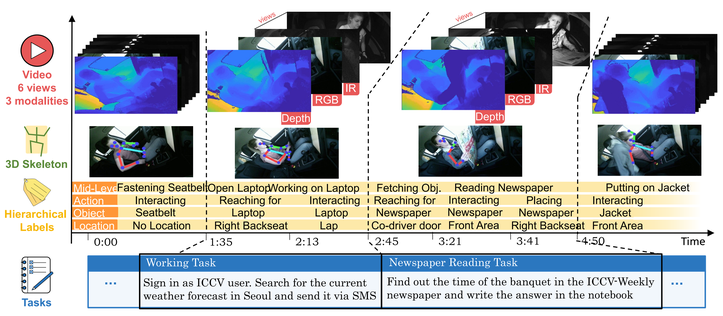
Drive&Act: A Multi-modal Dataset for Fine-grained Driver Behavior Recognition in Autonomous Vehicles

Manuel Martin*, Alina Roitberg*, Monica Haurilet, Matthias Horne, Simon Reiß, Michael Voit, Rainer Stiefelhagen

www.driveandact.com *equal contribution, alphabetical order

Fine-Grained Driver Behaviour Recognition



Motivation

- Looking at humans *inside* the cabin crucial for human-vehicle communication, dynamic driving adaptation and safety
- Lack of large-scale public datasets for driver-activity recognition

Main Contribution

- First large-scale dataset for fine-grained driver behavior recognition in context of manual and autonomous driving
- Twelve hours (>9.6 Mio. frames) annotated and publicly available

Drive&Act in Numbers and Comparison to Previous Datasets

	SoA conven. AR	Multi-mod. AR	Driver Activity Recognition Datsets						
	Kinetics [7]	NTU [43]	HEH [36]	Ohn et al. [35]	Brain4Cars [19]	D.P.-Night [50]	D.P.-Real [50]	AUC-D.D. [2]	Drive&Act
Year	2017	2016	2014	2014	2015	2016	2016	2017/18	2019
Publicly available	✓	✓	✓	–	✓	–	–	✓	✓
Manual driving	–	–	✓	✓	✓	✓	✓	✓	✓
Autonomous driving	–	–	–	–	–	–	–	–	✓
RGB/Grayscale	✓	✓	✓	✓	✓	✓	✓	✓	✓
Depth	–	✓	✓	N/A[b]	–	–	–	–	✓
NIR	–	✓	–	–	–	✓	–	–	✓
Skeleton	–	✓	–	–	–	–	–	–	✓
Video	✓	✓	✓	N/A[b]	✓	✓	✓	N/A[b]	✓
N° images	>76M	4M	N/A[b]	11K	2M	29K	18K	17K	> 9.6M
N° synch. views	1	3	1	2	2	1	1	1	6
Resolution	N/A[c]	1920×1080[a]	680×480	N/A[b]	1920×1088	640×480	640×480	1920×1080	1280×1024[a]
N° subjects	N/A[b]	40	8	4	10	20	5	31	15
Female / male	N/A[b]	N/A[b]	1 / 7	1 / 3	N/A[b]	10 / 10	N/A[b]	9 / 22	4 / 11
N° Classes	400	60	19	3	5	4	4	10	83
Multi-level annot.	–	–	–	–	–	–	–	–	✓
N° Levels	1	1	1	1	1	1	1	1	3
Continuous labels	–	–	–	N/A[b]	–	✓	✓	N/A[b]	✓
Object annot.	✓	–	–	–	–	–	–	–	✓

[a] RGB resolution, IR/Depth resolution is 512×424
[b] information not provided by the authors
[c] variable resolution
[d] NIR-camera resolution

Manuel Martin
Fraunhofer IOSB
Interactive Analysis and Diagnosis (IAD)
Phone +49 721 6091-630
manuel.martin@iosb.fraunhofer.de

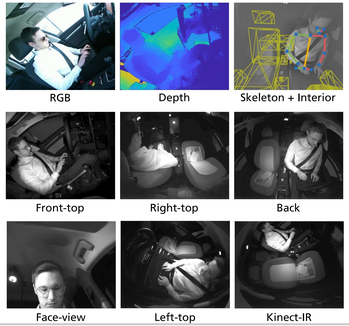
Alina Rolitberg
Karlsruhe Institute of Technology (KIT)
Computer Vision for HCI Lab (cv:hci)
Phone +49 721 608-44735
alina.roitberg@kit.edu

Drive&Act – Key Properties

- Activities during both, **autonomous** and **manual** driving (83 classes in total at all levels of abstraction)
- **Multi-modality**: color-, depth-, infrared- and body pose
- **Multi-view**: six synchronized and calibrated camera views
- **Hierarchical** activity labels on three levels of abstraction
- **Fine-grained** actions (e.g. *opening bottle* and *closing bottle*)
- **Diversity of action duration/complexity** (e.g. *opening door from inside* – seconds; *reading a magazine* – minutes).

