https://doi.org/10.1145/3536221.3556566

Towards Commensal Activities Recognition

Radoslaw Niewiadomski
University of Trento
Rovereto, Italy
r.niewiadomski@unitn.it

Gabriele De Lucia
Sapienza University of Rome
Rome, Italy
delucia.1811284@studenti.uniroma1.it

Gabriele Grazzi
University of Trento
Rovereto, Italy
gabriele.grazzi@studenti.unitn.it

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

ABSTRACT

Eating meals together is one of the most frequent human social experiences. When eating in the company of others, we talk, joke, laugh, and celebrate. In the paper, we focus on commensal activities, i.e., the actions related to food consumption (e.g., food chewing, in-taking) and the social signals (e.g., smiling, speaking, gazing) that appear during shared meals. We analyze the social interactions in a commensal setting and provide a baseline model for automatically recognizing such commensal activities from video recordings. More in detail, starting from a video dataset containing pairs of individuals having a meal remotely using a video-conferencing tool, we manually annotate commensal activities. We also compute several metrics, such as the number of reciprocal smiles, mutual gazes, etc., to estimate the quality of social interactions in this dataset. Next, we extract the participants' facial activity information, and we use it to train standard classifiers (Support Vector Machines and Random Forests). Four activities are classified: chewing, speaking, food in-taking, and smiling. We apply our approach to more than 3 hours of videos collected from 18 subjects. We conclude the paper by discussing possible applications of this research in the field of Human-Agent Interaction.

CCS CONCEPTS

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

KEYWORDS

computational commensality, facial expressions, activity recognition, social signal processing

ACM Reference Format:
Radoslaw Niewiadomski, Gabriele De Lucia, Gabriele Grazzi, and Maurizio Mancini. 2022. Towards Commensal Activities Recognition. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '22), November 7–11, 2022, Bengaluru, India.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3536221.3556566

ICMI '22, November 7–11, 2022, Bengaluru, India
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9390-4/22/11…$15.00
https://doi.org/10.1145/3536221.3556566

1 INTRODUCTION

Eating together in a group, often called "commensality" [36], is one of the most profound social experiences. It does not matter whether it is a business lunch, a romantic date in a cozy restaurant, or a family Christmas dinner; commensal events are occasions to consume food and talk, meet new people, share experiences, socialize, and socially entertain ourselves. Several positive impacts of commensality can be observed, including balanced food in-taking and better food choices, positive triggered emotions, and well-being [4, 13, 16]. Conversely, a lack of social eating may negatively affect mental and physical health [19, 45].

Recently, researchers in HCI started to investigate how interactive technologies and artificial intelligence could contribute to improving commensality experiences [34, 41]. Examples of such technologies include using video-conferencing tools to share eating moments (e.g., video-conferencing tools [11]) or developing Artificial Commensal Companions [17, 29] - embodied agents (e.g., social avatars and robots) that provide company to humans during meals.

The interaction around the table is quite peculiar, as the involved partners constantly shift their attention between the conversation and the food. Not only are we interacting with one or more partners simultaneously, but we are also focusing on them and the consumed food and drinks. Moreover, commensal events follow specific social rules and culture-based rituals (such as passing food, serving the wine, positions at the table, etc.) that make commensal interaction distinctive and unique. Finally, commensal interactions are profoundly multimodal, since facial expressions (e.g., chewing, smiling), gaze, and body actions (e.g., food in-taking) accompany speech and other sounds (e.g., slurping).

Considering the ubiquity of the commensal experience, it is surprising that commensal activity recognition models are still scarce. The paper addresses this research gap by providing a first commensal activity recognition model that considers various activities during shared meals. Notably, the term "commensal activities" includes actions related to the consumption of the food (e.g., chewing, in-taking) and the social signals (e.g., smiling, speaking, gazing, passing the food, cheering...). So far, researchers have only addressed these activities separately. Our work is probably the first attempt to address the variety of (nonverbal) behaviors that may appear during the commensal experience.

To reach this goal, we use data containing several pairs of friends sharing meals online. This form of commensality became popular during the Covid-19 pandemic. For this work, we perform manual annotation of nine shared meals. This allows us to analyze the social signals of remote commensal partners. The paper's final part illustrates a first "baseline" recognition model. While the commensality experience is undoubtedly multimodal, we start our exploration by focusing on video data only. We extract information about facial activity from videos and apply a supervised feature-based machine learning approach for commensal activity classification. Our goal is also in line with previous studies on automated detection of multimodal behaviors, e.g., speaking [5] and laughter detection [35], using video data only.

The rest of the paper is organized as follows. In the next section, we briefly present previous works on eating and social signal detection. In Section 3 we introduce the dataset and the annotation. Section 4 explains details of the recognition models and the validation results. In Section 5 we present possible applications.

2 STATE OF THE ART

While we are unaware of any research on video-based activity recognition in commensality settings, some related works were realized in assistive technology, in which robot-based systems were used, e.g., to feed a person. For instance, the system in [37] delivers food from a bowl to the user's mouth. The user selects a preferred task via the GUI on a tablet. The robot automatically estimates food location, scoops it and places it inside the user's mouth. The system uses a 3D mouth pose estimator and a wrist-mounted RGB-D camera. In the same line, [10] use depth images to track the user's mouth and a separate vision system mounted on the robot to collect information about the amount of acquired food on the robotic spoon.

A few more works focus on tracking users to measure and improve their eating habits. Rouast and Adam [39] propose video-based detection of food in-taking gestures using deep learning. Cadavid and Abdel-Mottaleb [9] discriminate between chewing and non-chewing facial actions, such as talking, by exploiting spectral analysis of facial features. The approach builds on the observation that chewing movements are usually periodic while non-chewing ones are not. In a recent technical report [6], Bi and Kotz describe a system with a head-mounted camera to discriminate eating vs. non-eating behaviors. Recently, Hossain et al. [22] proposed an automatic bite and chews counter using a pre-trained AlexNet network. It is worth noticing that this work uses the data collected in a commensal setting, where three persons are sitting at the table.

