Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

OREBA: A Dataset for Objectively Recognizing Eating Behaviour and Associated Intake

PHILIPP V. ROUAST[1], (Student Member, IEEE), HAMID HEYDARIAN[1], MARC T. P. ADAM[1,3], and MEGAN E. ROLLO[2,3]

[1]School of Electrical Engineering and Computing, The University of Newcastle, Callaghan, NSW 2308, Australia
[2]School of Health Sciences, The University of Newcastle, Callaghan, NSW 2308, Australia
[3]Priority Research Centre for Physical Activity and Nutrition, The University of Newcastle, Callaghan, NSW 2308, Australia

Corresponding author: Marc T. P. Adam (e-mail: marc.adam@newcastle.edu.au).

We gratefully acknowledge the support by the Bill & Melinda Gates Foundation [OPP1171389]. Philipp Rouast and Hamid Heydarian were supported by an Australian Government Research Training (RTP) Scholarship.

ABSTRACT Automatic detection of intake gestures is a key element of automatic dietary monitoring. Several types of sensors, including inertial measurement units (IMU) and video cameras, have been used for this purpose. The common machine learning approaches make use of the labelled sensor data to automatically learn how to make detections. One characteristic, especially for deep learning models, is the need for large datasets. To meet this need, we collected the Objectively Recognizing Eating Behavior and Associated Intake (OREBA) dataset. The OREBA dataset aims to provide a comprehensive multi-sensor recording of communal intake occasions for researchers interested in automatic detection of intake gestures. Two scenarios are included, with 100 participants for a discrete dish and 102 participants for a shared dish, totalling 9069 intake gestures. Available sensor data consists of synchronized frontal video and IMU with accelerometer and gyroscope for both hands. We report the details of data collection and annotation, as well as technical details of sensor processing. The results of studies on IMU and video data involving deep learning models are reported to provide a baseline for future research.

INDEX TERMS Dietary monitoring, eating behaviour assessment, accelerometer, communal eating, gyroscope, 360-degree video camera

I. INTRODUCTION

TRADITIONAL dietary assessment methods are reliant on self-report data. While data captured with active methods such as self-report and 24-hr recall are widely used in practice, they are not without limitations (e.g., human error, time-consuming manual process) [1]. Automatic dietary monitoring, where data is collected and processed independent of the individual, has the potential to complement data from traditional methods and reduce associated biases [2]. In addition, such systems have the potential to support personal self-monitoring solutions by providing individuals with targeted eating behaviour recommendations.

A key element of automatic dietary monitoring is the detection of *intake gestures* (i.e., the process of moving food or drink towards the mouth). Recent research on this task focuses mainly on machine learning approaches which are characterized by a need for large amounts of labeled data. This is especially true in conjunction with deep learning, which has been applied in this context since 2017 [3]. However, collecting, synchronizing, and labeling data of eating occasions is a work-intensive process. Hence, there is a need for more public datasets to reduce barriers for researchers to create new machine learning models, and to objectively compare the performance of existing approaches [4], [5].

At the same time, current research on dietary monitoring identified a gap in research on shared plate eating [6]. Communal eating (i.e., eating occasions involving more than one person) is not yet well understood, let alone the impact it has on accuracy of automatic dietary monitoring. Hence, existing research on capturing dietary intake from discrete dishes needs to be complemented and contrasted with research on the detection of intake from shared dishes.

In order to address these gaps, the present paper introduces the Objectively Recognizing Eating Behavior and Associated Intake (OREBA) dataset. OREBA was developed as part of a research project investigating eating behavior and associated

TABLE 1. Public datasets of intake gestures with synchronized sensor data and annotations available.

Dataset	Synchronized sensors: Video	Audio	IMU	Scale	Ground truth	Participants (Recordings)[a]	Intake events	Annotations
ACE [7]	-	Earbud	Both hands 15 Hz	-	Video multi[b]	7 (13)	1492	Chews, swallows with type and amount of food and drink.
Clemson [8]	-	-	Dominant hand 15 Hz	Tray 15 Hz	Video ceiling[b]	264 (488)	20644	Intake gestures, utensiling. Intake annotated with hand, utensil, container, and food.
FIC [9]	-	-	Dominant hand 100 Hz	-	Video frontal[b]	12 (21)	1332	Plain intake gestures; further annotation of micromovements.
OREBA-DIS	Frontal 24 fps	-	Both hands 64 Hz	-	Video frontal	100 (100)	4790	Intake gestures annotated with eat/drink, hand and utensil.
OREBA-SHA	Frontal 30 fps	-	Both hands 64 Hz	Communal 1 Hz	Video multi	102 (102)	4279	Intake gestures annotated with eat/drink, hand and utensil.

