https://doi.org/10.1145/3686215.3686219

Automatic Recognition of Commensal Activities in *Co-located* and *Online* settings

Kheder Yazgi
DIPSCO
University of Trento
Rovereto, Italy
kheder.yazgi@unitn.it

Cigdem Beyan
Department of Computer Science
University of Verona
Verona, Italy
cigdem.beyan@univr.it

Maurizio Mancini
Department of Computer Science
Sapienza University of Rome
Rome, Italy
m.mancini@di.uniroma1.it

Radoslaw Niewiadomski
Department of Informatics, Bioengineering, Robotics and Systems Engineering
University of Genoa
Genoa, Italy
radoslaw.niewiadomski@unige.it

ABSTRACT

Technological advancement has profoundly impacted how people share meals, fostering research interest in new forms of commensality such as tele-dining and eating with artificial companions. Consequently, there is a need to develop computational methods for recognizing commensal activities, that is, actions related to food consumption and social signals displayed during meal-time. This paper introduces the first dataset that consists of synchronized video data of co-located dining dyads. The dataset is annotated with key social signals such as speaking activity, smiling, and food-related activities like chewing and food intake. Unlike previous studies that use remote settings, this work emphasizes the differences between online and co-located setups. A set of machine learning experiments is conducted on our and existing datasets, reaching the best F-score of 0.82. The cross-dataset analysis between co-located and online datasets also evidences the significant disparity between these two settings. While mixing co-located and online recordings may increase the model's generalizability, we notice strong differences between the two settings, highlighting the importance of in-person data recordings for accurate recognition.

CCS CONCEPTS

• **Computing methodologies** → *Computer vision*; • **Human-centered computing** → *Interaction techniques.*

KEYWORDS

Activity recognition, commensality, datasets, social interactions, co-located, in-person

ACM Reference Format:
Kheder Yazgi, Cigdem Beyan, Maurizio Mancini, and Radoslaw Niewiadomski. 2024. Automatic Recognition of Commensal Activities in *Co-located* and *Online* settings. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI Companion '24), November 4–8, 2024, San Jose, Costa Rica.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3686215.3686219

1 INTRODUCTION

Commensality refers to the act of eating and sharing food, typically in a social setting [21]. There has been growing interest in developing computational methods to study commensality, leading to the emergence of "computational commensality", which uses computational approaches to explore various aspects of food and eating behaviors [18]. Research in these fields has led to various applications spanning healthcare, well-being, and entertainment. These applications require models capable of accurately detecting human actions during mealtime. For instance, a chewing tracker can monitor chewing activities, supporting individuals coping with obesity [15]. Robot-assisted feeding systems assist people with physical or perceptual disabilities by closely monitoring the user's condition and delivering food from plate to mouth [2]. Additionally, Artificial Commensal Companions (ACCs), like the robot FoBo, facilitate social interactions during eating to mitigate the negative impacts of solitary dining on mental health [13, 16].

Studies such as [19], which focus on automatically recognizing commensal activities (defined as key actions such as food intake, chewing, and social signals like speaking and smiling), used the data collected during video calls. In contrast, other work [4, 8] concentrated on co-located settings, which allows for natural and spontaneous interaction without the delays and interruptions typical of internet-based commensality. However, these works [4, 8] are limited in scope, primarily focusing on healthcare and/or assistive purposes and often overlooking social aspects of eating. To the best of the authors' knowledge, an in-person dataset specifically focusing on the synchronized video data of activities, such as consuming food, mastication, and social signals like speaking and smiling, has not yet been developed. Notice that the synchronization of the recordings is an important feature that permits the study of the interaction dynamics and complexity of the relations between the commensal partners.

This research aims to fill the above-mentioned gaps through the following objectives:

ICMI Companion '24, November 4–8, 2024, San Jose, Costa Rica

ACM ISBN 979-8-4007-0463-5/24/11
https://doi.org/10.1145/3686215.3686219

(1) to collect and annotate a new dataset from in-person commensal events;
(2) to conduct baseline experiments using handcrafted features extracted from spatiotemporal Facial Action Units (AUs) with Support Vector Machines (SVM) and raw AUs data modeled with Long Short-Term Memory (LSTM) on the new dataset to demonstrate the feasibility of automatic recognition;
(3) to investigate whether the performance of the aforementioned models varies across the datasets of co-located and online recordings.