Several techniques exist for eating-related activities (such as chewing and swallowing) detection, which use wearable devices (e.g., Amft and Tröster [2], Bi et al. [7], Fontana et al. [15]). For instance, Fontana et al. [15] developed a wearable system composed of a jaw motion sensor, a hand gesture sensor, and an accelerometer placed on the chest. The system is integrated into a smartphone equipped with a food in-taking recognition module that uses dedicated sensor fusion and pattern recognition techniques. It can detect food in-taking with an average accuracy of 89.8%.

In Mendi et al. [31], bites-taken rate and eating speed are measured using an accelerometer placed on the user's wrist. Rahman et al. [38] use Google Glasses to track head movements and to show that inertial data from the device and standard machine learning techniques can be used to recognize human's eating activities.

Regarding audio processing, Hantke et al. [20] classify normal speech and eating speech and detect the type of food consumed while speaking. Individual bite weight prediction is carried out from acoustic data through an ear-pad sound sensor by Amft and colleagues [1]. Finally, a multimodal approach for eating recognition combining head and wrist motion (captured with Google Glass and smartwatches on each wrist) with audio (custom earbud microphone) is proposed in [32].

Regarding the social dimension of eating, Kiriu et al. [27] detect whether a person is eating alone or in the company of others, using a smartwatch accelerometer data and several metrics of a smartphone. Several works in the field of Social Signal Processing [8] focus on detecting single social behavior such as smiling, laughing, and speaking. Still, none of them is performed in a commensality setting. For instance, Beyan and colleagues [5] detect the speaker in the group by processing video data only. Unlike traditional audio processing approaches, they perform upper body motion analysis using deep learning techniques. Support-Vector Machine (SVM) combined with hand-crafted features is proposed in [14] for the automatic classification of spontaneous vs. posed enjoyment smiles from a video. The latter is a type of smile frequently playing the role of a social signal [30]. Similarly, Griffin and colleagues [18] distinguish between hilarious and social laughter using motion capture data of body movements, hand-crafted features, and traditional machine learning techniques, e.g., Random Forests.

From the above summary, it is evident that, although there exist some works that focus on single activity detection/recognition, they do not focus on the variety of activities present in a commensal setting. Eating recognition is usually done in a single-person setting (and under lab conditions). At the same time, social signals are analyzed in various contexts, but none of the works was realized in a commensal setting. We introduce the first commensal activities recognition model to address this gap in the following sections.

3 DATASET AND ANNOTATION

One possible reason for the lack of commensal activities recognition models in the literature could be the lack of suitable datasets. Such datasets should contain audio and video data (although other sensors might also be considered, see the previous section) of at least two commensal partners sharing a meal in an ecological setting (e.g., kitchen, restaurant, etc.). Their data should be collected possibly in a low-invasive manner, not to compromise the experience (both in a social and gastronomical sense).

3.1 Dataset

We collected recordings of participants pairs (i.e., dyads) eating together in a video call. We asked them to prepare a single-course meal (e.g., some pasta) in advance and to eat in the company of a friend or relative. Due to the Covid-19 restrictions in most of the world in 2021, participants used online meeting software to share their meals. Apparently, this form of commensality may still provide a sense of belonging and togetherness to remote commensal partners [11]. At the same time, this setup allowed us to collect audiovisual data of spontaneously behaving people eating in an ecological setting. Indeed, they consumed meals in their natural environment (e.g., kitchen at their homes) and in the company of people they knew well. Moreover, we expected that, due to the Covid-19 restrictions, the use of a teleconference tool for this purpose was not unusual for the participants as it could have been some time ago. In more detail, we recruited participants pairs by asking them if they were interested in the study and willing to participate. After receiving their confirmation, we emailed each participant the instructions and an identification code. Then, we fixed an appointment for each pair, and, on the appointment date and time, we recorded them sharing a meal in a video conference.

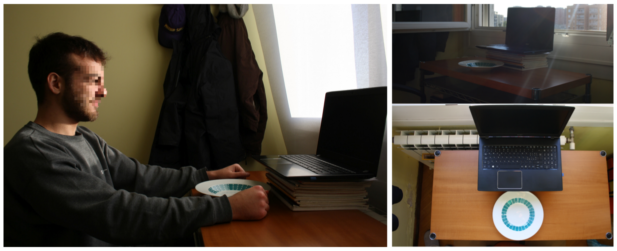

Figure 1: The data collection setup

Figure 1depicts the technical setup. Participants sat in front of their computer's camera, so their faces and upper bodies were visible to the interaction partner. One of the experimenters was present during the call to supervise the recording process, but the experiment's camera and microphone were turned off, so as not to interfere with the interaction. The video-conference software used for data collection permits the creation of a single video file of the two participants in sync (see Figure 2). The videos are at 25 fps and with a resolution of 1280 × 720 pixels (synchronized view).

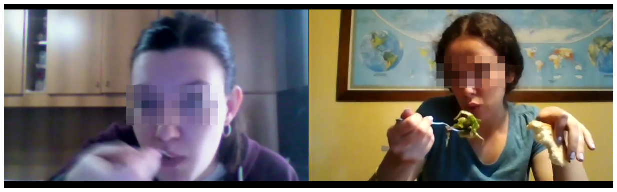

Figure 2: Participants 1 and 2 chatting while remotely eating.

Participants could freely choose any conversation topic they wanted during the meals. However, we provided a subsidiary list of conversation topics to facilitate the chat in case they did not feel at ease finding a discussion topic.

Twenty-two people participated in the data collection. Before the data collection, we asked participants to provide basic personal data. The participants' gender was balanced (with 56% of females), and the majority (72%) were between 18 and 24 years old. Most of them declared to know very well their interaction partners. Before the recordings, participants signed a formal consent allowing us to collect their data and share it in an anonymized version (following the EU GDPR rules).