[a] One recordings equals one person consuming one meal.
[b] Used for ground truth annotation, but not available for download as of early 2020.

intake in communal eating across two scenarios (discrete dish and shared dish). A total of 202 meal recordings were completed, with 180 unique individuals participating. The total number of intake gestures across both recordings is 9069. A single spherical video camera positioned in the center of the table made it possible to capture the entire scene of up to four participants. The video features the frontal perspective of each participant, offering a full view of all relevant gestures. The dataset includes synchronized frontal video and inertial measurement unit (IMU) of both hands, along with labels for each intake gesture. To our best knowledge, this is the first dataset to make such data publicly available for communal eating from both discrete and shared dishes.

In the following, Section II gives an overview of the related work and existing datasets, Section III introduces the data collection and annotation process of the OREBA dataset in detail, Section IV provides results from our initial studies as baselines, and we conclude with a discussion in Section V.

II. RELATED WORK

A. AUTOMATIC DIETARY MONITORING

Automatic dietary monitoring encompasses three major goals: (i) detecting the timing of intake events, (ii) recognizing the type of food or drink, and (iii) estimating the weight consumed. Detection of intake behavior, which is associated with intake gestures, chews, and swallows, can be considered as part of the first goal. Researchers have leveraged various different sensor types for this purpose. While chews and swallows can be detected using audio signals [10], intake gestures are typically handled using an IMU including accelerometer and gyroscope sensors [11]. Before the application of deep learning architectures, the traditional approach in this field reduced the dimensionality of the raw sensor data by extracting features handcrafted using expert knowledge. Deep learning methods have been explored to detect individual intake gestures with inertial sensor data since 2017 [3], [12] and with video data since 2018 [2], [4], [5], whereby large amounts of labelled examples are leveraged to let algorithms learn the features automatically.

B. EXISTING DATASETS

To date, most published studies on recognition of intake behaviour rely on dedicated, private datasets collected for a specific purpose. Considering the shift towards adoption of deep learning techniques, we expect an increasing need for large, public datasets that existing and emerging machine learning approaches can objectively be benchmarked on. Similar developments can be observed across several related fields such as action recognition [13], affect recognition [14], and object recognition [15], [16].

Table 1 provides an overview of publicly available datasets on intake gestures which feature synchronized sensor data of eating occasions with labels for individual gestures (intake gestures or other eating related gestures)[1].

- The *accelerometer and audio-based calorie estimation (ACE) dataset*[2] [7] contains seven participants with audio and IMU data for both hands and the head. Annotations of type and amount of food and drink are available for chews and swallows.
- The *Clemson Cafeteria dataset*[3] [8] contains 264 participants and 488 recordings. IMU data is available at 15 Hz for the dominant hand, along with scale measurements for the tray. Each intake gesture is annotated with hand, utensil, container, and food.
- The *Food Intake Cycle (FIC) dataset*[4] [9], which consists of 12 participants and 21 recordings, includes IMU data for the dominant hand. The focus is on the micromovements during intake gestures.

While video is commonly used as ground truth, none of the existing datasets currently include video data as part of the synchronized sensor data for analysis. In terms of IMU data and quantity of recorded intake events, we find that the existing datasets are restricted either to data from only one hand, a relatively low recording frequency (15 Hz), or few participants. We aim to further the field by establishing the

[1] A related dataset is iHEARu-EAT [17], however we did not include it here since it does not focus on intake events.
[2] See http://www.skleinberg.org/data.html
[3] See http://cecas.clemson.edu/~ahoover/cafeteria/. Recordings with missing annotations are excluded here.
[4] See https://mug.ee.auth.gr/intake-cycle-detection/

