variability not analyzed in prior works due to lack of such data. Recent works on detecting brushing patterns from wrist-worn inertial sensors [15, 27] collected multiple



(a) Distribution of stroke count differences between detected stroke counts and ground-truth

(b) Scatter plot of actual count vs. detected count of brushing strokes in a brushing surface

Fig. 14. Performance of the stroke detection method.

episodes from the same participants, but used 10-fold cross-validation. Hence, between-person generalizability of a machine learning model for detecting brushing surfaces has not yet been studied. The dataset we use has a larger number of episodes compared with [10, 35], more participants as compared with [15, 27], and is unique in representing natural brushing patterns in the users' home environment, without any specific brushing instructions (as in [15, 27]) that may reduce the natural between-person variability. To evaluate the between-person generalizability of our model, we start with Leave-One-Subject-Out-Cross-Validation (SCV), but also present 10-fold cross-validation (10CV) results to both allow a comparison with recent works on brushing surface detection and to show the impact of between-person variability in natural brushing on the model's performance. In addition, to study the impact of within-person variability among brushing episodes, we also perform Leave-One-Episode-Out-Cross-validation (ECV), where for each participant, we take one brushing episode as test data and the remaining episodes from that participant as training data, yielding a personalized model for each participant.

We start by evaluating the accuracy of detecting brushing strokes, which is a key enabling and distinguishing aspect of our model. Next, we evaluate the performance of our model for surface detection via all three validation methods. Finally, we study the impact of time synchronization on model performance, and conclude our evaluation by reporting the accuracy of estimating the total brushing duration on each surface, that can improve oral care.

10.1 Accuracy in Detecting Brushing Strokes

To evaluate the accuracy of our brushing stroke detection method (see Section 6.1), we compare the number of brushing strokes detected in each brushed surface from sensor data with that from video annotation. As annotating each brushing stroke (lasting only milliseconds) for all the episodes is even more arduous than annotating each brushing surface, we limit our annotation to 1,456 brushing strokes from 100 surfaces. For each surface, we count each and every periodic movement (valley-peak-valley) with no condition imposed on brushing strokes for each of the x, y and z-axis of the accelerometer separately and also through our brushing stroke detection model discussed in Section 6.1. We calculate the difference between the counts of brushing strokes from sensor data and that from video annotation for each annotated episode and present the results in Figure 14a. We observe from the distributions that our proposed method results in the lowest error (mean absolute error is

Table 3. Classification performance for identifying broad teeth surface categories. Median values are reported here.

	$M_{IO}(f)$		$M_{CLR\|I}(f)$		$M_{CLR\|O}(f)$		$M_{UD\|I,C}(f)$		$M_{UD\|I,L}(f)$		$M_{UD\|I,R}(f)$	
	SCV	**10CV**	**SCV**	**10CV**	**SCV**	**10CV**	**SCV**	**10CV**	**SCV**	**10CV**	**SCV**	**10CV**
Recall(%)	77.57	94.05	83.85	94.12	85.82	96.75	86.84	100	65.43	96.55	77.48	97.06
Precision(%)	79.29	94.07	84.09	94.38	88.92	96.86	100	100	81.33	96.87	78.53	97.37
F1-Score(%)	77.4	93.93	82.23	94.15	84.98	96.75	91.67	100	60.98	96.51	72.13	97.07

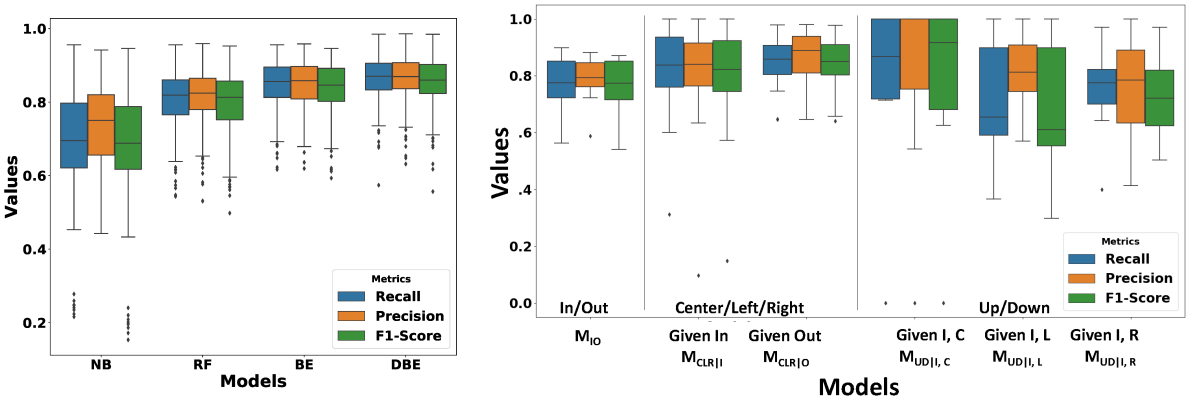
1.5). In Figure 14b, the scatter plot shows the counts of brushing strokes from the video and sensor data using our proposed method. We observe that most of the errors are limited to no more than 2 strokes, even when the number of strokes are as high as 40 (in a brushing surface). We note that [15] estimated the number of brushing strokes using acoustic sensors with an average error of 10.3%, leaving the task of stroke detection using inertial sensors open. Our stroke detection algorithm solves this open problem with less than 4.2% error.