3.2 Manual Annotation

A trained annotator manually annotated the commensal activities of the 18 participants using the ELAN software [44] (due to technical issues, the audio recordings of two pairs were lost). Even if the goal of the annotation process was to annotate only the visible behaviors of the participants, the annotator used both audio and video information to perform this task. We annotated the following commensal activities:

- activity 1 - speaking
- activity 2 - food/drink in-taking
- activity 3 - chewing
- activity 4 - smiling/laughing
- activity 5 - gazing

As for the gaze behavior (activity 5), the annotation distinguished between a) "gaze at plate" and b) "gaze at commensal partner".

This set of labels is a combination of activities related to food consumption (activity 2, activity 3, activity 5a) and social signals (activity 1, activity 4, activity 5b). In particular, smiling is an important social signal [28] that may serve several functions, such as regulating the interaction flow and turn-taking [25]. It may also help to estimate the quality of interaction [12]. For example, annotations of *speaking* and *smiling* can be used to estimate the social bonds between partners [23]. At the same time, detecting food in-taking (activity 2) might be used to estimate the rhythm and velocity of eating (see, e.g., [39]). In the annotation, the food in-taking event starts when the fork movement begins to be visible in the camera frame and ends when the food is put into the mouth. Thus, it differs from activity 3, i.e., chewing the food.

The annotator annotated each person and activity separately. Indeed, different commensal activities may be performed simultaneously by the same person (e.g., speaking and eating), or, the same activity (e.g., chewing) can be performed by two or more partners simultaneously. Thus, annotations of the activities can overlap, e.g., when the person i speaks and eats simultaneously or when persons i and j speak simultaneously.

The average duration of annotated videos is 9 minutes and 23 seconds (with a minimum of 5 minutes and 18 seconds and a maximum of 16 minutes and 17 seconds). The total time of the annotated videos is 95 minutes and 57 seconds (each video shows two persons, see Figure 2).

The manual annotation provides interesting information about pairs eating remotely. Figure 3 shows the results of the annotation process. As expected, the two main activities are *speaking* and *chewing*. The percentage of time dedicated to speaking or chewing varies a lot between participants, from 19% (speaking) and 21.1% (chewing) to 51.4% (speaking) and 60.9% (chewing). The average percentage for speaking is 34% (SD=9.1%), and the one for chewing is 40.1% (SD=11.5%). The smiling and food in-taking activities appear more rarely. The minimum percentage of smiling is 1.2%, the maximum is 29.1%, and average is 14.5% (SD=7%). Finally, the percentage of time dedicated to food in-taking is stable across participants, from 6.5% to 17%, with an average of 11.6% and a standard deviation of 3.3%. Regarding gaze behavior, surprisingly, participants spent a high percentage of time looking at each other, as the label *look towards other person* was used between a minimum of 28.6% and a maximum of 73.6% of the time (average 51.3%, SD 14.2%). That is, the participants look at the screen displaying the commensal partners for around half of the meal time. Finally, the *look at plate* action was annotated between a minimum of 19.7% and a maximum of 62.9% of the time (average 37%, SD 14.9%).

Starting from the above annotation, for each pair, we computed metrics to inform us of the quality of the social interaction. We checked the percentage of time participants speak simultaneously (H1). High values may indicate low cohesion of the partners [23], but, in the case of remote meetings, it can also be caused by a poor Internet connection. In both cases, it may negatively influence the perception and satisfaction of the interaction. We also computed the percentage of mutual (H2) and non-mutual gazes (H3) and their ratio (H4), as well as the percentage of mutual (H5) and non-mutual smiling time (H6) and their ratio (H7). To check whether or not the interaction between commensal partners is balanced, we also computed speech imbalance H8: we take the absolute difference between the number of frames in which the first participant speaks and the second participant does not speak and the number of frames in which the second participant speaks and the first participant does not speak, normalized by the total number of frames in the segment. We applied a similar approach to compute smile imbalance (H9). The idea is that if both partners speak/smile about the same amount of time, H8/H9 get close to 0; when one speaks/smiles longer than the other, H8/H9 gets higher. The speech and/or smiles imbalance may be an essential social feature indicating, e.g., the dominance of one of the partners [21]. Results are reported in Table 1.

Again, interesting differences can be observed between the participants pairs. The interaction of the pair P1-P2 is characterized by the highest gaze ratio (149.18) and the second highest smile ratio (42.16), which means that the mutual gaze and smile are remarkably long for them. In particular, the mutual gazes are longer than the sum of non-mutual gazes. The pair P3-P4 could be placed at the other extreme, with the lowest gaze and smile ratios (33.17 and 1.10). When summing H8 with H9, P3-P4 also appears to be the most imbalanced pair, while P7-P8 is the most balanced one.

In general, from these results, it can be deduced that the interactions are smooth and rich. The percentage of overlapping speech time is only 1.6%, on average (with a maximum of 3.6% of the time); the duration of the mutual gaze is, on average, 27.85% (with a minimum of 17%), while at least one person of the pair is smiling, 22% of the time (with a maximum of 35.42%).

Results also show that the range of activities considered in the annotation might be very useful to describe the quality of the interaction, and in particular, the social dynamics. Being able to detect these activities automatically would help to improve such analyses in the future. In the next section, we provide a baseline method for classifying 4 out of the 5 activities considered in this section.

4 COMMENSAL ACTIVITIES CLASSIFICATION

In this section, we aim to show that using video data only is feasible for commensal activity recognition. Our baseline model uses a feature-based approach applied in the past to solve other research questions in Affective and Social Computing [35]. The main novelty is applying such an approach to the specific domain of commensal activity recognition. For this reason, we use only two classification algorithms and apply standard methods for extracting nonverbal behaviors from video data. While other communication channels (e.g., gaze and upper body movements) are also relevant for commensal activity recognition, in this first attempt, we focus on facial expression only.