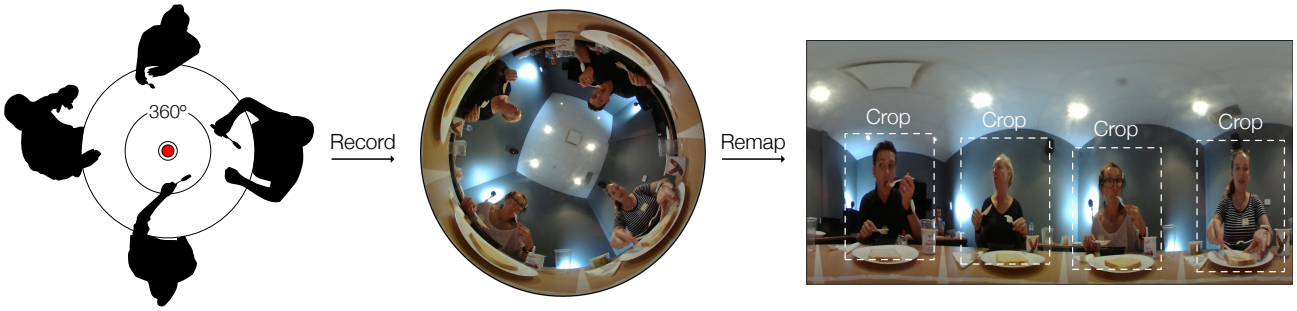

FIGURE 1. The spherical video is remapped to equirectangular representation, cropped, and reshaped to square shape.

FIGURE 2. Study setup for OREBA-SHA. One camera in the center of the table, IMU on each wrist, and four scales.

OREBA dataset, which includes video and IMU from both hands, at a quantity of intake events sufficient to train deep learning models for both video and inertial modalities.

III. THE OREBA DATASET

The OREBA dataset aims to provide a comprehensive multi-sensor recording of communal intake occasions for researchers interested in automatic detection of intake gestures and other behaviours associated with intake (e.g., serving food onto a plate). Available sensor data consists of synchronized frontal video and accelerometer and gyroscope for both hands in two different scenarios (i.e., discrete dish and shared dish). IRB approval was given (H-2017- 0208), and the data was recorded between Mar 2018 and Oct 2019.

A. SCENARIOS

The OREBA dataset consists of two separate communal eating scenarios. In each scenario, groups of up to four participants were simultaneously recorded consuming a meal at a communal table:

1) OREBA-DIS: In the first scenario, foods were served in discrete portions to each participant. The meal consisted of lasagna (choice between vegetarian and beef), bread, and yogurt. Additionally, there was water available to drink, and butter to spread on the bread. The study setup for OREBA-DIS is shown in Fig. 1.
2) OREBA-SHA: In the second scenario, participants consumed a communal dish of vegetable korma or butter chicken with rice and mixed vegetables. Additionally, there was water available to drink. The study setup for OREBA-SHA is shown in Fig. 2.

Lasagna and rice-based dishes were chosen since they are amongst the most common dishes in similar studies [11]. All participants in each scenario are unique, however 22 participants participated in both scenarios.

B. SENSORS

For each group, video was recorded using a spherical camera placed in the center of the shared table (360fly-4K[5]). This allowed video recording to occur in a simultaneous and unobtrusive way for all participants engaging in the communal eating occasion around the table. The sampling rates are 24 fps for OREBA-DIS, and 30 fps for OREBA-SHA. Each participant wore two IMU units, one on each wrist (Movisens Move 3+ [6]). The IMU included an accelerometer and a gyroscope with a sampling rate of 64 Hz. For OREBA-SHA, four scales additionally recorded the weight of the communal dishes (two rices dishes, one wet dish, one vegetable dish) at 1 Hz (Adam Equipment CBK 4).

C. SENSOR PROCESSING

1) Video

As shown in Fig. 1, we first mapped the spherical video from the 360-degree camera to equirectangular representation[7]. Then, we separated the equirectangular representation into individual participant videos by cropping the areas of interest. We further resized each participant video to a square shape. The two spatial resolutions 140x140 (e.g., `<id>_video_140p.mp4`) and 250x250 pixels (e.g., `<id>_video_250p.mp4`) are included. All videos are encoded using the H.264 standard and stored in mp4 containers.