10.2 Impact of Between- and Within-Person Variability on Brushing Surface Detection

We trained Naive Bayes (NB), Random Forest (RF), Bayesian Ensemble (BE), and Dynamic Bayesian Ensemble (DBE) for brushing surface detection (see Figure 15a) and find that the DBE produces the best performance (for 10 CV). Hence, we use DBE for subsequent evaluations. As brushing recommendations are usually based on broad surface category, we begin by evaluating the performance for detecting the three surface layers. Recall that our Bayesian Ensemble model consists of six models each of which is a Random-forest classifier, as shown in Table 2. We evaluate the performance of each model for classification at each teeth surface layer and present the results in Figure 15b. Note that any model in the form of $M_{x|y}$ is trained on a filtered dataset belonging to surface label y from the upper layer. For example, $M_{CLR|I}$ is trained on only the feature set from inner surface. Table 3 shows the results for both SCV and 10CV. The $M_{UD|I,C}$ model for inner-center surfaces achieves the best performance.

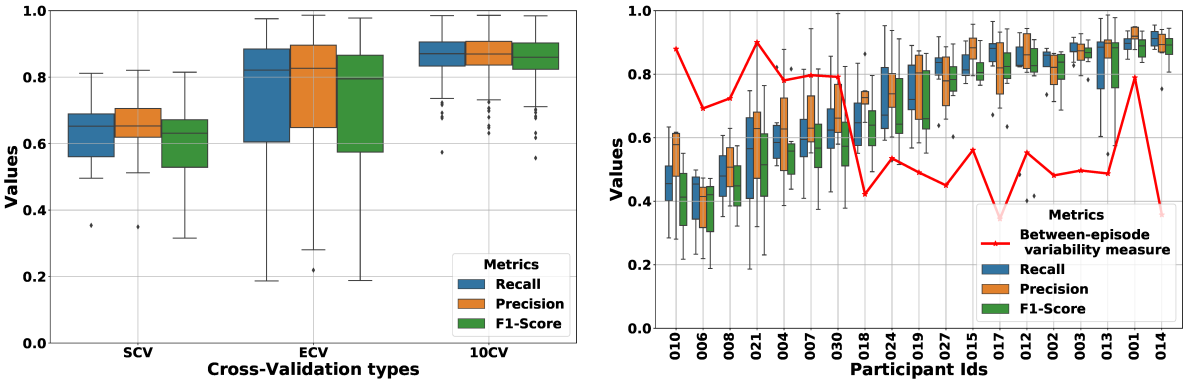
For nine-surface classification, the model obtains median recall, precision and F1-score of 65.26%, 65.30% and 63.14% for SCV, which improves to 82.14%, 82.66% and 79.50% for ECV, and further improves to 87.06%, 86.96% and 86.02% for 10CV. Low performance for SCV as compared with 10CV (used in prior works on brushing surface detection [15, 27]) can be explained by wide between-person variability. As described in Section 4.2, there does not exist a population profile or even clusters of participants with similar brushing patterns (see Figures 5 and 6). Hence, a model trained on other participants' data performs poorly when tested on a different participant.