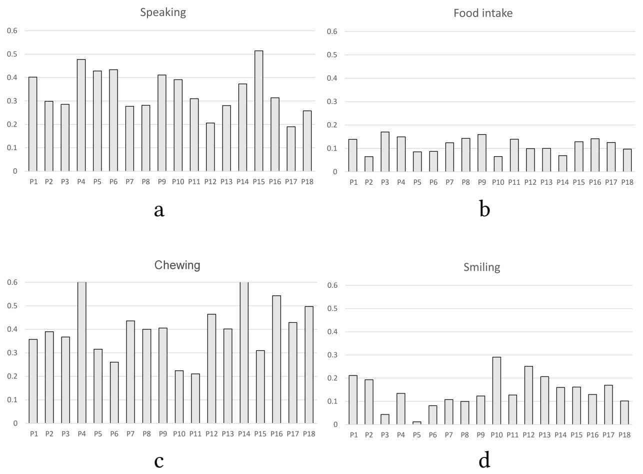

Figure 3: The percentage of time each participant dedicates to one of four annotated activities: 1) speaking, 2) food intaking, 3) chewing, 4) smiling. P1-P18 (x-axis) is the participants' ID.

4.1 Data Preparation

A single video frame offers a snapshot of reality, from which, however, it is difficult, if not impossible, to understand whether the recorded person is eating, speaking, or not doing any of the activities of interest. For example, let us imagine that, in one frame, we see a person with his mouth closed (see Figure 4). Their mouth could be in that position because of different reasons: 1) they were not speaking for an extended period (e.g., they are listening to

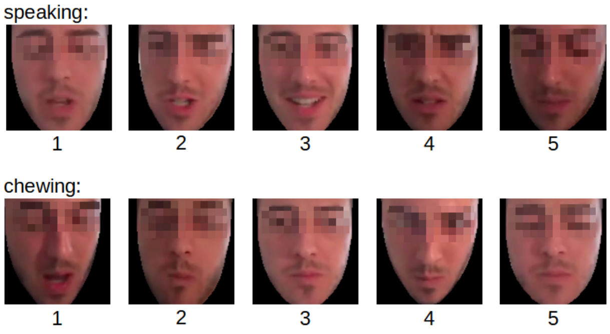

Figure 4: Sample frames of participant 5, classes 2 - speaking (upper row) and 4 - chewing (lower row). These samples show that, by considering single frames instead of time windows, the two types of activities will look very similar either between them (see, for example, frames 1 and 5 of each row) or compared to other activities (e.g., with class 5 - smiling, see frame 3 in the upper row).

Table 1: Social metrics on the dataset. The values of H1-H3, H5-H6, and H8-H9 are expressed as time percentages. We also report the identifier of the corresponding pair for minimum and maximum values.

ID	Name	Minimum	Min. Pair	Maximum	Max. Pair	Average
H1	Overlapping speech	0	P3-P4	3.6	P5-P6	1.67
H2	Mutual gaze	17.11	P1-P2	49.95	P3-P4	27.85
H3	Non-mutual gaze	33.49	P1-P2	51.57	P3-P4	43.22
H4	Ratio mutual/non-mutual gaze	33.18	P3-P4	149.18	P1-P2	69.11
H5	Mutual smile	0.19	P3-P4	12.00	P1-P2	6.14
H6	Non-mutual smile	6.73	P5-P6	35.43	P11-P12	22.64
H7	Ratio mutual/non-mutual smile	1.10	P3-P4	46.79	P15-P16	27.53
H8	Speech imbalance	0.45	P7-P8	20.1	P15-P16	8.77
H9	Smile imbalance	0.84	P7-P8	16.76	P9-P10	7.1

the other person speaking); 2) their lips are in a short pause between two words (or even syllables); 3) they are in a pause between one chewing and another. Thus, we focus on sequences of frames annotated as activities 1-4, and we create fixed-length segments attributing them labels of classes 1-4. But, as mentioned earlier, a participant can perform more than one activity simultaneously. So, we decided to skip all the annotated video frames with more than one label. Thus, when two activities partially overlap, we only consider the video frames from the beginning of the first activity to the beginning of the second activity (that is, from the point in time in which the first and second activity starts to overlap) and from the end of the first activity to the end of the second one. That is, for the moment, we skip the video frames in which both activities appear together.

Regarding the length of the segments, we consider three different segment lengths: 10, 25, and 50 frames. We choose the segment length of 25 as it corresponds to 1 second of the data. We believe that the activities we are considering (smiling, chewing, speaking) usually last about, at minimum, 1 second. We compare them with 2 seconds segments (50 frames), which can be seen as an upper bound, as some of the activities, such as short utterances or food in-taking, might last at most 2 seconds. However, we obtained several relatively short frame sequences in segmenting the data due to dropping overlapping annotations. Thus, we also check whether more fine-grained segments of 400ms (10 frames) can allow for a more (or less) effective activity recognition. Due to the segmenting process, the activities lasting less than a given segment length (i.e., 10, 25, 50 frames) are discarded (12% for 10 frames sequences; 30% for 25 frames sequences; 52% for 50 frames sequences).

Table 2 shows the final number of segments per class. We can see that classes are imbalanced. During a meal, the most frequently appearing activity is speaking (class 1, 45.9%, 45.0%, and 43.3%), and the next one is chewing (class 3, 31.9%, 33.8%, and 38.7%). That is not surprising, as the video recordings cover only the actual meal time (they were immediately interrupted when participants finished eating). Less frequent activities are food in-taking (class 2, 11.6%, 10.6%, and 6.9%) and laughing/smiling (class 4, 10.5%, 10.6%, and 11.1%). In particular, a low number of 2 seconds segments of class 2 shows that this activity might often be shorter than 50 frames.

At the same time, we also observe significant differences between participants. For instance, in the case of 10 frame segments, the number of class 1 segments varies across participants, from 107 to 629. Class 2 segments range from 34 to 217, class 3 from 73 to 575, and class 4 from 16 to 202.

