Deep Affordance-grounded Sensorimotor Object Recognition

Spyridon Thermos,[1,2] Georgios Th. Papadopoulos,[1] Petros Daras,[1] Gerasimos Potamianos[2]

[1] Information Technologies Institute, Centre for Research and Technology Hellas, Greece
[2] Department of Electrical and Computer Engineering, University of Thessaly, Greece

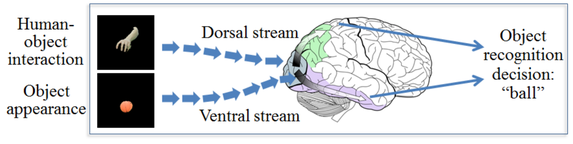
IEEE 2017 Conference on Computer Vision and Pattern Recognition
CVPR July 21-26 2017 HONOLULU

Motivation

- Object recognition constitutes an open research challenge
  - ✓ Almost exclusively use of static appearance features
  - ✓ Challenges: object appearance variance, occlusions, deformations, illumination variation
- Findings in cognitive neuroscience
  - ✓ Object perception is based on the fusion of sensory (object appearance) and motor (human-object interaction) information
  - ✓ Object affrdances: the types of actions that humans typically perform when interacting with them

Proposed approach

- Accumulated evidence on ventral and dorsal stream interaction



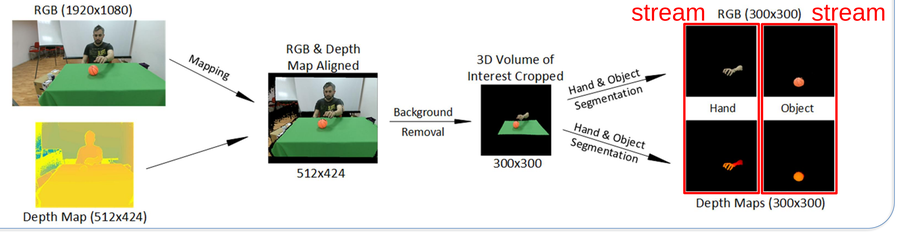
- Novel contributions on sensorimotor object recognition
  - ✓ Neurobiologically and neurophysiologically grounded Neural Network architectures
  - ✓ Evaluation of multiple recent neuro-scientific findings
  - ✓ Large number of complex affordance types
  - ✓ Large-scale public RGB-D object recognition dataset

Introduced SOR3D dataset

- 14 object categories
- 13 affordance types
- 54 object-affordance combinations
- 105 subjects
- Over 20,800 instances
- 3 synchronized Kinect II
- http://sor3d.vcl.iti.gr

Visual front-end

- Scope: Hand and object segmentation
- Output: a) colorized object depth maps, b) colorized hand depth maps, and c) colorized hand 3D flow magnitude fields



Generalized Template Matching (GTM) architecture

- Scope: Estimation of complex multi-level affordance-related patterns along the spatial dimensions
- Based solely on the use of CNNs

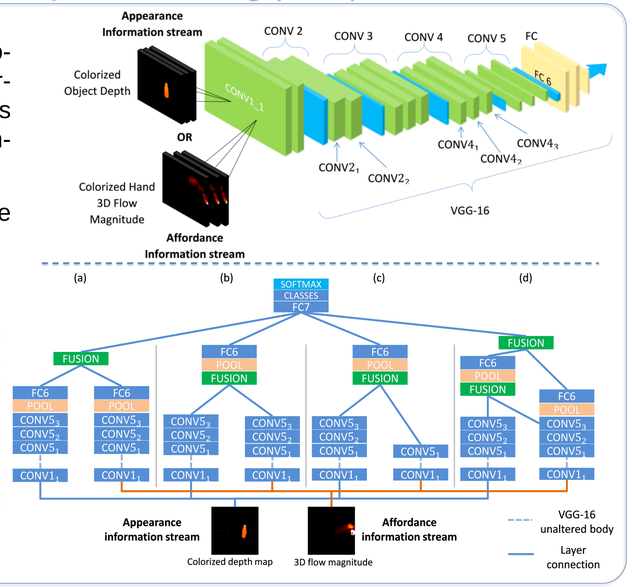


Figure Top: appearance CNN for object recognition and affordance CNN (single-stream model). Bottom: Detailed topology of the GTM architecture for: a) late fusion at FC layer, b) late fusion at last CONV layer, c) single-level slow fusion and d) multi-level slow fusion.

Generalized Spatio-Temporal (GST) architecture

- Scope: Encoding of the time-evolving procedures of the performed human actions
- Composite CNN-LSTM NN considered
- Support of asynchronous fusion

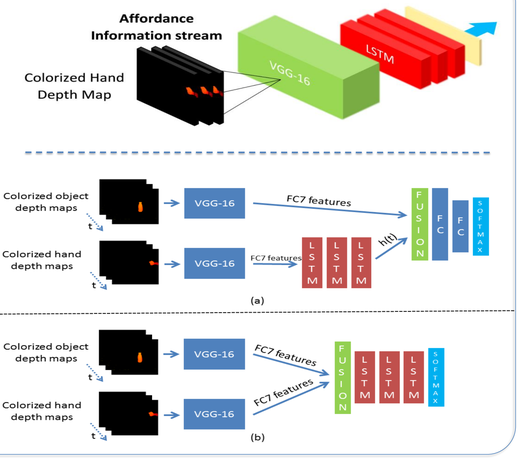


Figure Top: affordance CNN-LSTM (single-stream model). Bottom: Detailed topology of the GST architecture for: a) late fusion and b) slow fusion.

Experimental results

- Up to 29% relative error reduction, compared to the baseline model
- Recognition performance of all supported object types boosted

Fusion architecture	*Accuracy (%)*
Baseline CNN (appearance only)	**85.12**
GTM late	88.24
GTM slow single level	88.13
GTM slow multi-level	**89.43**
GST late	86.50
GST slow single level	79.65
Product Rule	73.45
SVM	83.43
Bayes	75.86

Conclusions & future work

- The proposed NN-based sensorimotor approach outperforms similar probabilistic approaches of the literature.
- Future work will investigate the modeling of the human-object interactions in more details and the application of the proposed methodology to "in the wild" object recognition scenarios.