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

Kheder Yazgi, Cigdem Beyan, Maurizio Mancini, Radoslaw Niewiadomski

UNIVERSITÀ DI TRENTO

UNIVERSITÀ di VERONA

SAPIENZA UNIVERSITÀ DI ROMA

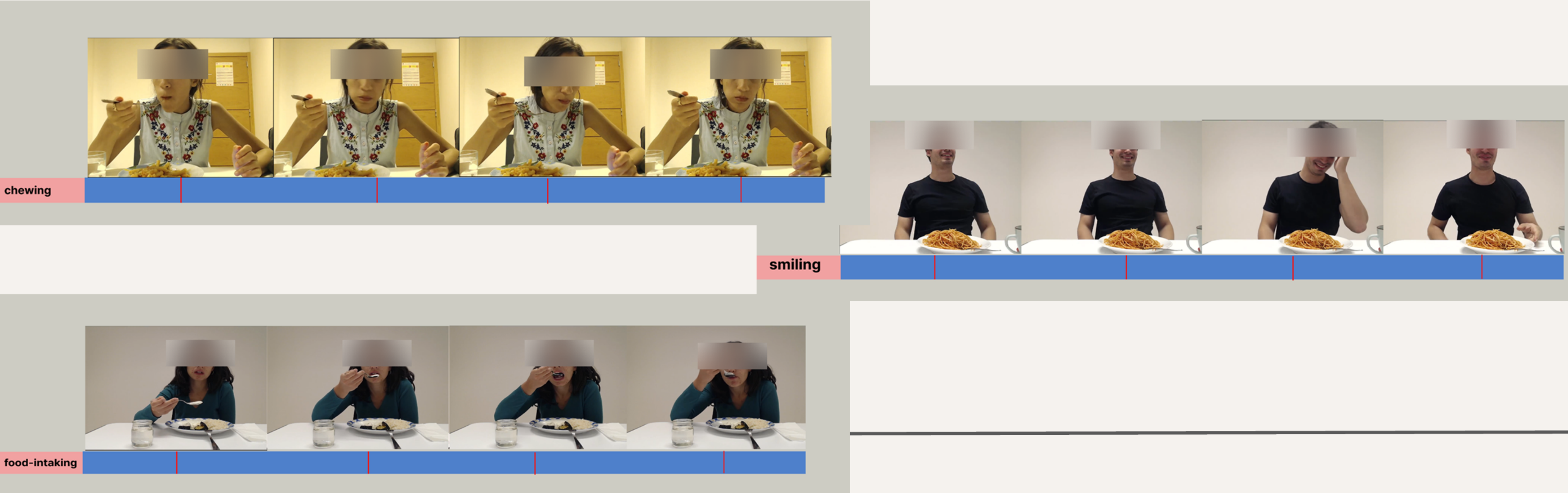
Università di Genova

COCOA

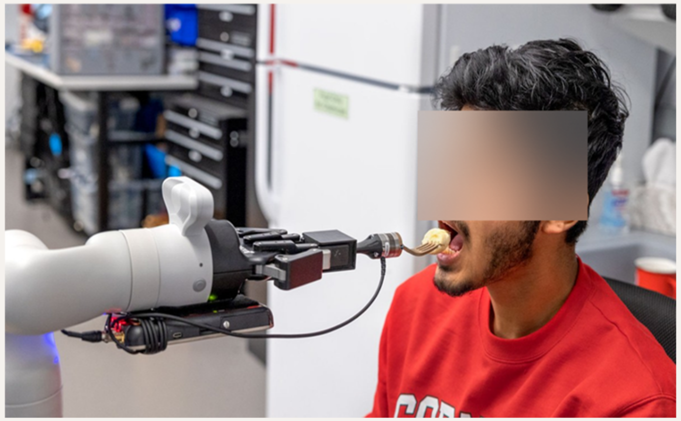
COCOA-PROJECT.GITHUB.IO

Commensal Activities

- actions that often performed during shared meals
- they include:
  - actions related to food consumption, e.g., food intake, chewing, drinking, passing the plate ..
  - social signals displayed by the eaters: gaze, smiles…



