Deep Learning for Action Recognition in Augmented Reality Assistance Systems

Matthias Schröder
Neuroinformatics Group, Bielefeld University
maschroe@techfak.uni-bielefeld.de

Helge Ritter
Neuroinformatics Group, Bielefeld University
helge@techfak.uni-bielefeld.de

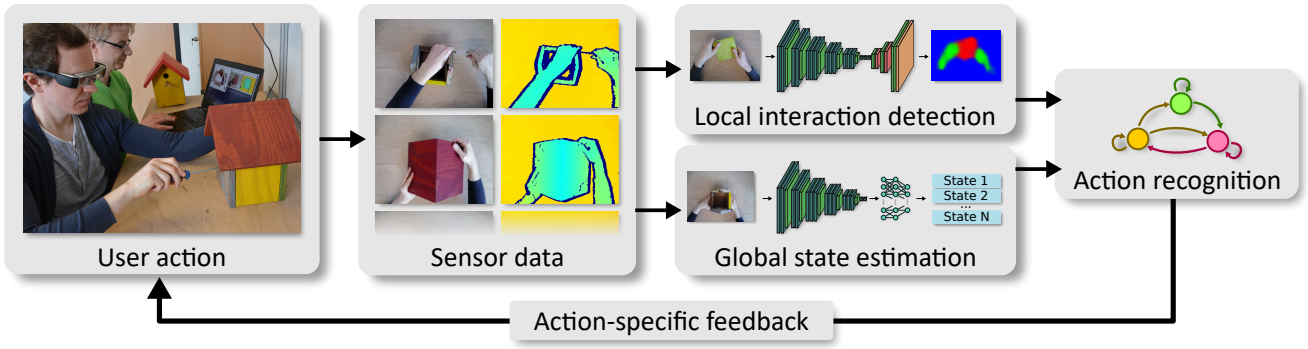

Figure 1: Overview of our system. User actions are captured by an RGBD sensor and recognized using convolutional neural networks (CNNs) and Bayesian inference. Based on this, the user is given action-specific feedback via a head-mounted display.

CCS CONCEPTS

•**Computing methodologies** → **Activity recognition and understanding;** •**Human-centered computing** → *Human computer interaction (HCI); Mixed / augmented reality;*

KEYWORDS

action recognition, deep learning, augmented reality

ACM Reference format:
Matthias Schröder and Helge Ritter. 2017. Deep Learning for Action Recognition in Augmented Reality Assistance Systems. In *Proceedings of SIGGRAPH '17 Posters, Los Angeles, CA, USA, July 30 - August 03, 2017,* 2 pages.
DOI: 10.1145/3102163.3102191

1 INTRODUCTION

Recent advances in the development of optical head-mounted displays (HMDs), such as the Microsoft HoloLens, Google Glass, or Epson Moverio, which overlay visual information directly in the user's field of vision, have opened up new possibilities for augmented reality (AR) applications. We propose a system that uses such an optical HMD to assist the user during goal-oriented activities (e.g. manufacturing work) in an intuitive and unobtrusive way (Essig et al. 2016). To this end, our system observes and recognizes the user's actions and generates context-sensitive feedback. Figure 1 shows an overview of our approach, exemplified with the task of assembling a bird house.

SIGGRAPH '17 Posters, Los Angeles, CA, USA
 978-1-4503-5015-0/17/07…$15.00
DOI: 10.1145/3102163.3102191

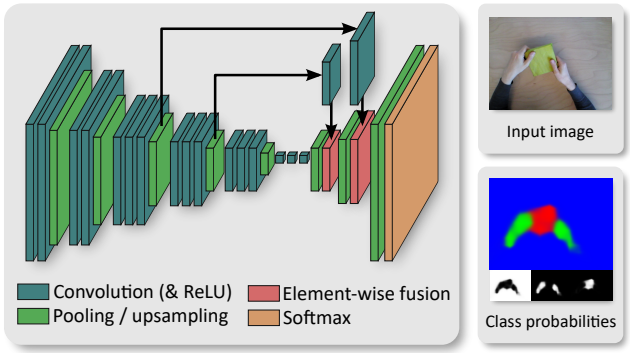

Figure 2: Dense classification of background, hand, and object pixels using a fully convolutional neural network.

User actions captured by a static RGBD sensor are recognized in our system by combining information about *local interactions* and *global progress* extracted from the sensor data. We use deep learning to detect local hand-object interactions and to estimate the overall state in the activity progress. The user's action is recognized by combining this local and global information in a probabilistic state machine using Bayesian inference. Based on this recognition, the user is presented with action- and context-sensitive prompts in real-time, which provide assistance for performing the next step or recovering from errors.

2 LOCAL INTERACTION DETECTION

To accurately localize hand-object interactions, we adapt the approach of (Long et al. 2015) and train a fully convolutional network (FCN) to densely discriminate between hand and object pixels. The FCN (illustrated in Figure 2) uses the VGG-16 architecture

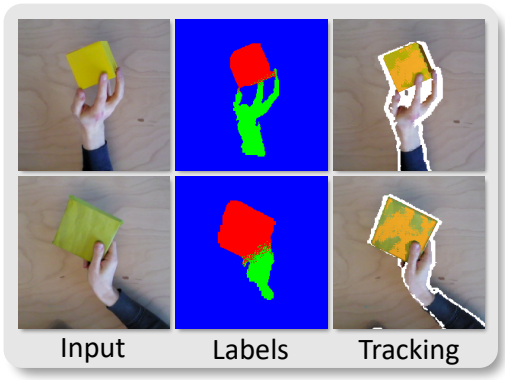

Figure 3: Object tracking examples. Left: input image, center: hand-object labeling, right: model fitted to point cloud.