As our dataset consists of video frame sequences, we use the freely available software called OpenFace [3] to extract facial features. In particular, we focus on the intensity of the 17 Action Units (AUs) computed by OpenFace on each video frame. Then, we calculate six statistical measures on the AUs of each segment: min, max, mean, standard deviation, skewness, and kurtosis. Using these six measures for each feature, the input vector of the classifier has a size of $6 \times 17 = 102$.

4.2 Experiments

The presented study uses two machine learning approaches: Random Forest (RF) and Support Vector Machines (SVM). We choose these two techniques as they have been widely used in the past to classify human internal states and nonverbal behavior (e.g., [33, 35]). We perform a set of experiments using different segment sizes and validation methods. All the data is normalized before training using *z-normalization*.

In the first step, we tune, train, and test the two models on original dataset and using cross-validation method. The first one is an SVM with a Radial Basis Function kernel on the original unbalanced dataset. We use grid research for parameters tuning with C being consecutive powers of 10 in the range $0..4$ and γ being consecutive powers of 10 in the range $-6..-1$; the parameter search uses a 5-fold cross-validation with $F\_weighted$ score. We also use a 5-fold cross-validation for the outer loop. We repeat the same training and testing procedure for the 3 segment lengths (10, 25, 50). The second model is RF, for which we use a Random search for parameters tuning in a 3 folds cross-validation, limited to 100 randomly picked parameters combinations. Again, we use a 5-fold cross-validation for the outer loop, and we repeat the procedure for 3 segment lengths.

In a second step, we train and test SVM and RF models using the leave-one-subject-out method.

In a third step, we repeat the approach described in the first step on a balanced dataset. Balancing was performed by combining two different undersampling algorithms and an oversampling one: 1) Random UnderSampler and SMOTE OverSampler; 2) NearMiss UnderSampler and SMOTE OverSampler. As a result, in both cases, the samples number for all classes becomes identical to the initial average (computed on all classes) number of samples.

Table 2: Number of segments per class

	Class 1	Class 2	Class 3	Class 4	Total
10 frames	7594	1920	5273	1741	16528
25 frames	2636	622	1978	621	5857
50 frames	987	158	882	253	2280

Table 3: Classification results using cross-validation on the original dataset

	Length	Macro-Avg Precision	Macro-Avg Recall	Accuracy	Macro-Avg F-score	Weighted F-score
RF	10	68.38	56.31	70.09	59.36	68.29
SVM	10	66.9	56.54	69.1	59.38	67.62
RF	25	68.29	57.01	71.02	60.04	69.34
SVM	25	68.4	59.71	71.76	62.50	70.59
RF	50	70.18	56.22	73.81	58.65	71.91
SVM	50	66.34	61.64	73.64	63.50	73.07

4.3 Results

In Tables 3 and 6 we report the macro-average precision, macro-average recall, macro-average, and weighted F-score, as well as accuracy obtained with two validation methods on the original dataset. In Table 7 we report the results for the class-balanced dataset. From Table 3 can be seen that the results in terms of accuracy and F-score are similar for SVM and RF (e.g., Macro F-score: 58.7 − 63.5 and Weighted F-score: 67.6 − 73.1). When comparing the accuracy of SVM with RF for each segment size separately, no significant differences were observed for 50 frames and for 25 frames (measured with Wilcoxon signed rank). Still, the accuracy of 10 frames RF was better than the accuracy of 10 frames SVM ($Z = -2.023$, $p < 0.05$). The difference between Macro and Weighted F-score shows that unrepresented classes obtain a lower F-score. Despite that, the results are above the chance level in all the experiments. As expected, the results with 10 frames segments are, in general, slightly worse than the results for the other 2 segment sizes (25 and 50).

The confusion matrices for RF and SVM for 25 frame segments are presented in Tables 4 and 5. In these two tables, the results for all the testing sets are aggregated. As shown by the tables, better results are obtained for the most numerous classes. The best result is obtained for speaking (i.e., class 1, 84.9% and 82.9%), while the worst one is obtained for smiling (i.e., class 4, 29.8% and 36.1%). Smiling (class 4) and food in-taking (class 2) are very often wrongly classified as speaking (class 1). In several cases, chewing (i.e., class 3) is wrongly classified as speaking (about 21.0% and 18.8%).

Table 4: Confusion Matrix for RF, segments of 25 frames.

	Class 1	Class 2	Class 3	Class 4
Class 1	2239	33	289	75
Class 2	216	232	150	24
Class 3	416	29	1504	29
Class 4	291	38	107	185

Table 5: Confusion Matrix for SVM, segments of 25 frames

	Class 1	Class 2	Class 3	Class 4
Class 1	2186	47	309	94
Class 2	181	265	150	26
Class 3	373	47	1528	30
Class 4	253	31	114	224

Results of leave-one-subject-out validation (see Table 6) are worse in terms of accuracy and F-score. This is not surprising, considering the different number of segments per class of each participant. Results are still above the chance level.

Considering that unbalanced classes might negatively influence the training, we perform another set of experiments using two under-/over-sampling techniques on the training set in each of five folds of the cross-validation (and keeping the original test sets). Here, we balance the global number of samples per class, but we do not balance the number of samples per participant. Results are in the Table 7.

Balancing the train set resulted in a slight improvement of the RF results (but not those of SVM). The best F-score is 64.81, and the best accuracy is 72.14. When comparing the accuracy of different machine learning techniques and balancing approaches, significant differences were observed for all three segments' sizes (measured with Friedman tests): $\chi^2(3) = 8.04$, $p < 0.05$ for 50 frames, $\chi^2(3) = 13.776$, $p < 0.05$ for 25 frames, and $\chi^2(3) = 11.88$, $p < 0.05$ for 10 frames. To sum up, we can say that, although two commensal activities, *speaking* and *chewing*, are recognized quite well using 25 and 50 frames segments, we can still observe some confusion between these two classes. Unfortunately, the remaining two classes obtained much worse results. Even balancing the training set did not help. The substantial differences in F-score between the most and most minor numerous classes suggest that we should extend our dataset. For example, we could ask the same participant pairs to share another online meal.