Related works

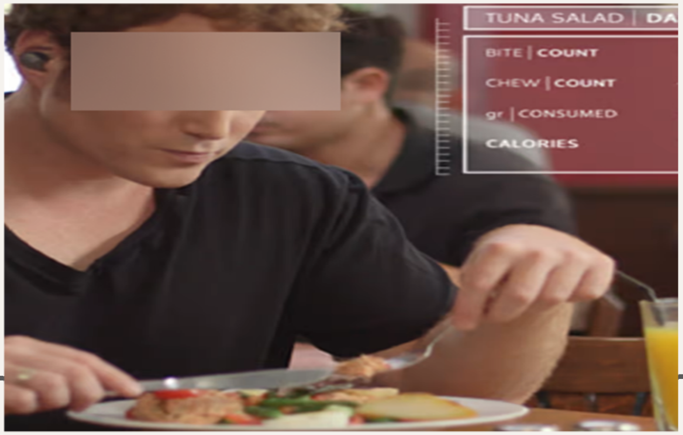


Physical assistance

e.g., robot aims to aid individuals with physical disabilities in overcoming obstacles related to their disabilities while eating

Artificial Commensal Companion

is a social robot or virtual agent able to maintain social interaction with human partner during food consumption



Health and well-being

e.g., chewing tracker counting the number of chews or the duration of chewing

Images taken from https://news.cornell.edu/stories/2022/01/robot-assisted-feeding-focus-15m-nsf-grant

Shortcoming: datasets for (computational) commensality

- No audiovisual dataset has been collected in-person from commensal activities targeting the food/eating related actions and social signals.
- Existing in-person datasets target eating actions only primarily.

food-intaking

chewing

speaking

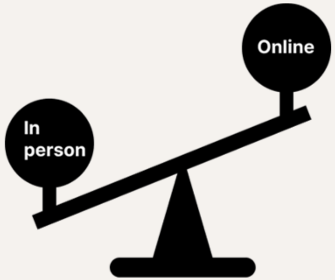
smiling

Objectives

1. Collect and annotate a new audio-visual dataset from in-person commensal events

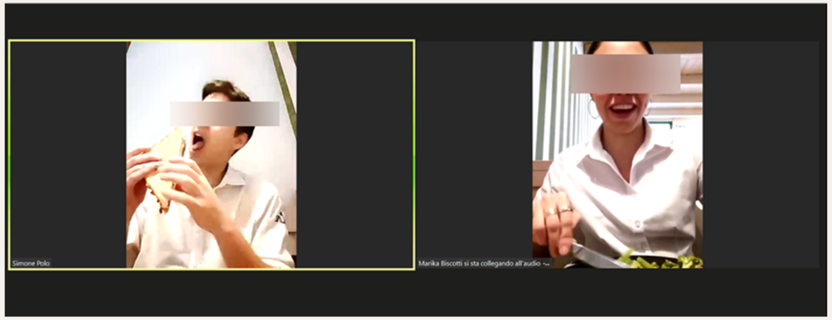
2. Perform baseline machine learning experiments on the new dataset to demonstrate recognition feasibility

3. Investigate whether the performance of models varies across different dataset types (online and in-person)

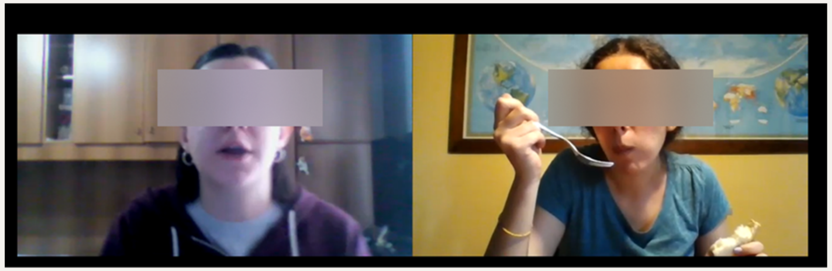


Existing datasets (online setting)