We expect that models that deal with the temporal evolution of facial activity (e.g., LSTM) will outperform standard methods. This is because several commensal activities display periodic or repetitive movements and other particularities in the time domain that are difficult to address with standard approaches (see [19]).

Collecting in-person data (i.e., co-located setting) often requires more effort than collecting it online (i.e., remote setting). Additionally, it is not unusual for some activities, especially professional ones, to move often online. Consequently, several online datasets have been proposed recently on various aspects of human social behavior [24, 25], while other researchers studied differences between online and co-located interactions [9, 23]. With another set of experiments, we will see whether there is a domain gap across the co-located and online datasets. Answering this question may shed more light on whether there is a need to collect data in co-located settings or if online data collection can be used instead.

2 RELATED WORK

Past research mainly focuses on detecting specific behaviors such as chewing and food intake, typically for health-related applications, utilizing diverse sensory types and modalities. Commensal actions are rarely addressed in their variety. Certain studies employ visual data to automatically identify activities such as speaking, chewing, and smiling [3, 19]. For instance, in [11], researchers explore the automatic detection of children's smiles and gaze activity.

Audio data is also employed in recognition of commensal activities. For example, [8] introduces a method for recognizing eating habits using a necklace-like device that analyzes throat acoustic signals to detect chewing, swallowing, and breathing. Motion data is another area of focus (e.g., [10, 14]) with accelerometer and gyroscope sensors in smartwatches distinguishing eating, drinking, and smoking.

Finally, [17] is an example of a multimodal model that integrates data from body-worn motion and audio sensors to detect eating activities. The iHEARu-EAT [12] identifies eating conditions from audio-visual data. Other related studies, such as [28] and [20] utilize audio-visual data to detect laughter and smiles.

The bite timing prediction model by [22] is used to control robot-assisted feeding during shared meals. The model was specifically developed to be used in social dining scenarios, and it used multimodal data from all commensal partners. The authors also provide a Human-Human Commensality Dataset (HHCD) containing 30 groups of three people eating together. While HHCD is most related to the work presented in this paper, unfortunately, their data lacks annotation of social behaviors.

Other studies make use of various sensors, including Frequency Modulated Continuous Wave (FMCW) radar sensors [27] for recognizing eating and drinking gestures, and pressure sensors integrated into tables [29] for detecting food-related actions such as cutting and scooping.

3 DATASETS

In this work, we utilized two different datasets: a new dataset, introduced in this paper, aimed at capturing co-located (i.e., in-person) dining, and an existing dataset of remote (i.e., online) dining interactions [19].

3.1 New Dataset: In-person (P) Commensal Activities

Our dataset consists of audio-visual recordings of co-located pairs sharing meals. The collection comprises 22 participants (8 females, 14 males, average age 24) across 12 recording sessions, featuring pairs who either knew each other well or met for the first time. One pair was recorded twice.

The recordings were conducted in a 3×3 square meter room. Participants generally consumed similar food, such as pasta or rice, which required using a fork. Additionally, water and napkins were provided. Such a setup can be considered a more controlled setting with limited meal options and minimal background noise. We believe that these conditions permit the participants to behave spontaneously and naturally. All participants signed the written consent before the recordings.

The videos were recorded using two cameras, as shown in Figure 1. Each recording includes the synchronized view of two participants facing each other. By using OBS Studio software, we guarantee that the recordings of both participants are synchronized. All the videos were recorded at a resolution of 970×710 with a frame rate of 25 fps. Some frames from the dataset are presented in Fig. 2. In total, 234 minutes were recorded, with the shortest session lasting 8 minutes and the longest 39 minutes.

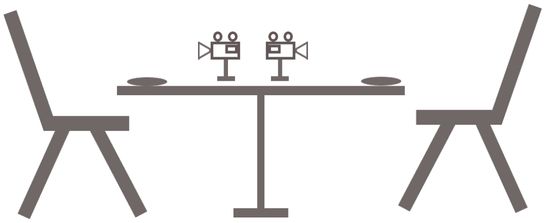

Figure 1: Data collection setup: including the table, the positioning of the cameras, and the arrangement of the plates.

One annotator conducted the annotations manually in ELAN [1]. The annotated food consumption-related behaviors include chewing and food intake, while social signals include speaking and smiling. The choice of the labels was inspired by the previous work [19].