The worse results for the class 2 segments can be explained by the face occlusions happening during that activity due to the fork moving towards the mouth. Additional hand movement tracking could help improve this class's recognition rate by adopting a multimodal approach in future work. At the same time, many smiles

Table 6: Classification results using leave-one-subject-out validation

	Length	Macro-Avg Precision	Macro-Avg Recall	Accuracy	Macro-Avg F-score	Weighted F-score
RF	10	55.18	46.61	60.07	48.65	58.12
SVM	10	55.88	49.1	62.06	50.8	60.53
RF	25	58.54	50.77	64.2	52.9	62.69
SVM	25	58.9	52.72	65.45	54.6	64.22
RF	50	59.25	50.47	67.62	50.73	65.54
SVM	50	53.75	51.75	65.93	52.12	65.3

Table 7: Classification results using cross-validation on the balanced dataset

Algorithm	Length	Undersampling	M.-Avg Pr.	M.-Avg Rec.	Accuracy	M.-Avg F-score	W. F-score
RF	10	NearMiss	59.39	64.65	65.55	60.78	66.63
RF	10	RandomUnderSample	62.72	65	68.85	63.57	69.18
SVM	10	NearMiss	56.35	60.59	62.72	57.55	63.67
SVM	10	RandomUnderSample	59.05	59.67	65.71	59.32	65.82
RF	25	NearMiss	61.35	66.13	68.14	62.98	68.92
RF	25	RandomUnderSample	63.53	66.87	70.01	64.81	70.49
SVM	25	NearMiss	58.52	62.70	65.44	59.95	66.23
SVM	25	RandomUnderSample	61.43	61.94	68.03	61.66	68.12
RF	50	NearMiss	61.6	64.25	70.3	62.05	70.96
RF	50	RandomUnderSample	62.42	65.27	72.14	63.54	72.61
SVM	50	NearMiss	62.07	63.96	68.02	61.7	69.33
SVM	50	RandomUnderSample	63.4	63.3	71.62	63.27	71.72

are wrongly labeled as speaking and chewing. Indeed, some Action Units appearing in a smile may also occur while chewing. The difference can be in the duration of the activation of this facial activity. We will extend features accordingly in future works.

5 POSSIBLE APPLICATION: ARTIFICIAL COMMENSAL COMPANIONS

In line with the research themes of ICMI 2022, we would like to discuss in more detail one application of the presented research: the creation of Artificial Commensal Companions, i.e., virtual agents or social robots that could enable companionship to humans while eating [29]. We expect that such companions will allow human interaction partners to benefit from the positive effects of commensality, even when physically alone [34].

Artificial agents need to sense (nonverbal) behaviors from, e.g., video data to build natural interaction with human partners. They detect relevant human nonverbal behaviors and react to them accordingly. For example, the SAL character [40] can detect the nonverbal behavior of a human speaker and generate appropriate nonverbal feedback (i.e., backchannels, such as head nods). Interaction with a social agent develops in turns, so the agent needs to detect the most appropriate moment to take the turn to avoid interrupting the human interaction partner while they are speaking. So, Artificial Commensal Companions should be able to detect and appropriately react to significant events in a commensality setting, e.g., to detect if the human partner is chewing food, so they might not be able to respond to the companion's utterance immediately. Consequently, novel computational models are needed, considering these aims; we need models for recognizing the activities that most commonly appear in commensality settings.

While some first attempts at Artificial Commensal Companions were already made (e.g., [17, 26, 42]), their sensing and interaction skills are still quite limited. For example, in [17] the companion can track the user's hand position to understand when the human is taking food. Recognizing a larger set of commensal activities is essential for developing more complex interactions with Commensal Companions.

6 CONCLUSION AND FUTURE WORK

In this paper, we focused on commensal activities analysis and recognition. The manual annotation of 9 pairs remotely eating was used to 1) analyze the social interaction in a remote commensality setting and 2) to develop a baseline recognition model.

Regarding the first point, we showed that such a type of interaction could be smooth and entertaining for the commensal partners. While it was not possible (due to COVID19 restrictions) to collect the data of the same participants in the presence and compare them to the ones collected online, we noticed several positive aspects of remote commensality. Participants maintain gaze contact and talk to each other most of the time during the meal; they display a relatively high number of smiles/laughter. At the same time, significant individual differences were observed between the participants in terms of eating-related activities.

Moving to the second point, we used the manual annotation to train models for automatic recognition of commensal activity. We showed that it is possible to classify four activity types, namely chewing, speaking, food in-taking, and smiling, above the chance level by processing face low-quality video data of naturalistic interactions and applying standard video processing methods and machine learning. To the best of our knowledge, this is the first attempt to classify these four commensal activities using video data only. Differently from previous works that focused on recognizing either eating-only activity (i.e., chewing) or social signals (i.e., smiling, laughing) only, we consider these activities together, providing a first attempt to classify them in a real-life commensality setting. We applied our approach to more than 3 hours of data and 18 subjects.

It is essential to notice that our aim was not to obtain the highest classification results, as this is an ongoing work. We acknowledge that more advanced machine learning techniques need to be explored in the future. The results reported are just a starting point toward systems able to recognize commensal activities. A critical aspect of commensality interaction is that two or more activities often overlap (e.g., speaking and chewing, or speaking and smiling). The current model does not address these more complex cases. We consider using Multi-Label classification for this purpose after extending the dataset.

Moreover, other or more fine-grained commensal activities need to be introduced. For instance, distinguishing between different types of laughter and smile (e.g., hilarious Vs. feedback) in the commensality setting can be of high importance. Last but not least, in this work, we focus on facial expressions only, but we will add upper body movements and gaze.