We observe that training a personalized model for ECV improves the performance substantially from SCV, but still falls short of the 10CV performance due to within-person between-episode variability exhibited in natural brushing habits of participants (see Section 4.2). Figure 16b displays breakdown of the result by participant with individual precision, recall and F1-score to show participant wise between-episode variability. We compute between episode variability as follows: as discussed in Section 4.2, each episode is represented as a duration vector of all the brushing surfaces, i.e., a nine value vector, and we use the Euclidean distance metric to compute the distance between two such vectors. We compute Euclidean distance of all pair-wise combinations of a participant's episodes and take the mean of the distances as a representative of the between-episode variability for that participant. To show the relative variation over the participants, we plot the normalized measurements of all the participants. We observe that between-episode variation in the total time spent in the nine surfaces highly affects the performance of correctly identifying the surfaces. As discussed in Section 4.2, we observe that some participants completely miss some surfaces in many of the brushing episodes due to their personal brushing habits. So, when the model is trained with mostly missing data for a surface from most of the episodes and asked to detect the surface when it is present in the test episode, it fails to do so.



(a) Performance of different classification models (b) Classification performance at each layer (for models defined in Table 2)

Fig. 15. a) The DBE model outperforms other models on 10CV; b) Performance using Leave-one-subject-out cross-validation (SCV) for brushing surface identification for broad categories.



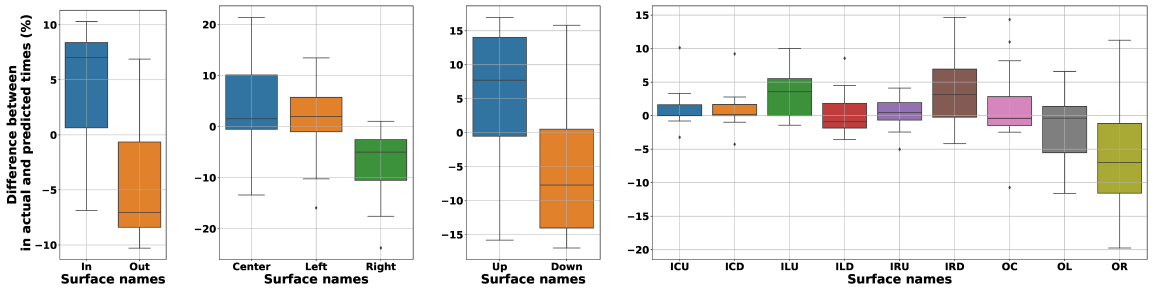
(a) Different types of cross-validation results

(b) Leave-one-episode-out cross validation results

Fig. 16. Nine surfaces classification results for different validation types.

10.3 Impact of Time Synchronization on Nine Surface Classification

We first manually check if the proposed method correctly synchronizes the sensor data to the video-obtained labels. We find that 101 out of 114 episodes (88.59%) are correctly synchronized. We carefully analyze the remaining 16 episodes and observe that wrist movement during brushing is too slow to detect the significant rotation required to identify transitions and make a cluster. We manually synchronized these 16 episodes. Next, to analyze the impact of time synchronization, we train a model without performing the time synchronization step. We find that the surface detection accuracy (using 10CV) drops significantly to recall, precision, and F1-score of 75.09%, 74.02%, and 73.34% respectively, showing a drop in F1-score by almost 13%.



(a) In vs. Out (b) Center vs. left vs. right (c) Up vs. down (d) Nine surfaces

Fig. 17. Accuracy of estimating total duration of brushing in different surfaces

10.4 Accuracy of Estimating Total Brushing Duration on Each Surface

Thus far, we have presented the accuracy of detecting when each surface is being brushed. As we present in Section 4, participants switch frequently between surfaces, coming back to a surface multiple times. For oral health purposes, both users and their providers may be interested in determining the total time a user spends in brushing of each surface in a brushing episode. Figure 17 shows the percent error (as compared with labeled data from video) in estimating the total duration of brushing on surface groups and for each of the nine surfaces.

We observe that the median absolute error is $< 7.5\%$ for in vs. out, $< 2.5\%$ for center vs. left vs. right, $< 7.5\%$ for up vs. down and $< 7\%$ for all of the nine surfaces. Even the first and third quartiles are $< 5\%$ in most cases. We notice that most errors are from confusing some instances of out with in, right with left, and down with up.

Finally, we note that using mTeeth model can improve the estimates of start/end times of brushing (and total duration of brushing, a widely-used clinical variable) by models such as mORAL [5]. Our assumption is that mTeeth will be triggered upon detection of the start of a brushing event by mORAL. mORAL considers the start of the event from when the hand is in the upward direction, which includes putting the paste on the brush-head and preparing to brush; and end of the event when the hand moves to the downward position. Error in start/end times is 4.1% for the mORAL model. Once mTeeth model is activated by mORAL, by using our stroke-based approach, the error in estimating the total brushing duration will be reduced from 4.1% to $< 0.5\%$.