- Existing audiovisual datasets (D1 - D2) contain recordings of ONLINE commensal activities
- Each of them has twenty-two individuals (11 recordings)
- Almost all participants are Italian, gender distribution balanced
- All videos are annotated for speaking, smiling, chewing, and food intaking
- Recording takes in free-living conditions (e.g., at home)
- Most of the pairs knew each other well



Dataset D2: 95 minutes and 20 seconds



Dataset D1: 95 minutes and 57 seconds

Niewiadomski, R., De Lucia, G., Grazzi, G., & Mancini, M. (2022, November). Towards Commensal Activities Recognition. In *Proceedings of the 2022 International Conference on Multimodal Interaction* (pp. 549-557). https://doi.org/10.1145/3536221.3556566

New data collection

Examples of synchronized video-frames from two recordings

- 12 recording sessions were conducted
- 22 participants, avg. age 24 years, 8 females
- Eight Italian
- Eating in pairs, selection was based solely on the availability of time
- In five recording sessions, individuals meet each other for the first time
- A total of 234 minutes were recorded

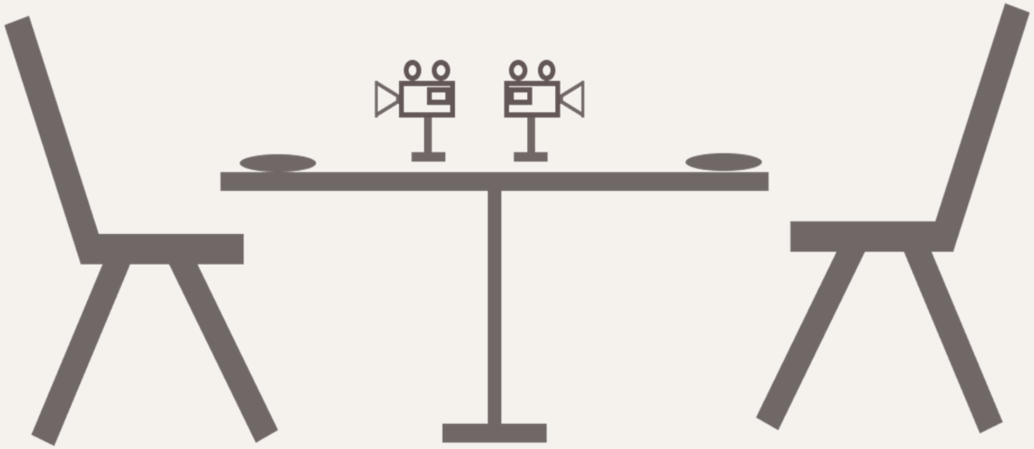
Setting

- Participants facing each other at a table
- One camera positioned in front of each participant

- Controlled environment:
  - 3x3 square meter room in a student dormitory
  - participants consume the similar food, primarily pasta

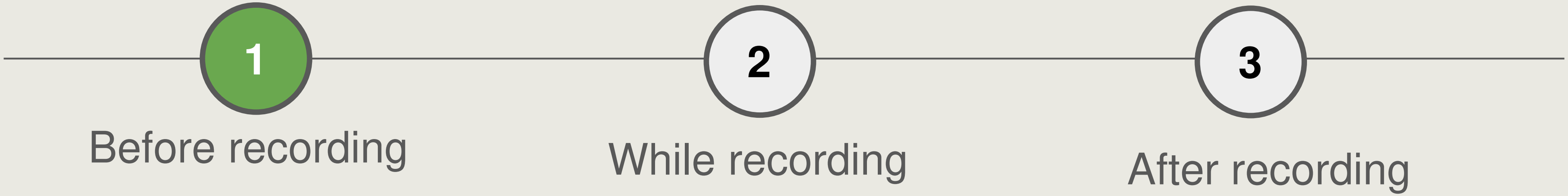
- Videos:
  - recorded two Logitech cameras
  - OBS Studio software,
  - resolution of 970 x 710, a frame rate of 25