The dataset used in this work has some limitations. First, the number of participants is relatively small, and the labels are imbalanced. We are currently working on extending the number of recordings and annotations that will share the same protocol. Second, the dataset's recording conditions were relatively stable: for example, the participants were instructed about where to place their laptops/webcams. So, in the future, the dataset could be extended by considering less constrained setups in which, e.g., interaction partners are present in the same physical space (public areas, canteens, restaurants).

We are convinced that commensal activity recognition models may have several applications. Apart from creating Artificial Commensal Companions (see the previous section), we aim to use them to study the dynamics of the human-human interaction in commensality settings. For instance, such models could be used to improve and tune the "hand-made" analyses presented in this paper. Also, they will allow us to build models of social interactions, e.g., to quantify the relationship between interaction partners automatically or to classify the type of the commensality event(s) (e.g., business lunch vs. romantic dinner). Moreover, the quantitative measures in traditional commensality research are often based, e.g., on food in-taking or similar metrics. Their manual annotation is very time-consuming, and thus we could replace it with automatic systems. Other applications go beyond commensality settings. Apps designed to improve eating habits (e.g., technology-based interventions for fast-eaters, see, e.g., [24]), assistive technology for people with motor/physical disabilities [10, 37] and technology aiming to enhance the gastronomic experience (e.g., by introducing the multi-sensory food interaction [43]) would also benefit from the automatic recognition of commensal activities. We hope this work may boost the research in commensal activity recognition and thus enhance our knowledge about this universal human experience. For this reason, the annotation and the AUs extracted from the videos will be made available for research purposes after the paper publication.

REFERENCES

[1] 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.

[2] Oliver Amft and Gerhard Tröster. 2008. Recognition of dietary activity events using on-body sensors. *Artificial intelligence in medicine* 42, 2 (2008), 121–136.

[3] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. 2016. OpenFace: An open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 1–10.

[4] Rick Bell and Patricia L Pliner. 2003. Time to eat: the relationship between the number of people eating and meal duration in three lunch settings. *Appetite* 41, 2 (2003), 215–218.

[5] Cigdem Beyan, Muhammad Shahid, and Vittorio Murino. 2021. RealVAD: A Real-World Dataset and A Method for Voice Activity Detection by Body Motion Analysis. *IEEE Transactions on Multimedia* 23 (2021), 2071–2085.

[6] S. Bi and D. Kotz. 2021. Eating detection with a head-mounted video camera. *Computer Science Technical Report* TR2021-1002 (2021), 1–15.

[7] 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* (Oldenburg, Germany) *(PervasiveHealth '14)*. ICST, ICST, Brussels, Belgium, Belgium, 174–177.

[8] J. K. Burgoon, N. Magnenat-Thalmann, M. Pantic, and A. Vinciarelli. 2017. *Social Signal Processing*. Cambridge University Press, Cambridge.

[9] Steven Cadavid and Mohamed Abdel-Mottaleb. 2010. Exploiting Visual Quasi-periodicity for Automated Chewing Event Detection Using Active Appearance Models and Support Vector Machines. In *2010 20th International Conference on Pattern Recognition*. 1714–1717.

[10] Alexandre Candeias, Travers Rhodes, Manuel Marques, Jo P. ao Costeira, and Manuela Veloso. 2018. Vision Augmented Robot Feeding. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*. 50–65.

[11] Eleonora Ceccaldi, Gijs Huisman, Gualtiero Volpe, and Maurizio Mancini. 2020. Guess who's coming to dinner? Surveying Digital Commensality During Covid-19 Outbreak. In *Companion Publication of the 2020 International Conference on Multimodal Interaction*. 317–321.

[12] Samuele Centorrino, Elodie Djemai, Astrid Hopfensitz, Manfred Milinski, and Paul Seabright. 2015. Honest signaling in trust interactions: smiles rated as genuine induce trust and signal higher earning opportunities. *Evolution and Human Behavior* 36, 1 (2015), 8–16.

[13] Giada Danesi. 2018. A cross-cultural approach to eating together: Practices of commensality among French, German and Spanish young adults. *Social Science Information* 57, 1 (2018), 99–120.

[14] Hamdi Dibeklioğlu, Albert Ali Salah, and Theo Gevers. 2012. Are You Really Smiling at Me? Spontaneous versus Posed Enjoyment Smiles. In *Computer Vision – ECCV 2012*, Andrew Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato, and Cordelia Schmid (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 525–538.

[15] J. M. Fontana, M. Farooq, and E. Sazonov. 2014. Automatic Ingestion Monitor: A Novel Wearable Device for Monitoring of Ingestive Behavior. *IEEE Transactions on Biomedical Engineering* 61, 6 (June 2014), 1772–1779.

[16] Jayne A. Fulkerson, Nicole Larson, Melissa Horning, and Dianne Neumark-Sztainer. 2014. A Review of Associations Between Family or Shared Meal Frequency and Dietary and Weight Status Outcomes Across the Lifespan. *Journal of Nutrition Education and Behavior* 46, 1 (2014), 2–19.

[17] Conor Patrick Gallagher, Radoslaw Niewiadomski, Merijn Bruijnes, Gijs Huisman, and Maurizio Mancini. 2020. Eating with an Artificial Commensal Companion. In *Companion Publication of the 2020 International Conference on Multimodal Interaction* (Virtual Event, Netherlands) *(ICMI '20 Companion)*. Association for Computing Machinery, New York, NY, USA, 312–316.

[18] Harry J. Griffin, Min. S. Hane Aung, Bernadino Romera-Paredes, Ciaran McLoughlin, Gary McKeown, William Curran, and Nadia Bianchi-Berthouze. 2015. Perception and Automatic Recognition of Laughter from Whole-Body Motion: Continuous and Categorical Perspectives. *IEEE Transactions on Affective Computing* 6, 2 (2015), 165–178.

[19] Amber J. Hammons and Barbara H. Fiese. 2011. Is Frequency of Shared Family Meals Related to the Nutritional Health of Children and Adolescents? *Pediatrics* 127, 6 (06 2011), e1565–e1574.