2) Inertial Measurement Unit

Raw accelerometer data is measured in g, while gyroscope data is measured in deg/s. The OREBA dataset includes (i)

[5]See https://www.360fly.com/
[6]See https://www.movisens.com/en/products/activity-sensor-move-3/
[7]See https://github.com/prouast/equirectangular-remap

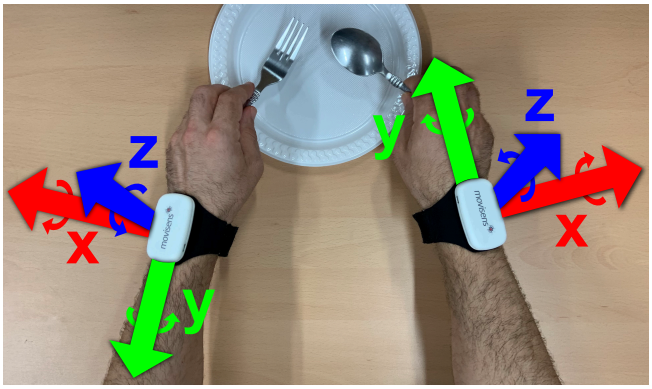

FIGURE 3. The wrist-worn sensors with their internal coordinate frames. Additionally, the direction of positive rotations are indicated for each axis.

raw sensor data without any processing for left and right hand (e.g., `<id>_inertial_raw.csv`), and (ii) processed sensor data for dominant and non-dominant eating hand (e.g., `<id>_inertial_processed.csv`). Processed data is generated from the raw data according to the following steps:

1: Removal of gravity effect. The raw accelerometer reading is subject to acceleration from participants' wrist movements as well as the earth's gravitational field. We remove this gravity effect by estimating sensor orientation using sensor fusion with Madgwick's filter [18], rotation of the acceleration vector with the resulting quaternion, and deduction of the gravity vector (see [3] for a similar approach).

2: Standardization. Each column (meaning each axis for each modality and hand) is standardized by subtracting its mean and dividing by its standard deviation (see [9] for a similar approach). Processed data can hence be regarded as unitless.

3: Transforming from left and right hand to dominant and non-dominant hand. To achieve data uniformity, we report hands in the processed data as *dominant* and *non-dominant*. A similar approach was chosen for the FIC dataset [12]. All data reported as dominant hands correspond to right hands, and non-dominant hands to left hands; for left-handed participants data for both hands has been transformed to achieve this. We can simulate a recording on the opposite wrist (mirrored horizontally, e.g., transforming data from left wrist as if it had been recorded on right wrist). Due to the way the sensors are mounted on the wrist (see Fig. 3), the horizontal direction corresponds to the x axis. For accelerometer data, we estimate mirroring by flipping the sign of the x axis, and for gyroscope by flipping the signs of the y and z axis. Further, we also flip the signs of the x and y axis to compensate for the different sensor orientations on the wrists, yielding transformation (1) for accelerometer and (2) for gyroscope.

$$[a'_x, a'_y, a'_z] = [-(-a_x), -a_y, a_z] = [a_x, -a_y, a_z] \quad (1)$$

$$[g'_x, g'_y, g'_z] = [-g_x, -(-g_y), -g_z] = [-g_x, g_y, -g_z] \quad (2)$$

TABLE 2. The labeling scheme.

Category	Possible values
Main	Intake, Serve
Sub	Intake-Eat, Intake-Drink, Serve-Self, Serve-Other
Hand	Left, Right, Both
Utensil	Fork, Spoon, Hand, Knife, Finger, Cup, Bottle

Note that the mirroring technique propopsed here could also be of use for data augmentation pipelines.

3) Synchronization

Ground truth for sensor synchronization was acquired by asking participants to clap their hands before starting, and after finishing their meal (see [3] for a similar approach). The clapping creates a distinct signature in both the video recording and the accelerometer. All sensors were trimmed in time and synchronized for each participant based on these two reference points.

4) Scales

The shared plate setting in OREBA-SHA included four scales that measured the weight of the two rice dishes at two corners of the table as well as the wet dish and the vegetable dish in the centre of the table (see Figure 2). These scales recorded the weight of the four dishes in grams at a sampling rate of 1 Hz. The scale recordings were time-synchronized by means of a time-lapse camera and a 200g calibration weight. At the start of the recording, a research assistant removed the calibration weight from the scale. This was captured by the scale recordings as well as the time-lapse camera. Further, the time-lapse camera also captured the clapping at the start of the recording. Based on this, each scale recording includes a reference in seconds to the clapping at the start of a recording. Further, the dataset provides a mapping of each participant number to the closest rice dish.

D. ANNOTATION

All relevant gestures were labelled and cross-checked by two independent annotators using ChronoViz [8]. Each gesture is defined by a start and an end timestamp:

- The start timestamp is the point where the final uninterrupted movement to execute the gesture starts;
- the end timestamp is the point when the participant has finished returning their hand(s) from the movement or started a different gesture.