11 LIMITATIONS AND FUTURE WORK

The work presented here has several limitations that open up opportunities for future work. First, our video annotation did not disambiguate between occlusal and lingual surfaces. As users are known to spend more time brushing occlusal and less time on lingual surfaces [35], future work can improve on video annotation and model training to seperately estimate the time spent on these two kinds of surfaces.

Second, this work did not estimate the pressure being applied during brushing, which is also an important component of brushing efficiency. Future work can develop methods that can leverage the stroke detection and characterization approach presented here to estimate the pressure applied during brushing of different surfaces.

Third, this work analyzed a week worth of daily brushing data from 19 participants during their natural brushing sessions. This work found significant variability between episodes of the same participant, and even greater variability among participants. As a result, although it achieved very high accuracy of classification in 10-fold cross-validation, but found the accuracy drop for leave-one-episode-out training and drop even further for leave-one-subject-out training. As between-person generalizability is important for real-life adoption of machine learning models, future work can increase the number of episodes per person and the number of participants

to determine the level at which clusters emerge among both episodes and among users that exhibit sufficient similarity. These clusters can then be used to develop group specific models which can more accurately detect brushing surface for each brushing episode and for each person.

Fourth, the time synchronization method presented here missed 16 out of 122 episodes, due to very slow brushing pace of some participants. Future work can investigate better methods that can automatically synchronize video labels and sensor data for all episodes without manual intervention.

Fifth, the algorithms presented here for stroke detection and time synchronization found specific thresholds that was suitable for the current dataset. Future work can develop adaptive thresholds or other adaptive algorithms that can generalize to unseen datasets without retraining. Finally, this work only observed the natural daily brushing behavior of participants and did not attempt to teach them better brushing habits. Future work can leverage the mTeeth model to develop interventions that can help users self-reflect on their brushing habits, detect regularly missed surfaces, and present personalized behavioral nudges to help individuals optimize their oral self-care routines and proactively tackle teeth surfaces at-risk for plaque accumulation.

12 CONCLUSIONS

The orientation of a toothbrush changes noticeably when brushing different tooth surfaces, resulting in detectable changes if inertial sensors are embedded in or attached to the brush itself, as in smart or instrumented toothbrushes. However, inferring tooth surface coverage from wrist-worn sensors is much more challenging because the changes are very subtle as the general orientation of the hand does not change much when transitioning from one teeth surface to another. Give that most brushes are manual and lack sensors, we develop a model for leveraging sensor data from ubiquitous smartwatches to infer brushing coverage. This work presents several insights for detecting these subtle signatures and constructs a model to distinguish among teeth surfaces and transitions between them. By doing so, it opens up a new frontier in the detection of rare daily events such as brushing, flossing, eating, drinking, and smoking by allowing finer-grained characterization (i.e., detecting even more ephemeral embedded micro-events) of self-care activities in natural environments. This may motivate new methods for successful characterization of other rare events such as detecting smoking with e-cigarettes that only consists of one or two puffs at a time, classifying among different kinds of food or drink in an eating or drinking episode by distinguishing the subtle differences in the hand-to-mouth gesture involved, and similar other daily behaviors.

ACKNOWLEDGEMENTS

We thank the anonymous reviewers for improving the paper. Research reported here was supported by the National Institutes of Health (NIH) under award R01DE024244 by the National Institute of Dental and Craniofacial Research (NIDCR), and awards P41EB028242, R01CA224537, R01MD010362, R01CA190329, R24EB025845, U01CA229437 and U54EB020404. It was also supported by the National Science Foundation (NSF) under awards IIS-1722646, ACI-1640813, and CNS-1823221. The authors thank Shahin Samiei from the University of Memphis for study support and Hansjoerg Reick and Ingo Vetter from Oral-B/Procter & Gamble for the material support. The opinions expressed in this article are the authors' own and do not reflect the views of the NIH, NSF, or Oral-B.

REFERENCES

[1] Accessed August, 2020. *Bass tooth brushing technique.* https://tube.medchrome.com/2013/04/brushing-technique-bass-and-modified.html
[2] Accessed August, 2020. *ELAN software.* https://archive.mpi.nl/tla/elan/
[3] Accessed August, 2020. *Oral B Genius brush.* https://oralb.com/en-us/products/electric-toothbrushes/genius-8000-rechargeable-electric-toothbrush/
[4] Accessed August, 2020. *Sonic-powered automatic toothbrush.* https://bfbe9e.kckb.st
[5] Sayma Akther, Nazir Saleheen, Shahin Alan Samiei, Vivek Shetty, Emre Ertin, and Santosh Kumar. 2019. 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* 3, 1 (2019), 1–25.