3.2 Online (O) Commensality Dataset

An extended version of the dataset presented in [19] was used for comparison. Compared to the original version described in [19],

Figure 2: Examples of synchronized video frames from two different recordings.

this extended version contains nearly half more recordings done using the same setup. Thus, the extended version consists of 22 videos by 44 participants (26 female) who consumed their meals online. The total duration of the recordings is 191 minutes and 17 seconds. The participants were predominantly aged between 18 and 30 years. The longest video is 18 minutes and 42 seconds; the shortest is 5 minutes and 18 seconds. The recordings capture participants consuming meals together in front of a laptop equipped with a webcam in a natural setting such as the kitchen or dining room. The participants were typically well-acquainted, enhancing the naturalness of the interactions. All videos were standardized to a frame rate of 25 fps with resolutions ranging from 1920 x 1080 to 1280 x 316 pixels (synchronized view). The dataset shares the same annotation schema as the in-person dataset.

3.3 Data preparation

The distribution of commensal activities in both datasets is presented in Table 1. The percentage of smiling in the in-person dataset is much lower than in the online dataset, whereas the percentages for eating, intake, and speaking are similar across the datasets. We suppose that the small number of smiles can be related to the fact that in the co-located dataset, some dyads were composed of people who met themselves for the first time.

The data include the 17 Action Units (AUs) extracted from the recordings using the OpenFace toolkit [5]. The behavior annotations are aligned with the extracted AUs to ensure accurate correspondence between the detected AUs and the behaviors. After merging the annotations and AUs, each recording is segmented into segments of a fixed length of 50 frames (without overlapping).

4 EXPERIMENTS

To address the aims described in Section 1, we performed some experiments on the two datasets introduced in the previous Section.

Commensal Activity	[19] (Online)	OURS (Co-located)
Speaking	1966 (42.6%)	879 (49.4%)
Smiling	643 (13.9%)	64 (3.5%)
Eating	1776 (38.5%)	763 (42.88%)
Intaking	225 (4.9%)	73 (4%)
Total	4610	1779

Table 1: The number of segments per commensal activity

.

4.1 Architecture and training

We applied two approaches: (a) Support Vector Machines and (b) Long Short-Term Memory. The SVM was chosen as it was used in the reference work [19]. We compare the obtained results with a popular method that can handle sequential data, e.g., time series of facial activity. For this reason, we chose LSTM.

A 5-fold cross-validation was used in line with [19]. We conducted a grid search to optimize hyperparameters for both SVM and LSTM, using a fixed subset (equivalent to 10% of the entire dataset) of a single training fold. The SVM hyperparameters included the RBF kernel. The parameter C, which balances between maximizing the margin and minimizing classification error, was tested with values 0.1, 1, 10, 100, 1000, and 10000. Similarly, γ, determining the RBF kernel's reach and fit training data, was explored with values: 0.0001, 0.001, 0.01, 0.1, 1.

The LSTM architecture used in our experiments consisted of three LSTM layers. The first LSTM layer has 64 units and takes an input shape of (50, 17), where 50 represents the time steps, and 17 denotes the number of AUs. The second LSTM layer has 32 units, and the third LSTM layer has 16 units. The output layer is a Dense layer with 4 units corresponding to the number of classes, and it utilizes the softmax activation function. The training was conducted using the ADAM optimizer for 64 epochs, batch size 32, and the categorical cross-entropy loss function.

We conducted evaluations using two approaches: within-dataset and cross-dataset. The training, validation, and test splits in the within-dataset approach all came from the same dataset. In the cross-dataset approach, we trained the models on one dataset (including the hyperparameter tuning based on the performance of the validation split) and tested them on a different dataset. Finally, we merged the two datasets and conducted a within-dataset evaluation on the combined data. As evaluation metrics, we report precision (Pre), recall (Rec), weighted (w-F1), and macro F1-score (m-F1), and accuracy (Acc).

4.2 Results

The results of the within-dataset and cross-dataset evaluations are presented in Table 2. The LSTM outperforms the SVM, indicating that the raw AU values are more effectively learned and used for differentiating various commensal activities. As anticipated, the cross-dataset experiments yield lower performance due to the domain gap. However, the hand-crafted features modeled with SVM generalize better than LSTM, as evidenced by the higher SVM scores in the P→O and O→P evaluations. This difference may stem from the fact that the LSTM architecture, including the number of LSTM layers, ideally requires adjustment to optimize performance for different training data sizes. However, in our study, we maintained a fixed LSTM architecture across all experiments, potentially limiting its adaptability and performance in varying data scenarios. Once the datasets are merged, the size of the training data increases importantly. As expected, both models perform numerically better; however, LSTM surpasses SVM also in this scenario.