Additionally, each gesture is assigned four labels according to our labelling scheme as listed in Table 2. Besides the *Main* identification as an Intake or Serve gesture, this scheme allows to further specify a *Sub* category for each gesture, as well as indicating the *Hand* and *Utensil* (e.g., `<id>_annotations.csv`).

The scheme is designed to be extendable with more categories in possible extensions of the dataset. The discrete dish

[8] See http://chronoviz.com

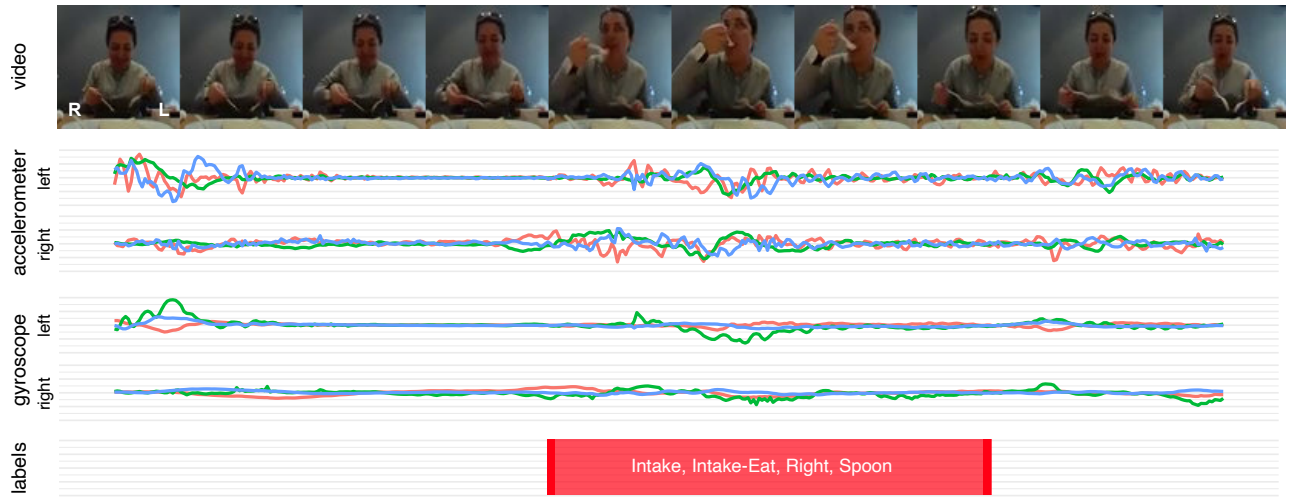

FIGURE 4. Example of a labeled intake gesture with video, accelerometer, and gyroscope sensor data. For easier display, the video framerate has been reduced.

scenario OREBA-DIS includes Intake labels. Correspondingly, the shared dish scenario OREBA-SHA includes both Intake and Serve labels. Figure 4 depicts an example of a labeled intake gesture and associated IMU sensor data.

E. SPLITS

For machine learning problems with time-intensive training and evaluation, the best practice is to train, validate, and test using three separate sets of data [19]. Models are trained with the training set, hyperparameters are tuned using the validation set, and reported results are based on the test set. We choose a split of approximately 3:1:1, such that each participant only appears in one of the three subsets; this is to ensure that we are measuring the model's ability to generalise and avoid data leakage. The recommended split is included in the dataset download. Table 3 summarises high-level statistics on these splits.

F. DEMOGRAPHICS

Out of 180 participants in total, 161 agreed to complete a demographics questionnaire. Across the dataset, 67% identified as male and 33% as female. The median age is 24, with the minimum and maximum age being 18 and 54 years respectively. Reported ethnicities in the dataset include White Australian (52.2%), White other European (9.9%), Chinese (8.7%), Other Asian (8.7%), Persian (5.6%), Arabic (3.1%), White British (3.1%), African (2.5%), and South East Asian (1.8%). About 10% reported being left-, and 90% right-handed.

G. AVAILABILITY

The OREBA dataset is available on request to research groups at academic institutions. Please visit http://www.newcastle.edu.au/oreba to download the data sharing agreement and get access.

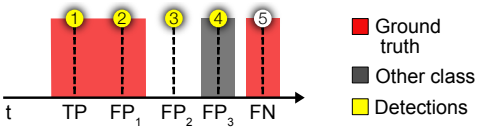

FIGURE 5. The evaluation scheme (proposed by [9]; figure from [21] extended here). (1) A true positive is the first detection within each ground truth event; (2) False positives of type 1 are further detections within the same ground truth event; (3) False positives of type 2 are detections outside ground truth events; (4) False positives of type 3 are detections made for the wrong class; (5) False negatives are non-detected ground truth events.