- The output is synchronized view of 2 participants



The setting used during data collection

Procedure



1 Before recording

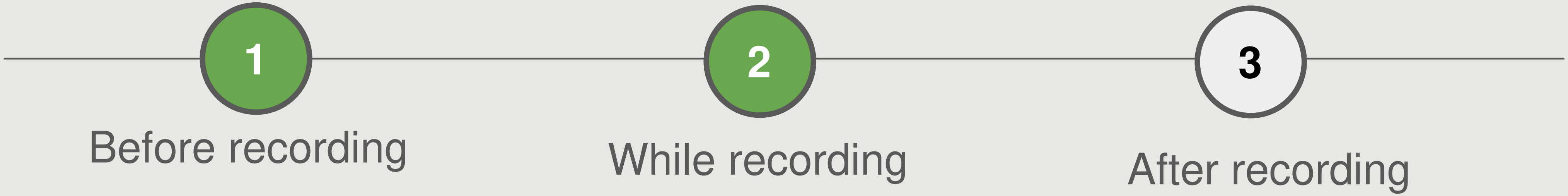
2 While recording

3 After recording

A set of questionnaires is used to assess:

- the general attitude toward eating together (ad-hoc questionnaire)
- the strength of the relationship with the commensal companion (two standard questionnaires)
- affective state and attitude toward the companion immediately before the commensal experience (ad-hoc questionnaire)

Procedure



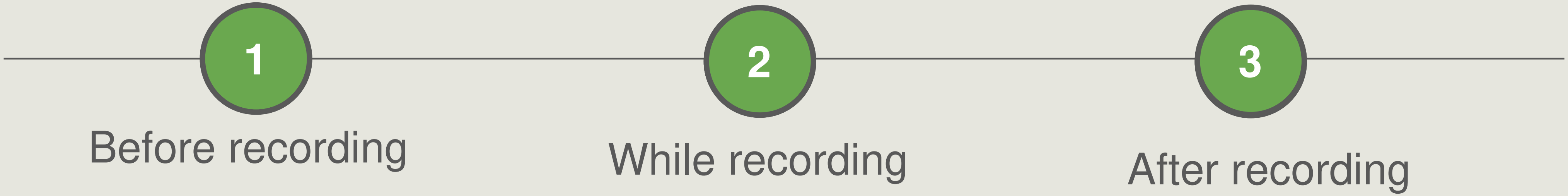
1 Before recording

2 While recording

3 After recording

- Participants remain alone in the room
- They can speak any language they want
- Typically the session is about 10 minutes