[6] Miguel Bruns Alonso, Jelle Stienstra, and Rob Dijkstra. 2014. Brush and learn: transforming tooth brushing behavior through interactive materiality, a design exploration. In *Proceedings of the 8th International Conference on Tangible, Embedded and Embodied Interaction*. ACM, 113–120.

[7] Thomas Attin and E Hornecker. 2005. Tooth brushing and oral health: how frequently and when should tooth brushing be performed? *Oral Health & Preventive Dentistry* 3, 3 (2005).

[8] Barbara Bruno, Fulvio Mastrogiovanni, Antonio Sgorbissa, Tullio Vernazza, and Renato Zaccaria. 2012. Human motion modelling and recognition: A computational approach. In *2012 IEEE International Conference on Automation Science and Engineering (CASE 2012)*. IEEE, 156–161.

[9] Yu-Chen Chang, Jin-Ling Lo, Chao-Ju Huang, Nan-Yi Hsu, Hao-Hua Chu, Hsin-Yen Wang, Pei-Yu Chi, and Ya-Lin Hsieh. 2008. Playful toothbrush: ubicomp technology for teaching tooth brushing to kindergarten children. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 363–372.

[10] Renate Deinzer, Stefanie Ebel, Helen Blättermann, Ulrike Weik, and Jutta Margraf-Stiksrud. 2018. Toothbrushing: to the best of one's abilities is possibly not good enough. *BMC Oral Health* 18, 1 (2018), 1–7.

[11] Christian Graetz, Jule Bielfeldt, Lars Wolff, Claudia Springer, Karim M Fawzy El-Sayed, Sonja Sälzer, Sabah Badri-Höher, and Christof E Dörfer. 2013. Toothbrushing education via a smart software visualization system. *Journal of Periodontology* 84, 2 (2013), 186–195.

[12] Taku Hachisu and Hiroyuki Kajimoto. 2015. Modulating tooth brushing sounds to affect user impressions. *International Journal of Arts and Technology* 8, 4 (2015), 307–324.

[13] Takashi Hamatani, Moustafa Elhamshary, Akira Uchiyama, and Teruo Higashino. 2018. FluidMeter: Gauging the human daily fluid intake using smartwatches. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 3 (2018), 1–25.

[14] Peng-Ju Hsieh, Yen-Liang Lin, Yu-Hsiu Chen, and Winston Hsu. 2016. Egocentric activity recognition by leveraging multiple mid-level representations. In *Multimedia and Expo (ICME), International Conference on*. IEEE, 1–6.

[15] Hua Huang and Shan Lin. 2016. Toothbrushing monitoring using wrist watch. In *Proceedings of the 14th ACM Conference on Embedded Network Sensor Systems*. ACM, 202–215.

[16] Kee-Deog Kim, Jin-Sun Jeong, Hae Na Lee, Yu Gu, Kyeong-Seop Kim, Jeong-Whan Lee, and Wonse Park. 2015. Efficacy of computer-assisted, 3D motion-capture toothbrushing instruction. *Clinical Oral Investigations* 19, 6 (2015), 1389–1394.

[17] Kyeong-Seop Kim, Tae-Ho Yoon, Jeong-Whan Lee, and Dong-Jun Kim. 2009. Interactive toothbrushing education by a smart toothbrush system via 3D visualization. *Computer Methods and Programs in Biomedicine* 96, 2 (2009), 125–132.

[18] Joseph Korpela, Ryosuke Miyaji, Takuya Maekawa, Kazunori Nozaki, and Hiroo Tamagawa. 2015. Evaluating tooth brushing performance with smartphone sound data. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 109–120.

[19] Joseph Korpela, Ryosuke Miyaji, Takuya Maekawa, Kazunori Nozaki, and Hiroo Tamagawa. 2016. Toothbrushing performance evaluation using smartphone audio based on hybrid HMM-recognition/SVM-regression model. *Journal of Information Processing* 24, 2 (2016), 302–313.

[20] Kang-Hwi Lee, Jeong-Whan Lee, Kyeong-Seop Kim, Dong-Jun Kim, Kyungho Kim, Heui-Kyung Yang, Keesam Jeong, and Byungchae Lee. 2007. 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, 4211–4214.