5 DISCUSSIONS AND CONCLUSIONS

In this paper, we have introduced a novel dataset collected in person and have utilized it to model the recognition of commensal activities. For this purpose, we employed SVM with hand-crafted features from the literature and LSTM to model AUs extracted with a standalone toolbox. Furthermore, we evaluated these models on a distinct dataset where interactions occur in an online environment, which significantly differs from in-person settings. Our findings highlight that models trained on data from one setting struggle to generalize well to data collected from a different context, emphasizing the critical role of context in behavioral analysis. LSTM performs better; however, neither method generalizes well across datasets. LSTM's performance in cross-dataset evaluations lags behind SVM.

When a larger dataset of online and in-person interactions is used in training, the inference performance of both methods improves, and LSTM outperforms SVM across all metrics.

Despite the online dataset used in this study being larger than the in-person dataset, the methods employed on the in-person dataset, especially LSTM, demonstrated superior performance overall. We hypothesize that the superior video quality in a co-located setting may enhance the accuracy of AU extraction. Also, online interactions may be impacted by delays or pixelation, which could reduce AU extraction accuracy. Moreover, the videos in the in-person dataset were taken in the same environment (lighting, camera position, etc.), while the online dataset is much more variable. On the other hand, a co-located setup allows the commensal partners much larger freedom and variety of movements, as the participants do not need to worry about moving out of the camera view, which is often a concern in online meetings (see, e.g., [30]). In co-located settings, they can interact physically, e.g., by passing objects such as cutlery or food or touching each other. These and many other behaviors may increase the difficulty of commensal activity recognition from co-located recordings. Nevertheless, the results highlight the importance of increasing the amount of data for better generalization.

Approach	Learning Method	Prec	Rec	w-F1	m-F1	Acc
O→O	SVM	0.66	0.57	0.73	0.62	0.74
	LSTM	0.83	0.80	0.84	0.81	0.85
P→P	SVM	0.69	0.65	0.83	0.67	0.83
	LSTM	0.87	0.87	0.93	0.87	0.93
O→P	SVM	0.46	0.56	0.66	0.46	0.64
	LSTM	0.42	0.55	0.60	0.41	0.55
P→O	SVM	0.53	0.44	0.61	0.45	0.64
	LSTM	0.48	0.40	0.59	0.41	0.61
M→M	SVM	0.65	0.55	0.74	0.57	0.76
	LSTM	0.79	0.86	0.87	0.82	0.87

Table 2: The best results for SVM and LSTM are reported. Here, *O* (i.e., online) refers to the extended version of [19], *P* (in-person) denotes our dataset, and *M* represents the merged dataset of *O* and *P*. In the notation used, the left side of → indicates the dataset used for training and parameter tuning (with validation split), while the right side indicates the dataset used for testing. The abbreviations Pre, Rec, W-F1, M-F1, and ACC stand for precision, recall, weighted F1-score, macro F1-score, and accuracy, respectively.

Our study is limited regarding the cues used to classify commensal activities, and it only focuses on facial Action Units. As a future study, we plan to incorporate features e.g., learned from the body pose [6], and upper body activity [26]. Additionally, we will explore multimodal approaches as described in [7]. The other research direction will address commensal partners' relations and interaction dynamics. For this reason, we will extract information about the nonverbal behaviors of individuals from the synchronized videos and focus on the temporal relations between their actions. Later on, the behavior data of other commensal partners will be used as context to build better recognition models. For example, it might be rare for all commensal partners to speak simultaneously, while it can be quite probable for all to be chewing simultaneously.

In conclusion, we envision that models in detecting social signals and food-related actions can potentially enhance the functionality of various technologies. Specifically, we anticipate that this contribution will improve the performance of chewing trackers or robot-assisted feeding. More importantly, it will contribute to the development of artificial commensal companions [16], enabling more natural and smoother interaction between humans and machines.

ACKNOWLEDGMENTS

This work is supported by the PRIN 2022 project COCOA (D53D23008850001), funded by the Next Generation EU (NGEU) Programme, National Recovery and Resilience Plan (PNRR) and by the Italian Ministry of University and Research.