Procedure



1

Before recording

2

While recording

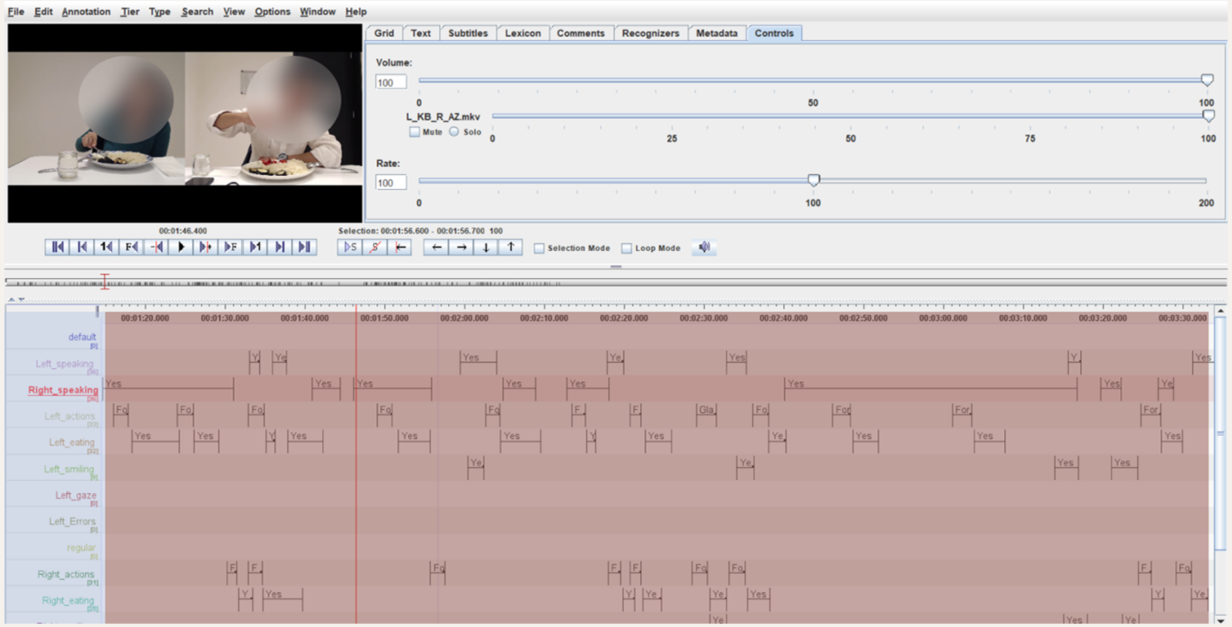
3

After recording

Following the recording, participants were given a post-recording questionnaire to evalue:

- affective state and attitude toward the companion immediately after the commensal experience (ad-hoc questionnaire)
- satisfaction with the meal (ad-hoc questionnaire)

Data annotation



Volume: 100

Rate: 100

00:01:46.400

Selection: 00:01:56.600 - 00:01:56.700 100

- Data annotation done with Elan software

- Four behaviors annotated in the videos: speaking, chewing, food intake, and smiling

- One annotator

Aguera, P. E., Jerbi, K., Caclin, A., & Bertrand, O. (2011). ELAN: a software package for analysis and visualization of MEG, EEG, and LFP signals. *Computational intelligence and neuroscience*, *2011*, 1-11. https://doi.org/10.1155/2011/158970