Data Collection: Environment

- Recorded in a static driving simulator
- 5 NIR cameras with active IR Illumination and band pass filter
- 1 Kinect for xBox One
- Tasks presented in random order on a touch screen
- 3D Body Pose by triangulation of OpenPose results

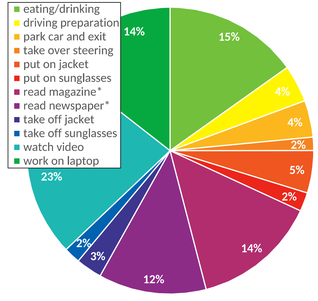


Data Collection: Annotation

Hierarchical Annotation Scheme:

- **Scenarios / Tasks**: 12 high-level tasks completed each session (e.g. read a magazine to answer a question)
- **Fine-grained Activities**: alternate freely i.e. the driver is not told how to execute the task (34 in total)
- **Atomic Action Units**: interactions with the environment: ***Action*** (5), ***Object*** (17), ***Location*** (14) triplets

Distribution of Scenarios/Tasks



Sample Frequency of Fine-grained Activities (Left) and Atomic Actions (Right) by Class

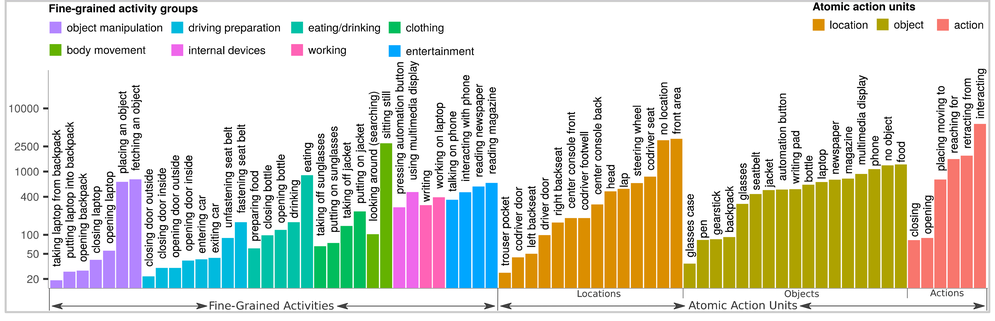
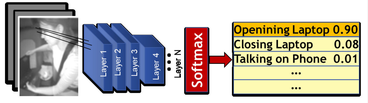


Image-based Recognition

Three CNN-based action recognition architectures trained end-to-end:

- **C3D** model [Tran et al., 2015]
- **Inflated 3D CNN** (I3D) [Carreira et al., 2017]
- **Pseudo 3D ConvNet** (P3D) [Qiu et al., 2017]
- Multi-Modal Recognition: **late fusion**



Pose-based Recognition

Based on mid-level representation:

- Body Pose
- Static car interior

LSTM-based independent streams with weighted late fusion:

- **Pose**: Body pose vector over time
- **Interior**: Pose-Interior distances
- **Structure**: Body structure using depth first walk of kinematic model

Results

Scenarios/Tasks

Type	Model	Validation	Test
Baseline	Random	8.33	8.33
Pose	Interior	35.76	29.75
	Pose	37.18	32.96
	Two-Stream	39.37	34.81
	Three-Stream	41.70	35.45
End-to-end	I3D Net	44.66	31.80

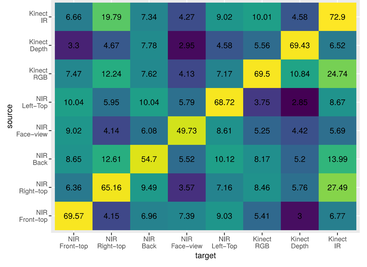
Fine-Grained Activities

Type	Model	Validation	Test
Baseline	Random	2.94	2.94
Pose	Interior	45.23	40.30
	Pose	53.17	44.36
	Two-Stream	53.76	45.39
	Three-Stream	55.67	46.95
End-to-end	C3D	49.54	43.41
	P3D ResNet	55.04	45.32
	I3D Net	69.57	63.64

Atomic Action Units

Model	Action		Object		Location		All	
	val	test	val	test	val	test	val	test
Random	16.67	16.67	5.88	5.88	7.14	7.14	0.39	0.31
Pose	57.62	47.74	51.45	41.72	53.31	52.64	9.18	7.07
Interior	54.23	49.03	49.90	40.73	53.76	53.33	8.76	6.85
Two-Stream	57.86	48.83	52.72	42.79	53.99	54.73	10.31	7.11
Three-Stream	59.29	50.65	55.59	45.25	59.54	56.5	11.57	8.09
I3D Net	62.81	56.07	61.81	56.15	47.70	51.12	15.56	12.12

Cross View Evaluation (Fine-Grained Act.)



Fine-Grained Activities on different Views

Camera	View	Validation	Test
NIR Cameras	front top	69.57	63.64
	right top	65.16	60.80
	back	54.70	54.34
	face view	49.73	42.98
	left top	68.72	62.83
	combined	72.70	67.17
Kinect Color	right top	69.50	62.95
Kinect Depth		69.43	60.52
Kinect IR		72.90	64.98
Combined		73.80	68.51
All combined (score averaging)		74.85	69.03