REFERENCES

[1] Pierre-Emmanuel Aguera, Karim Jerbi, Anne Caclin, and Olivier Bertrand. 2011. ELAN: a software package for analysis and visualization of MEG, EEG, and LFP signals. *Computational intelligence and neuroscience* 2011, 1 (2011), 158970.

[2] Reem K Al-Halimi and Medhat Moussa. 2016. Performing complex tasks by users with upper-extremity disabilities using a 6-DOF robotic arm: a study. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 25, 6 (2016), 686–693.

[3] Sana Alshboul and Mohammad Fraiwan. 2021. Determination of chewing count from video recordings using discrete wavelet decomposition and low pass filtration. *Sensors* 21, 20 (2021), 6806.

[4] Oliver Amft, Martin Kusserow, and Gerhard Troster. 2009. Bite weight prediction from acoustic recognition of chewing. *IEEE Transactions on Biomedical Engineering* 56, 6 (2009), 1663–1672.

[5] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. OpenFace 2.0: Facial Behavior Analysis Toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. 59–66. https://doi.org/10.1109/FG.2018.00019

[6] Cigdem Beyan, Vasiliki-Maria Katsageorgiou, and Vittorio Murino. 2017. Moving as a leader: Detecting emergent leadership in small groups using body pose. In *Proceedings of the 25th ACM international conference on Multimedia*. 1425–1433.

[7] Cigdem Beyan, Alessandro Vinciarelli, and Alessio Del Bue. 2023. Co-Located Human–Human Interaction Analysis Using Nonverbal Cues: A Survey. *Comput.*

Surveys 56, 5 (2023), 1–41.

[8] Yin Bi, Wenyao Xu, Nan Guan, Yangjie Wei, and Wang Yi. 2014. Pervasive eating habits monitoring and recognition through a wearable acoustic sensor. In *Proceedings of the 8th International Conference on Pervasive Computing Technologies for Healthcare*. 174–177.

[9] Julie Boland, Pedro Fonseca, Ilana Mermelstein, and Myles Williamson. 2021. Zoom Disrupts the Rhythm of Conversation. *Journal of Experimental Psychology: General* 151 (11 2021). https://doi.org/10.1037/xge0001150

[10] Eleni Diamantidou, Dimitrios Giakoumis, Konstantinos Votis, Dimitrios Tzovaras, and Spiridon Likothanassis. 2022. Comparing deep learning and human crafted features for recognising hand activities of daily living from wearables. In *2022 23rd IEEE International Conference on Mobile Data Management (MDM)*. IEEE, 381–384.

[11] Dhia-Elhak Goumri, Thomas Janssoone, Leonor Becerra-Bonache, and Abdellah Fourtassi. 2023. Automatic Detection of Gaze and Smile in Children's Video Calls. In *Companion Publication of the 25th International Conference on Multimodal Interaction*. 383–388.

[12] Simone Hantke, Felix Weninger, Richard Kurle, Fabien Ringeval, Anton Batliner, Amr El-Desoky Mousa, and Björn Schuller. 2016. I hear you eat and speak: Automatic recognition of eating condition and food type, use-cases, and impact on asr performance. *PloS one* 11, 5 (2016), e0154486.

[13] Rohit Ashok Khot, Eshita Sri Arza, Harshitha Kurra, and Yan Wang. 2019. Fobo: Towards designing a robotic companion for solo dining. In *Extended abstracts of the 2019 CHI conference on human factors in computing systems*. 1–6.

[14] Konstantinos Kyritsis, Christos Diou, and Anastasios Delopoulos. 2020. A data driven end-to-end approach for in-the-wild monitoring of eating behavior using smartwatches. *IEEE Journal of Biomedical and Health Informatics* 25, 1 (2020), 22–34.

[15] Jie Li, Na Zhang, Lizhen Hu, Ze Li, Rui Li, Cong Li, and Shuran Wang. 2011. Improvement in chewing activity reduces energy intake in one meal and modulates plasma gut hormone concentrations in obese and lean young Chinese men. *The American journal of clinical nutrition* 94, 3 (2011), 709–716.

[16] Maurizio Mancini, Radoslaw Niewiadomski, Gijs Huisman, Merijn Bruijnes, and Conor Patrick Gallagher. 2020. Room for One More? - Introducing Artificial Commensal Companions. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems Extended Abstracts* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–8. https://doi.org/10.1145/3334480.3383027

[17] Mark Mirtchouk, Drew Lustig, Alexandra Smith, Ivan Ching, Min Zheng, and Samantha Kleinberg. 2017. Recognizing eating from body-worn sensors: Combining free-living and laboratory data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 1–20.