IV. BASELINE FOR INTAKE GESTURE DETECTION

Intake gesture detection refers to the task of detecting the times of individual intake gestures from sensor data. Similar to dataset papers in other areas [15] [20], we provide baseline results for this task on OREBA-DIS and OREBA-SHA. As baseline models, we use the ResNet-50 SlowFast model from Rouast et al. [21] for video data and the CNN-LSTM model from Heydarian et al. [22] for inertial data. To ensure comparability with future studies, we use the publicly available data splits introduced in Section III-E for training, validation, and test; we additionally report details on the used evaluation metric. Table 4 reports the test set results for the aforementioned models.

A. EVALUATION SCHEME

We extend the evaluation scheme proposed by Kyritsis et al. [9] as depicted in Fig. 5. The scheme uses the ground truth to translate sparse detections into measurable metrics for a given label category. As Rouast and Adam [21] report, one correct detection per ground truth event counts as a true positive (TP), while further detections within the same ground truth event are false positives of type 1 (FP_1). Detections outside ground truth events are false positives of type 2 (FP_2) and non-detected ground truth events count as false negatives (FN). The scheme has been extended here to support the multi-class case, where detections for a wrong

TABLE 3. Summary statistics for our dataset and the training/validation/test split.

Scenario	Type	Training			Validation			Test			Total		
		#	Mean [s]	Std [s]	#	Mean [s]	Std [s]	#	Mean [s]	Std [s]	#	Mean [s]	Std [s]
OREBA-DIS	Participants	61	804.98	238.64	20	793.14	254.36	19	875.77	217.40	100	816.07	237.47
	Intake Gest.	2907	2.36	1.04	943	2.24	0.98	940	2.29	1.01	4790	2.32	1.02
OREBA-SHA	Participants	63	838.55	259.76	20	811.31	221.04	19	824.60	201.68	102	830.61	240.80
	Intake Gest.	2574	2.44	1.16	896	2.23	1.17	809	2.24	1.02	4279	2.36	1.14
	Serve Gest.	337	9.87	4.52	107	10.99	6.56	112	9.05	4.34	556	9.92	4.97

TABLE 4. Baseline test set results for intake gesture detection. On OREBA-DIS, the video model performs better than the inertial model, while the opposite is true on OREBA-SHA. This indicates that the test set for OREBA-DIS is more challenging for using inertial data, while the test set for OREBA-SHA is more challenging for using video data.

Dataset	Modality	Precision	Recall	F_1
OREBA-DIS	Video	**0.912**	**0.801**	**0.853**
	Inertial	0.764	0.793	0.778
OREBA-SHA	Video	0.782	0.834	0.808
	Inertial	**0.800**	**0.912**	**0.852**

class are false positives of type 3. Based on the aggregate counts, precision ($\frac{TP}{TP+FP_1+FP_2+FP_3}$), recall ($\frac{TP}{TP+FN}$), and the F_1 score ($2 * \frac{Precision*Recall}{Precision+Recall}$) can be calculated.

B. INERTIAL SENSORS

Heydarian et al. [22] applied the two-stage approach proposed by Kyritsis et al. [9] to the inertial data in OREBA-DIS for detection of intake gestures. They ran multiple experiments benchmarking different deep learning models and pre-processing pipelines. The top model performance was achieved by a CNN-LSTM with earliest fusion through a dedicated CNN layer and target matching. Concerning preprocessing, their results show that applying a consecutive combination of mirroring, removing the gravity effect, and standardization was beneficial for model performance, while smoothing and downsampling had adverse effects.

From the results in Table 4, it appears that the test set for OREBA-DIS ($F_1 = 0.778$) is more challenging for inertial data than the test set for OREBA-SHA ($F_1 = 0.852$).

C. VIDEO

Rouast and Adam [21] applied several deep learning architectures established in the literature on video action recognition on the task of detecting intake gestures directly from the video data in OREBA-DIS. The best test set result was achieved using a SlowFast [23] network with ResNet-50 [24] as backbone. Further conclusions from the experiments are that appearance features are more useful than motion features, and that temporal context in form of multiple video frames is essential for top model performance.

