Towards Commensal Activities Recognition

R. Niewiadomski, G. De Lucia, G. Grazzi, M. Mancini
contact: r.niewiadomski@unitn.it

UNIVERSITÀ DI TRENTO
SAPIENZA UNIVERSITÀ DI ROMA

More about **computational commensality** and **artificial commensal companions**

Commensality

the practice of sharing food and eating together in a social group — *Ochs and Shohet, 2006*

- important social activity
- one of the most frequent and common human experiences
- time for celebration, making business, and creation of new social bounds
- food has a social and emotional meaning
- several impacts of commensality were observed

Aim

We aim to show the **feasibility** of using machine learning to **recognize a variety of commensal activities** from video data collected in a naturalistic setting

Interaction around the table is peculiar:
- partners shift their attention between the conversation and the food

Commensal activities include:
- actions related to food consumption (e.g., chewing, intaking, drinking)
- social signals (e.g., smiling, gazing, passing the food)

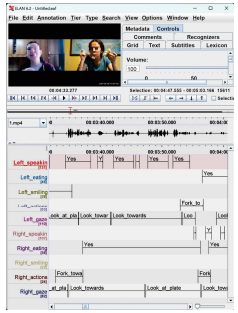
The previous works focus on a single activity recognition and/or different contexts

Data collection

Recordings of dyads eating in a video call:
- participants know each other well
- they eat at their home
- closeup framing including face and upper body
- 18 subjects

Videos:
- synchronized view
- 96 minutes of synchronized recordings
- average meal duration 9m 23s

Annotation

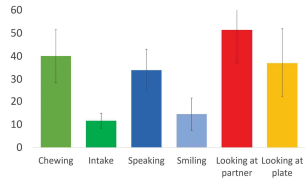


Five activities annotated:
- speaking
- food/drink in-taking
- chewing
- smiling/laughing
- gaze

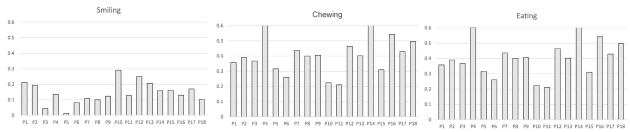
One expert:
- annotated all videos
- used audio and video for annotation
- annotated each activity independently

Different activities may be performed simultaneously

Annotation results

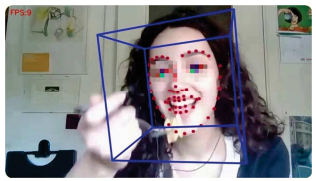


- classes are imbalanced
- social activities are frequent
- interactions are smooth and rich
- nine social metrics computed, e.g.:
  - avg. overlapping speech is 1.6%
  - avg. duration of mutual gaze is 27.85%
- differences between participants and dyads



Feature extraction

- 17 Action Units extracted with OpenFace



- three segment lengths: 10, 25, 50 frames
- only data segments with a single label
- six statistics computed for each AU: min, max, mean, standard deviation, skewness, kurtosis

	Length	Accuracy	F-score	Weigh. F-score
RF	10	70.09	59.36	68.29
SVM	10	69.1	59.38	67.62
RF	25	71.02	60.04	69.34
SVM	25	71.76	62.50	70.59
RF	50	73.81	58.65	71.91
SVM	50	**73.64**	**63.50**	**73.07**

Results using cross-validation

	Length	Accuracy	F-score	Weigh. F-score
RF	10	60.07	48.65	58.12
SVM	10	62.06	50.8	60.53
RF	25	64.2	52.9	62.69
SVM	25	65.45	**54.6**	64.22
RF	50	**67.62**	50.73	**65.54**
SVM	50	65.93	52.12	65.3

Results for leave-one-subject-out

	Length	Accuracy	F-score	Weigh. F-score
RF	10	68.85	63.57	69.18
SVM	10	65.71	59.32	65.82
Rf	25	70.01	**64.81**	70.49
SVM	25	68.03	61.66	68.12
RF	50	**72.14**	63.54	**72.61**
SVM	50	71.62	63.27	71.72

Results on balanced dataset

Classification

- best accuracy 73.6% (SVM, 50 frames)
- better results for longer segments
- differences between SVM and RF are minor
- accuracy drops with leave-one-subject-out to 67.6%
- when balancing the dataset, the results are better only for RF (4.6% on average)

	Speaking	Intaking	Chewing	Smiling	Total
10 frames	7594	1920	5273	1741	16528
25 frames	2636	622	1978	621	5857
50 frames	987	158	882	253	2280

Discussion and next steps

The first attempt to classify commensal activities from a video:
- standard video processing and machine learning techniques
- low-resolution videos of naturalistic interactions

Future extensions:
- other modalities, e.g., hands, gaze
- temporal features
- multi-label classification
- fine-grained activity recognition

References:
Ochs, E. and Shohet, M. (2006). The cultural structuring of mealtime socialization. New directions for child and adolescent development 2006, 111, 35–49.
ELAN software by Max Planck Institute for Psycholinguistics. The Language Archive, Nijmegen, The Netherlands (https://archive.mpi.nl/tla/elan); Sloetjes, H. & Wittenburg, P. (2008). Annotation by category - ELAN and ISO DCR. In: Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008).
Baltrušaitis, T., Zadeh, A., Chong Lim, Y., & Morency, L.P. (2018). OpenFace 2.0: Facial Behavior Analysis Toolkit. IEEE International Conference on Automatic Face and Gesture Recognition