Hochschule
Bonn-Rhein-Sieg
University of Applied Sciences

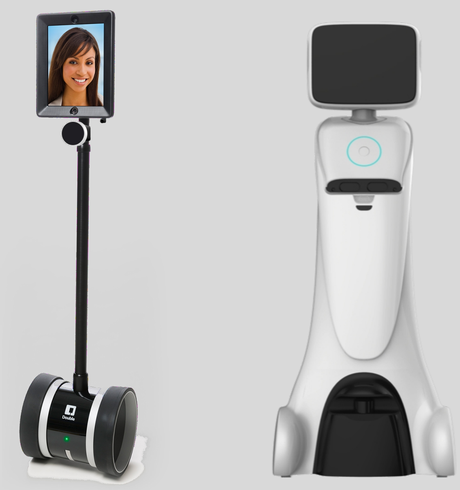
Weakly-supervised learning for daily-living action recognition

- **Motivation:**
  Enabling robots to understand and recognize human actions in domestic environments from processing video data.

  Enabling elderly people to live in their own homes longer by providing social, rather than physical robotic interaction. Examples: Intelligent reminding, cognitive stimulation and mobile video-telephony. [1]

- **The RoboLand project:**
  Telepresence (=mobile robotic video-telephony) for increasing social interactions of elderly people in rural environments of Germany.

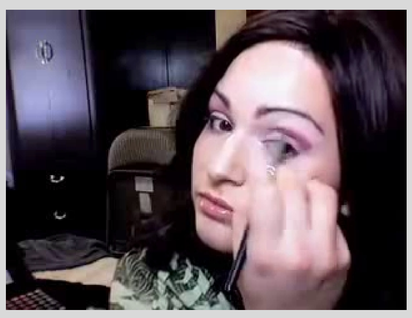
  Used telepresence robots:
  Double *(left)* and Amy A1 *(right)*.



- **Action Recognition:**
  Input: A video-clip with a person performing an action.
  Desired output: Name of that action (a.k.a the “label”).

  Neural Networks can be trained to map video-clips to action labels by processing a large amount of example training-data.

  The Training-dataset determines the action classes that can be recognized. More labelled training data means better performance!

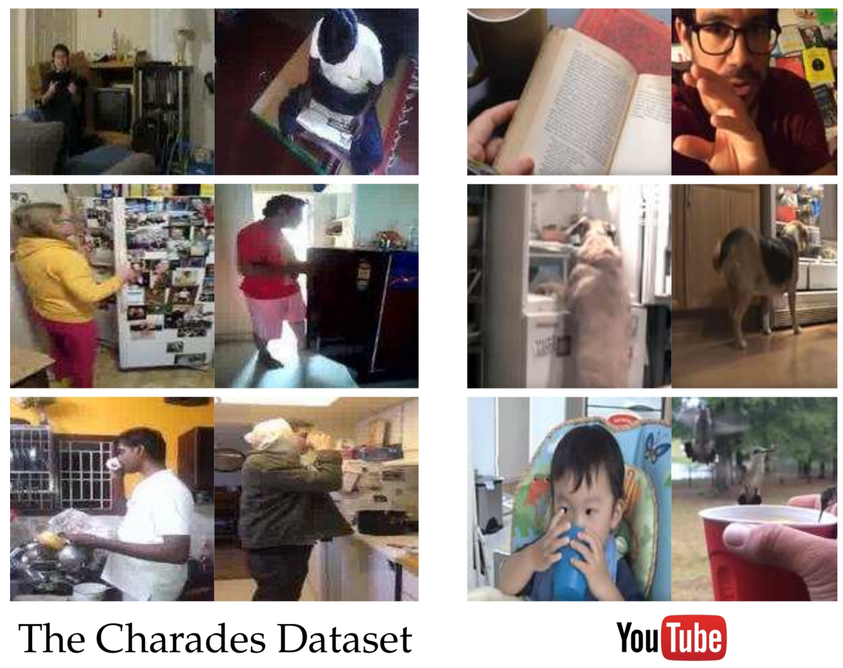


Action Classes:
BabyCrawling
Archery
ApplyEyeMakeUp
CliffDiving
GolfSwing
Fencing

- **Open Problems:**
  Video datasets for training neural networks are magnitudes smaller than their image counterparts.

  Daily-living actions datasets are even smaller, since these kinds of videos are “boring” and hard to obtain on a big scale.

  Current neural network approaches (most prominently Convolutional Neural Networks) “reason” about approx. half a second of input video only. Extension to longer temporal extends is needed. [2,3]



The Charades Dataset YouTube

The figure above shows sample frames from videos of a daily-living action dataset (Charades [4]) and videos sampled from YouTube.

YouTube videos are highly biased towards entertaining scenarios and settings.

- **Goals of this PhD:**
  Releasing a proper daily-living action recognition dataset by extending and fixing Charades.

  Training neural networks with multiple datasets (transfer learning, unsupervised learning, multi-task learning). [5,6,7]

  Adapting neural architectures for longer temporal inputs (using recurrent neural networks like LSTMs and GRUs).

References:
[1] David Feil-Seifer and Maja J. Mataric. “Defining Socially Assistive Robotics”. In: Rehabilitation Robotics, 2005. ICORR 2005. 9th International Conference On. IEEE, 2005.
[2] Jeffrey Donahue et al. “Long-Term Recurrent Convolutional Networks for Visual Recognition and Description”. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015.
[3] Gül Varol, Ivan Laptev, and Cordelia Schmid. “Long-Term Temporal Convolutions for Action Recognition”. In: arXiv preprint arXiv:1604.04494 (2016).
[4] Gunnar A. Sigurdsson et al. “Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding”. In: arXiv preprint arXiv:1604.01753 (2016)
[5] Andrej Karpathy et al. “Large-Scale Video Classification with Convolutional Neural Networks”. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014.
[6] Ishan Misra, C. Lawrence Zitnick, and Martial Hebert. “Shuffle and Learn: Unsupervised Learning Using Temporal Order Verification”. In: (Mar. 28, 2016). arXiv: 1603.08561.
[7] Rich Caruana. “Multitask Learning”. In: Machine learning (1997).
[8] Sepp Hochreiter and Jürgen Schmidhuber. “Long Short-Term Memory”. In: Neural computation (1997).

Hochschule Bonn Rhein-Sieg
Department of Computer Science
Grantham-Allee 20
53757 Sankt Augustin

Maximilian Schöbel
E-Mail: maximilian.schoebel@h-brs.de

Supervised by:
Prof. Dr. Jürgen Gall (University of Bonn)

b-it Bonn-Aachen International Center for Information Technology

Graduate Institute

www.h-brs.de/gi