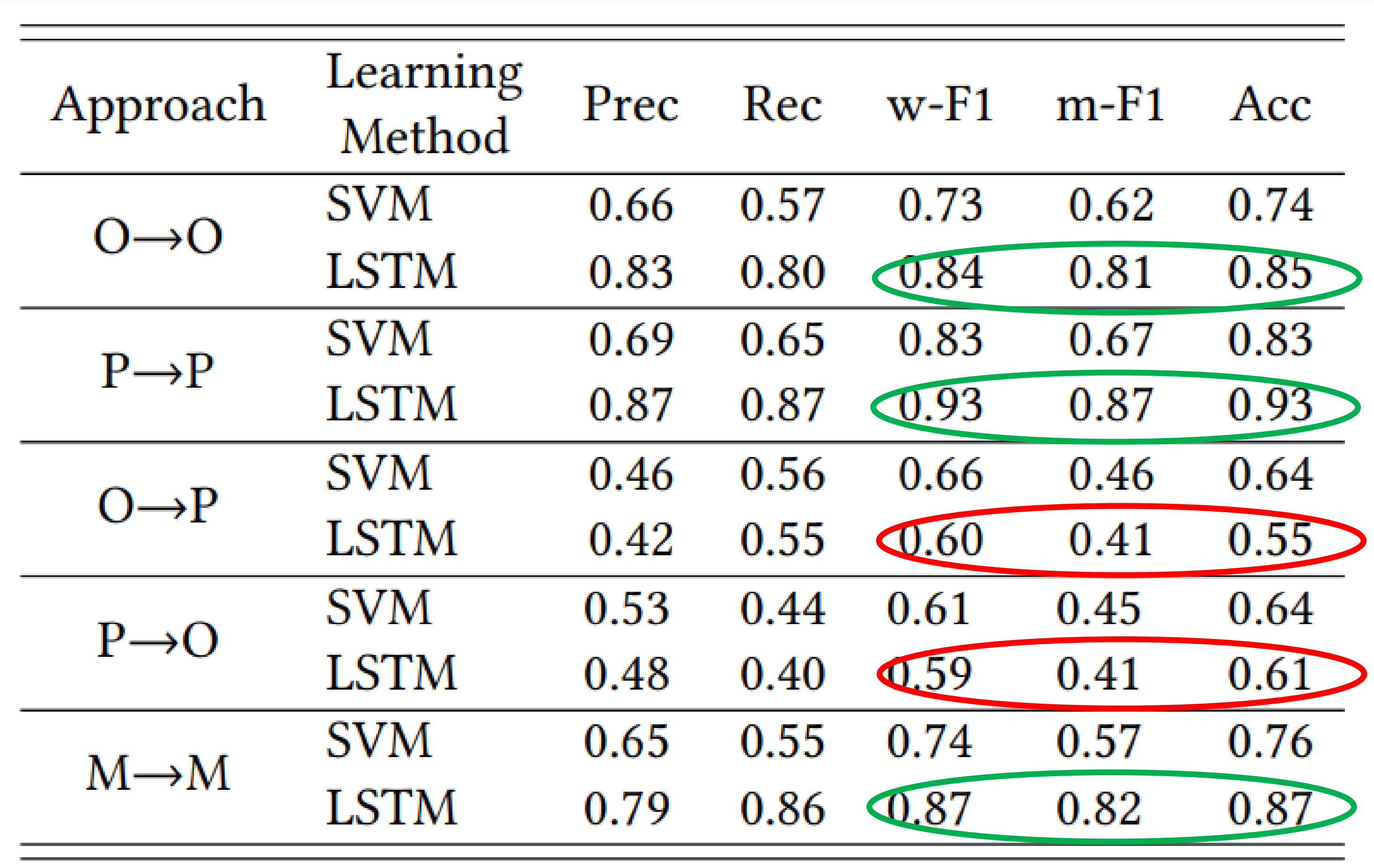
Models

- Two approaches: (a) Support Vector Machines and (b) Long Short-Term Memory
  - SVM used in previous work
  - LSTM for modeling temporal evolution of the signal

- The features are 17 AUs extracted using OpenFace 2.0
  - SVM: averages computed on the 50 frames segments
  - LSTM: 50 frames segments are used directly

- 5-fold cross-validation
- Grid search to optimize hyperparameters for both SVM and LSTM
- Evaluations using two approaches: within-dataset and cross-dataset

Experiments



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

- LSTM shows better performance than SVM even on small datasets
- performance is slightly better on in-presence dataset (even if there is less data)
- model struggles to generalize across different contexts

O – online, P – in-presence, M – merged datasets

Limitations

Lack of diversity in the dataset

Handling the overlapping actions

Missing essential comensal activities like gaze, drinking

The current models are off-line

Manual annotation performed by one person

Conclusions

- Baseline models for food-related actions and social signal classification:
  - good results for 4-class classification problem with LSTM and data collected in one setting
  - models trained on small datasets from one specific setting (online, in-person), can not generalize over data collected from different setting

- Future works:
  - data collection in public venues, such as restaurants or cafés,
  - sessions with more than two individuals,
  - considering body movements and vocal cues.

Thank you for your attention

The Potato Eaters by *Vincent van Gogh*

Let's discuss:

More data is needed:

- shared initiative to collect the data in different parts of the world
- shared data collection protocols
- data privacy, abstract representation of the data

Finanziato dall'Unione europea NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA

COCOA

This work supported by the PRIN2022 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 (MIUR)