The results in Table 4 indicate that the test set for OREBA-SHA ($F_1 = 0.808$) is more challenging when working with video data than the test set for OREBA-DIS ($F_1 = 0.853$).

V. DISCUSSION

In this paper, we have introduced the OREBA dataset, which provides a comprehensive multi-sensor recording with labeled gestures of communal intake occasions from discrete and shared meals. Building on a summary of related work on automatic dietary monitoring and an overview of existing public datasets in the field, we provided details on the data collection, sensor processing, and annotation methods employed in the creation of the OREBA dataset. Additionally, we reported baseline results on the task of intake gesture detection based on video and inertial sensor data.

Sensor-based, passive methods of dietary monitoring have the potential of complementing existing active methods such as food records and 24-hr recall. As seen in other fields such as object recognition [15] and action recognition [13], progress in the research of machine learning methods is tightly linked to the availability and ongoing development of datasets with labeled examples. In this light, we hope that the OREBA dataset will be able to support future developments in automatic dietary monitoring. Compared to existing public datasets of labeled intake gestures, the OREBA dataset is unique as it provides (i) synchronised frontal video data based on spherical video recordings, (ii) inertial data from both hands at 64 Hz, and (iii) a total amount of 202 recordings in two different communal eating scenarios.

While we have reported results from our initial studies on detecting intake gestures with either video or inertial sensor data, there are also several other directions that research on this dataset could go into. Sensor fusion of video and IMU for detecting intake gestures could be explored to combine the strengths of both approaches. Improvements could also be made by using transfer learning between the two different scenarios, allowing to contrast the difficulties of monitoring discrete versus shared dishes. Thanks to the availability of inertial data for both hands, studies could also explore how much information is retained on the dominant versus the non-dominant hand, which has implications for automatic dietary monitoring using commercial smartwatches. Further, while our initial studies focused on detecting individual intake gestures, future studies could explore the other label categories – for example, how well the video or inertial modalities perform at distinguishing between different utensils. Finally, sensor fusion of video and IMU is a further possibility that could be explored in the future. As such, and beyond researchers specifically interested in dietary monitoring, the OREBA dataset could also be a valuable resource for researchers interested in advancing machine learning models

for sensor fusion more broadly.

ACKNOWLEDGMENTS

We acknowledge the untiring support of Clare Cummings, Grace Manning, Alice Melton, Kaylee Slater, Felicity Steel and Sam Stewart in collecting and annotating the data.

REFERENCES

[1] S. W. Lichtman, K. Pisarska, E. R. Berman, M. Pestone, H. Dowling, E. Offenbacher, H. Weisel, S. Heshka, D. E. Matthews, and S. B. Heymsfield, "Discrepancy between self-reported and actual caloric intake and exercise in obese subjects," New England J. Medicine, vol. 327, no. 27, pp. 1893–1898, 1992.

[2] P. V. Rouast, M. T. P. Adam, T. Burrows, R. Chiong, and M. E. Rollo, "Using deep learning and 360 video to detect eating behavior for user assistance systems," in Proc. Europ. Conf. Information Systems, 2018.

[3] K. Kyritsis, C. Diou, and A. Delopoulos, "Food intake detection from inertial sensors using lstm networks," in Proc. Int. Conf. Image Analysis and Processing, 2017, pp. 411–418.

[4] J. Qiu, F. P.-W. Lo, and B. Lo, "Assessing individual dietary intake in food sharing scenarios with a 360 camera and deep learning," in Proc. IEEE International Conference on Wearable and Implantable Body Sensor Networks, 2019.

[5] D. Konstantinidis, K. Dimitropoulos, B. Langlet, P. Daras, and I. Ioakimidis, "Validation of a deep learning system for the full automation of bite and meal duration analysis of experimental meal videos," Nutrients, vol. 12, no. 1, p. 209, 2020.

[6] T. Burrows, C. Collins, M. T. P. Adam, K. Duncanson, and M. Rollo, "Dietary assessment of shared plate eating: A missing link," Nutrients, vol. 11, no. 4, pp. 1–14, 2019.

[7] C. Merck, C. Maher, M. Mirtchouk, M. Zheng, Y. Huang, and S. Kleinberg, "Multimodality sensing for eating recognition," in Proc. Int. Conf. Pervasive Computing Technologies for Healthcare, 2016, pp. 130–137.

[8] Y. Shen, J. Salley, E. Muth, and A. Hoover, "Assessing the accuracy of a wrist motion tracking method for counting bites across demographic and food variables," IEEE J. Biomedical and Health Informatics, vol. 21, no. 3, pp. 599–606, 2017.