(Simonyan and Zisserman 2014) as a downsampling *encoder* that extracts features from the input image, which are then incrementally upsampled in the *decoder* to produce class probability maps.

The resulting pixel-wise class predictions are used in combination with depth-based foreground masking to densely label the input data. Based on these labels, we detect hand-object contacts and track the object's rigid transformation while it is being manipulated, following the model-based approach of (Tagliasacchi et al. 2015). Figure 3 illustrates the pixel labeling and object tracking.

3 GLOBAL STATE ESTIMATION

An activity is composed of N discrete states $\mathcal{S} = \{S_1, \ldots, S_N\}$ in our system. To provide the user with context-sensitive feedback, our method estimates which of these states is currently most likely active. This is done using a CNN fine-tuned from VGG-16 (Simonyan and Zisserman 2014), which computes a probability distribution $\mathcal{P} = (p_1, \ldots, p_N)$ over all known states given the current input sensor image. Figure 4 illustrates the state estimation CNN architecture. Figure 5 shows some example images of states involved in the process of assembling a bird house.

4 ACTION RECOGNITION

In order to generate meaningful context-sensitive feedback, the frame-wise local and global information must be contextualized in the overall activity. To this end we use a state machine, which specifies the connections between states $\mathcal{S}$, as well as state transition probabilities $P_{ji} = P(S_i \mid S_j)$. These state transition probabilities are obtained in a data-driven manner using structural-dimensional analysis of mental representations (Essig et al. 2016).

This probabilistic state machine representation is used in conjunction with the output probabilities $\mathcal{P}$ of the global state estimation CNN to robustly recognize and update the current activity state. The prediction for the new state S_i given the current state S_j and the observed sensor data D is computed using Bayesian inference:

$$P(S_i \mid D, S_j) = \frac{P(D \mid S_j, S_i)P(S_i \mid S_j)}{\sum_{k=1}^{N} P(D \mid S_j, S_k)P(S_k \mid S_j)} = \frac{p_i P_{ji}}{\sum_{k=1}^{N} p_k P_{jk}}$$

The state prediction $\operatorname{argmax}_i P(S_i \mid D, S_j)$ and the detected interactions are then used to select a feedback message from a predefined table to assist the user to proceed with the activity.

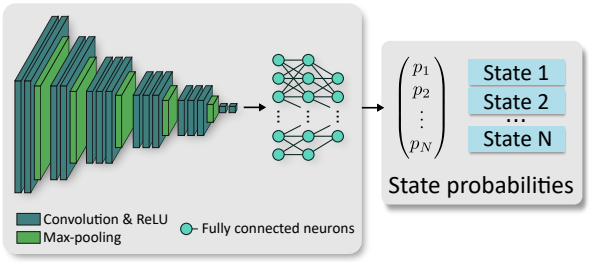

Figure 4: Global activity progress state estimation from an input sensor image using a convolutional neural network.

Figure 5: Examples for states that are distinguished during the global activity progress state estimation.

5 CONCLUSION

Experiments with study participants using our assistance system in a bird house assembly scenario showed generally positive results regarding the usefulness of the context-sensitive AR feedback. Our action recognition approach can be generally applied to goal-oriented tasks involving hand-object interactions, given suitable training data for the deep learning components, and a state-based description of the target activity.

The current implementation of our action recognition system runs in real-time on a laptop with an i7 CPU, 16 GB RAM, and a GTX 1070 GPU. Future work includes optimizing our CNNs for size and efficiency to allow for a completely mobile implementation, as well as the construction of a large-scale database of activities.

ACKNOWLEDGMENTS

This research is supported by the German Federal Ministry of Education and Research (BMBF) project ADAMAAS and the Cluster of Excellence Cognitive Interaction Technology CITEC (EXC 277) at Bielefeld University, which is funded by the German Research Foundation (DFG).

REFERENCES

Kai Essig, Benjamin Strenge, and Thomas Schack. 2016. ADAMAAS: Towards Smart Glasses for Mobile and Personalized Action Assistance. In *International Conference on Pervasive Technologies Related to Assistive Environments*. 46:1–46:4. DOI: http://dx.doi.org/10.1145/2910674.2910727

Jonathan Long, Evan Shelhamer, and Trevor Darrell. 2015. Fully convolutional networks for semantic segmentation. In *Conference on Computer Vision and Pattern Recognition*. 3431–3440. DOI: http://dx.doi.org/10.1109/CVPR.2015.7298965

Karen Simonyan and Andrew Zisserman. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. *CoRR* abs/1409.1556 (2014). http://arxiv.org/abs/1409.1556

Andrea Tagliasacchi, Matthias Schröder, Anastasia Tkach, Sofien Bouaziz, Mario Botsch, and Mark Pauly. 2015. Robust Articulated-ICP for Real-Time Hand Tracking. *Comput. Graph. Forum* 34, 5 (2015), 101–114. DOI: http://dx.doi.org/10.1111/cgf.12700