[20] 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 (05 2016), 1–24.

[21] Marvin A. Hecht and Marianne LaFrance. 1998. License or Obligation to Smile: The Effect of Power and Sex on Amount and Type of Smiling. *Personality and Social Psychology Bulletin* 24, 12 (1998), 1332–1342.

[22] Delwar Hossain, Tonmoy Ghosh, and Edward Sazonov. 2020. Automatic count of bites and chews from videos of eating episodes. *IEEE Access* 8 (2020), 101934–101945.

[23] Hayley Hung and Daniel Gatica-Perez. 2010. Estimating Cohesion in Small Groups Using Audio-Visual Nonverbal Behavior. *IEEE Transactions on Multimedia* 12, 6 (2010), 563–575.

[24] Azusa Kadomura, Cheng-Yuan Li, Yen-Chang Chen, Koji Tsukada, Itiro Siio, and Hao-hua Chu. 2013. Sensing Fork: Eating Behavior Detection Utensil and Mobile Persuasive Game. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems* (Paris, France) *(CHI EA '13)*. Association for Computing Machinery, New York, NY, USA, 1551–1556.

[25] Timo Kaukomaa, Anssi Peräkylä, and Johanna Ruusuvuori. 2013. Turn-opening smiles: Facial expression constructing emotional transition in conversation. *Journal of Pragmatics* 55 (2013), 21–42.

[26] 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* (Glasgow, Scotland Uk) *(CHI EA '19)*. ACM, New York, NY, USA, Article LBW1617, 6 pages.

[27] K. Kiriu, K. Ochiai, A. Inagaki, N. Yamamoto, Y. Fukazawa, M. Kimoto, T. Okimura, Y. Terasawa, T. Maeda, and J. Ota. 2017. Recognizing whether a person is eating alone or has company by using wearable devices. In *2017 Tenth International Conference on Mobile Computing and Ubiquitous Network (ICMU)*. 1–2.

[28] W. C. Mackey. 1976. Parameters of the smile as a social signal. *The Journal of Genetic Psychology* 129, 1 (1976), 125–130.

[29] 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.

[30] Jared Martin, Magdalena Rychlowska, Adrienne Wood, and Paula Niedenthal. 2017. Smiles as Multipurpose Social Signals. *Trends in Cognitive Sciences* 21 (09 2017).

[31] Engin Mendi, Ocal Ozyavuz, Emrah Pekesen, and Coskun Bayrak. 2013. Food intake monitoring system for mobile devices. In *Advances in Sensors and Interfaces (IWASI), 2013 5th IEEE International Workshop on*. IEEE, 31–33.

[32] Christopher Merck, Christina Maher, Mark Mirtchouk, Min Zheng, Yuxiao Huang, and Samantha Kleinberg. 2016. Multimodality Sensing for Eating Recognition. In *Proceedings of the 10th EAI International Conference on Pervasive Computing Technologies for Healthcare* (Cancun, Mexico) *(PervasiveHealth '16)*. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), Brussels, BEL, 130–137.

[33] Philipp Michel and Rana El Kaliouby. 2003. Real Time Facial Expression Recognition in Video Using Support Vector Machines *(ICMI '03)*. Association for Computing Machinery, New York, NY, USA, 258–264.

[34] 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).

[35] 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 (2016), 113–123.

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

[37] Daehyung Park, Yuuna Hoshi, Harshal P. Mahajan, Ho Keun Kim, Zackory Erickson, Wendy A. Rogers, and Charles C. Kemp. 2020. Active robot-assisted feeding with a general-purpose mobile manipulator: Design, evaluation, and lessons learned. *Robotics and Autonomous Systems* 124 (2020), 103344.

[38] Shah Atiqur Rahman, Christopher Merck, Yuxiao Huang, and Samantha Kleinberg. 2015. Unintrusive eating recognition using Google Glass. In *Proceedings of the 9th International Conference on Pervasive Computing Technologies for Healthcare*. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 108–111.

[39] Philipp V. Rouast and Marc T. P. Adam. 2020. Learning Deep Representations for Video-Based Intake Gesture Detection. *IEEE Journal of Biomedical and Health Informatics* 24, 6 (2020), 1727–1737.

[40] Marc Schroder, Elisabetta Bevacqua, Roddy Cowie, Florian Eyben, Hatice Gunes, Dirk Heylen, Mark ter Maat, Gary McKeown, Sathish Pammi, Maja Pantic, Catherine Pelachaud, Bjorn Schuller, Etienne de Sevin, Michel Valstar, and Martin Wollmer. 2012. Building Autonomous Sensitive Artificial Listeners. *IEEE Transactions on Affective Computing* 3, 2 (2012), 165–183.

[41] Charles Spence, Maurizio Mancini, and Gijs Huisman. 2019. Digital commensality: Eating and drinking in the company of technology. *Frontiers in psychology* 10 (2019), 2252.

[42] M. Takahashi, H. Tanaka, H. Yamana, and T. Nakajima. 2017. Virtual Co-Eating: Making Solitary Eating Experience More Enjoyable. In *Entertainment Computing – ICEC 2017*, N. Munekata, I. Kunita, and J. Hoshino (Eds.). Springer International Publishing, Cham, 460–464.

[43] Carlos Velasco, Marianna Obrist, Olivia Petit, and Charles Spence. 2018. Multisensory Technology for Flavor Augmentation: A Mini Review. *Frontiers in Psychology* 9 (2018).

[44] Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. ELAN: a Professional Framework for Multimodality Research. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. European Language Resources Association (ELRA), Genoa, Italy, 1556–1559.

[45] Vasoontara Yiengprugsawan, Cathy Banwell, Wakako Takeda, Jane Dixon, Samang Seubsman, and Adrian C Sleigh. 2015. Health, happiness and eating together: what can a large Thai cohort study tell us? *Global journal of health science* 7, 4 (2015), 270.