[9] K. Kyritsis, C. Diou, and A. Delopoulos, "Modeling wrist micromovements to measure in-meal eating behavior from inertial sensor data," IEEE J. Biomedical and Health Informatics, 2019.

[10] O. Amft, M. Stager, P. Lukowicz, and G. Troster, "Analysis of chewing sounds for dietary monitoring," in Proc. UbiComp, 2005, pp. 56–72.

[11] H. Heydarian, M. Adam, T. Burrows, C. Collins, and M. E. Rollo, "Assessing eating behaviour using upper limb mounted motion sensors: A systematic review," Nutrients, vol. 11, no. 5, p. 1168, 2019.

[12] K. Kyritsis, C. Diou, and A. Delopoulos, "A data driven end-to-end approach for in-the-wild monitoring of eating behavior using smartwatches," IEEE J. Biomedical and Health Informatics, 2020.

[13] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman, "The kinetics human action video dataset," arXiv preprint arXiv:1705.06950, 2017.

[14] P. V. Rouast, M. T. P. Adam, and R. Chiong, "Deep learning for human affect recognition: Insights and new developments," IEEE Trans. Affect. Comput., 2019.

[15] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and F.-F. Li, "Imagenet: A large-scale hierarchical image database," in Proc. CVPR, 2009, pp. 248–255.

[16] A. E. Ghahnavieh, M. Enayati, and A. A. Raie, "Introducing a large dataset of persian license plate characters," Journal of Electronic Imaging, vol. 23, no. 2, p. 023015, 2014.

[17] S. Hantke, F. Weninger, R. Kurle, F. Ringeval, A. Batliner, A. E.-D. Mousa, and B. Schuller, "I hear you eat and speak: Automatic recognition of eating condition and food type, use-cases, and impact on asr performance," PloS one, vol. 11, no. 5, p. e0154486, 2016.

[18] S. Madgwick, "An efficient orientation filter for inertial and inertial/magnetic sensor arrays," University of Bristol (UK), Tech. Rep., 2010.

[19] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. MIT Press, 2016.

[20] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild," IEEE Trans. Affect. Comput., 2017.

[21] P. Rouast and M. Adam, "Learning deep representations for video-based intake gesture detection," IEEE J. Biomedical and Health Informatics, vol. 24, no. 6, pp. 1727–1737, 2019.

[22] H. Heydarian, P. V. Rouast, M. T. P. Adam, T. Burrows, and M. E. Rollo, "Deep learning for intake gesture detection from wrist-worn inertial sensors: The effects of preprocessing, sensor modalities, and sensor positions," Working paper, 2020.

[23] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," arXiv preprint arXiv:1812.03982, 2018.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. CVPR, 2016, pp. 770–778.

PHILIPP V. ROUAST received the B.Sc. and M.Sc. degrees in Industrial Engineering from Karlsruhe Institute of Technology, Germany, in 2013 and 2016 respectively. He is currently working towards the Ph.D. degree in Information Systems and is a graduate research assistant at The University of Newcastle, Australia. His research interests include deep learning, affective computing, HCI, and related applications of computer vision. Find him at https://www.rouast.com.

HAMID HEYDARIAN received the B.Sc. in Computer Engineering (software) from Kharazmi University, Iran, in 2002. He is a senior software developer currently working towards the Ph.D. in Information Technology and is a casual academic at The University of Newcastle, Australia. His research interests include inertial signal processing using deep learning and its related applications in dietary intake assessment and passive dietary monitoring.

MARC T. P. ADAM is an Associate Professor in Computing and Information Technology at the University of Newcastle, Australia. In his research, he investigates the interplay of human users' cognition and affect in human-computer interaction. He is a founding member of the Society for NeuroIS. He received an undergraduate degree in Computer Science from the University of Applied Sciences Würzburg, Germany, and a PhD in Economics of Information Systems from Karlsruhe Institute of Technology, Germany.

MEGAN E. ROLLO is a Research Fellow in Nutrition and Dietetics within the School of Health Sciences and Priority Research Centre for Physical Activity and Nutrition at The University of Newcastle, Australia. She has received BAppSci, BHlthSci(Nutr&Diet), and PhD degrees from the Queensland University of Technology, Australia. She has research interests in technology-assisted dietary assessment and personalized behavioral nutrition interventions.