[21] Nahyeon Lee, Doyoung Jang, Yeji Kim, Byung-Cull Bae, and Jun-Dong Cho. 2016. Denteach: A device for fostering children's good tooth-brushing habits. In *Proceedings of the The 15th International Conference on Interaction Design and Children*. ACM, 619–624.

[22] Young-Jae Lee, Pil-Jae Lee, Kyeong-Seop Kim, Wonse Park, Kee-Deog Kim, Dosik Hwang, and Jeong-Whan Lee. 2011. Toothbrushing region detection using three-axis accelerometer and magnetic sensor. *IEEE Transactions on Biomedical Engineering* 59, 3 (2011), 872–881.

[23] Young-Jae Lee, Pil-Jae Lee, Jeong-Whan Lee, Kyeong-Seop Kim, Jin-Sun Jeong, Wonse Park, and Kee-Deog Kim. 2011. Quantitative assessment of toothbrushing education efficacy using smart toothbrush. In *2011 4th International Conference on Biomedical Engineering and Informatics (BMEI)*, Vol. 2. IEEE, 1160–1164.

[24] Hong Li, Shishir Chawla, Richard Li, Sumeet Jain, Gregory D Abowd, Thad Starner, Cheng Zhang, and Thomas Plötz. 2018. Wristwash: towards automatic handwashing assessment using a wrist-worn device. In *Proceedings of the 2018 ACM International Symposium on Wearable Computers*. ACM, 132–139.

[25] Yuan Liang, Hsuan Wei Fan, Zhujun Fang, Leiying Miao, Wen Li, Xuan Zhang, Weibin Sun, Kun Wang, Lei He, and Xiang'Anthony' Chen. 2020. OralCam: Enabling Self-Examination and Awareness of Oral Health Using a Smartphone Camera. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[26] Harald Löe. 2000. Oral hygiene in the prevention of caries and periodontal disease. *International Dental Journal* 50, 3 (2000), 129–139.

[27] Chengwen Luo, Xingyu Feng, Junliang Chen, Jianqiang Li, Weitao Xu, Wei Li, Li Zhang, Zahir Tari, and Albert Y Zomaya. 2019. Brush like a Dentist: Accurate Monitoring of Toothbrushing via Wrist-Worn Gesture Sensing. In *IEEE INFOCOM 2019*. IEEE, 1234–1242.

[28] Marco Marcon, Augusto Sarti, and Stefano Tubaro. 2016. Toothbrush motion analysis to help children learn proper tooth brushing. *Computer Vision and Image Understanding* 148 (2016), 34–45.

[29] Tatsuo Nakajima, Vili Lehdonvirta, Eiji Tokunaga, and Hiroaki Kimura. 2008. Reflecting human behavior to motivate desirable lifestyle. In *Proceedings of the 7th ACM conference on Designing interactive systems*. ACM, 405–414.

[30] Zhenchao Ouyang, Jingfeng Hu, Jianwei Niu, and Zhiping Qi. 2017. An asymmetrical acoustic field detection system for daily tooth brushing monitoring. In *IEEE Global Communications Conference*. IEEE, 1–6.

[31] Abhinav Parate, Meng-Chieh Chiu, Chaniel Chadowitz, Deepak Ganesan, and Evangelos Kalogerakis. 2014. Risq: Recognizing smoking gestures with inertial sensors on a wristband. In *Proceedings of the 12th Annual International Conference on Mobile Systems, Applications, and Services*. 149–161.

[32] Nazir Saleheen, Amin Ahsan Ali, Syed Monowar Hossain, Hillol Sarker, Soujanya Chatterjee, Benjamin Marlin, Emre Ertin, Mustafa Al'Absi, and Santosh Kumar. 2015. 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*. ACM, 999–1010.

[33] Edison Thomaz, Irfan Essa, and Gregory D Abowd. 2015. 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*. ACM, 1029–1040.

[34] Lahiru NS Wijayasingha and Benny Lo. 2016. A wearable sensing framework for improving personal and oral hygiene for people with developmental disabilities. In *2016 IEEE Wireless Health (WH)*. IEEE, 1–7.

[35] Tobias Winterfeld, N Schlueter, Daniela Harnacke, Jörg Illig, Jutta Margraf-Stiksrud, Renate Deinzer, and Carolina Ganss. 2015. Toothbrushing and flossing behaviour in young adults—a video observation. *Clinical Oral Investigations* 19, 4 (2015), 851–858.

[36] Takuma Yoshitani, Masa Ogata, and Koji Yatani. 2016. LumiO: a plaque-aware toothbrush. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 605–615.