[18] Radoslaw Niewiadomski, Eleonora Ceccaldi, Gijs Huisman, Gualtiero Volpe, and Maurizio Mancini. 2019. Computational Commensality: from theories to computational models for social food preparation and consumption in HCI. *Frontiers in Robotics and AI* 6 (2019), 119. https://doi.org/10.3389/frobt.2019.00119

[19] Radoslaw Niewiadomski, Gabriele De Lucia, Gabriele Grazzi, and Maurizio Mancini. 2022. Towards Commensal Activities Recognition. In *Proceedings of the 2022 International Conference on Multimodal Interaction*. 549–557. https://doi.org/10.1145/3536221.3556566

[20] R. Niewiadomski, M. Mancini, G. Varni, G. Volpe, and A. Camurri. 2016. Automated Laughter Detection From Full-Body Movements. *IEEE Transactions on Human-Machine Systems* 46, 1 (Feb 2016), 113–123. https://doi.org/10.1109/THMS.2015.2480843

[21] Elinor Ochs and Merav Shohet. 2006. The cultural structuring of mealtime socialization. *New directions for child and adolescent development* 2006, 111 (2006), 35–49.

[22] Jan Ondras, Abrar Anwar, Tong Wu, Fanjun Bu, Malte Jung, Jorge Jose Ortiz, and Tapomayukh Bhattacharjee. 2022. Human-robot commensality: Bite timing prediction for robot-assisted feeding in groups. In *6th Annual Conference on Robot Learning*.

[23] Grégoire Python, Cyrielle Demierre, Marion Bourqui, Angelina Bourbon, Estelle Chardenon, Roland Trouville, Marina Laganaro, and Cécile Fougeron. 2023. Comparison of In-Person and Online Recordings in the Clinical Teleassessment of Speech Production: A Pilot Study. *Brain Sciences* 13, 2 (2023). https://doi.org/10.3390/brainsci13020342

[24] Justine Reverdy, Sam O'Connor Russell, Louise Duquenne, Diego Garaialde, Benjamin R. Cowan, and Naomi Harte. 2022. RoomReader: A Multimodal Corpus of Online Multiparty Conversational Interactions. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association, Marseille, France, 2517–2527. https://aclanthology.org/2022.lrec-1.268

[25] Inka Schmitz and Wolfgang Einhäuser. 2023. Gaze estimation in videoconferencing settings. *Computers in Human Behavior* 139 (2023), 107517. https://doi.org/10.1016/j.chb.2022.107517

[26] Muhammad Shahid, Cigdem Beyan, and Vittorio Murino. 2019. Comparisons of visual activity primitives for voice activity detection. In *Image Analysis and Processing–ICIAP 2019: 20th International Conference, Trento, Italy, September 9–13, 2019, Proceedings, Part I 20*. Springer, 48–59.

[27] Chunzhuo Wang, T Sunil Kumar, Walter De Raedt, Guido Camps, Hans Hallez, and Bart Vanrumste. 2022. Eat-Radar: Continuous Fine-Grained Eating Gesture Detection Using FMCW Radar and 3D Temporal Convolutional Network. *arXiv preprint arXiv:2211.04253* (2022).

[28] Fan Yang, Mohamed A Sehili, Claude Barras, and Laurence Devillers. 2015. Smile and laughter detection for elderly people-robot interaction. In *Social Robotics: 7th International Conference, ICSR 2015, Paris, France, October 26-30, 2015, Proceedings 7*. Springer, 694–703.

[29] Bo Zhou, Jingyuan Cheng, Mathias Sundholm, Attila Reiss, Wuhuang Huang, Oliver Amft, and Paul Lukowicz. 2015. Smart table surface: A novel approach to pervasive dining monitoring. In *2015 IEEE International Conference on Pervasive Computing and Communications (PerCom)*. IEEE, 155–162.

[30] Julian Zubek, Ewa Nagórska, Joanna Komorowska-Mach, Katarzyna Skowrońska, Konrad Zieliński, and Joanna Rączaszek-Leonardi. 2022. Dynamics of Remote Communication: Movement Coordination in Video-Mediated and Face-to-Face Conversations. *Entropy* 24, 4 (2022). https://doi.org/10.3390/e24040559