RUPRECHT-KARLS-UNIVERSITÄT HEIDELBERG

Interdisciplinary Center for Scientific Computing

Computer Vision group heidelberg

LSTM Self-Supervision for Detailed Behavior Analysis

Heidelberg Collaboratory for Image Processing

ETH zürich

Biagio Brattoli*[1] Uta Büchler*[1] Anna-Sophia Wahl[2] Martin E. Schwab[2] Björn Ommer[1]

[1]Heidelberg University, HCI/IWR, Germany {biagio.brattoli, uta.buechler, bjoern.ommer}@iwr.uni-heidelberg.de

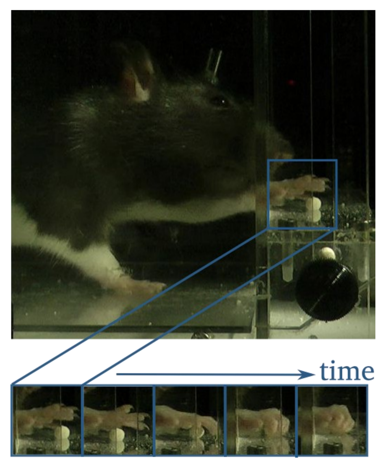
[2]ETH Zurich,Dept. of HST, Switzerland {wahl,schwab}@hifo.uzh.ch

1 Introduction

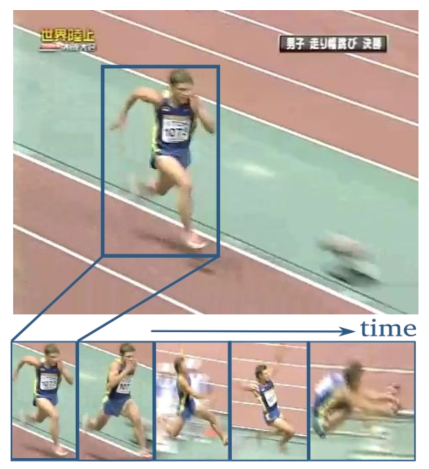
Motivation:

- End-to-End learning of individual postures & overall behavior from unlabelled videos
- Self-supervision by ordering entire sequences
- Learning fine-grained postures indirectly by training on sequences

Grasping



Long jump



Behavior analysis is a crucial, non-invasive diagnostic tool, revealing distinct functional deficits and their restoration during recovery/learning

Typical Approaches ✗

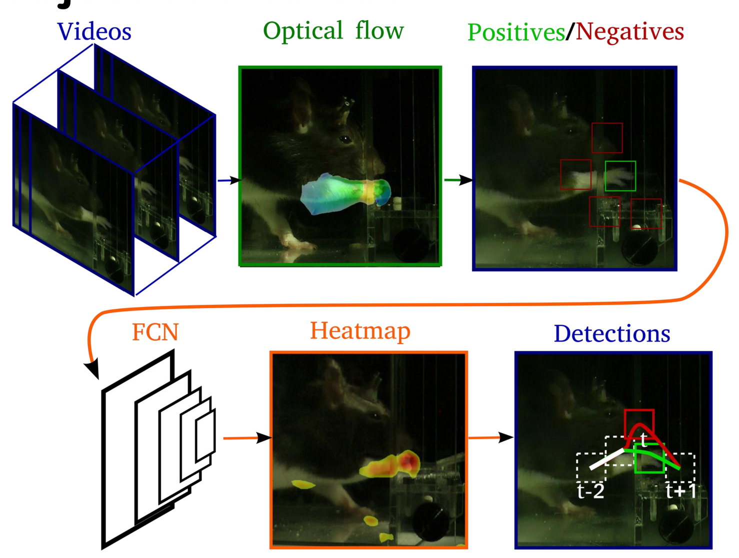
⇨Complex parametric model
⇨Time-intensive labeling
⇨Limited amount of data
⇨Experts required
⇨Subjective interpretation

Our automatic Approach ✓

⇨Model-free
⇨No labeling needed
⇨Scales to arbitrary size
⇨Fully automatic
⇨Objective evaluation

2 Initializing Automatic Detection

⇨ Extract the object of interest



⇨Optical flow is used to collect reliable pos/neg samples
⇨An FCN is trained using these samples
⇨The model extends beyond the original & produces more samples

Models	Accuracy(%)
OpticalFlow	40.2
FCN_0	58.0
FCN_1	81.4
FCN_2	**82.1**

Table 1: Accuracy of Detection

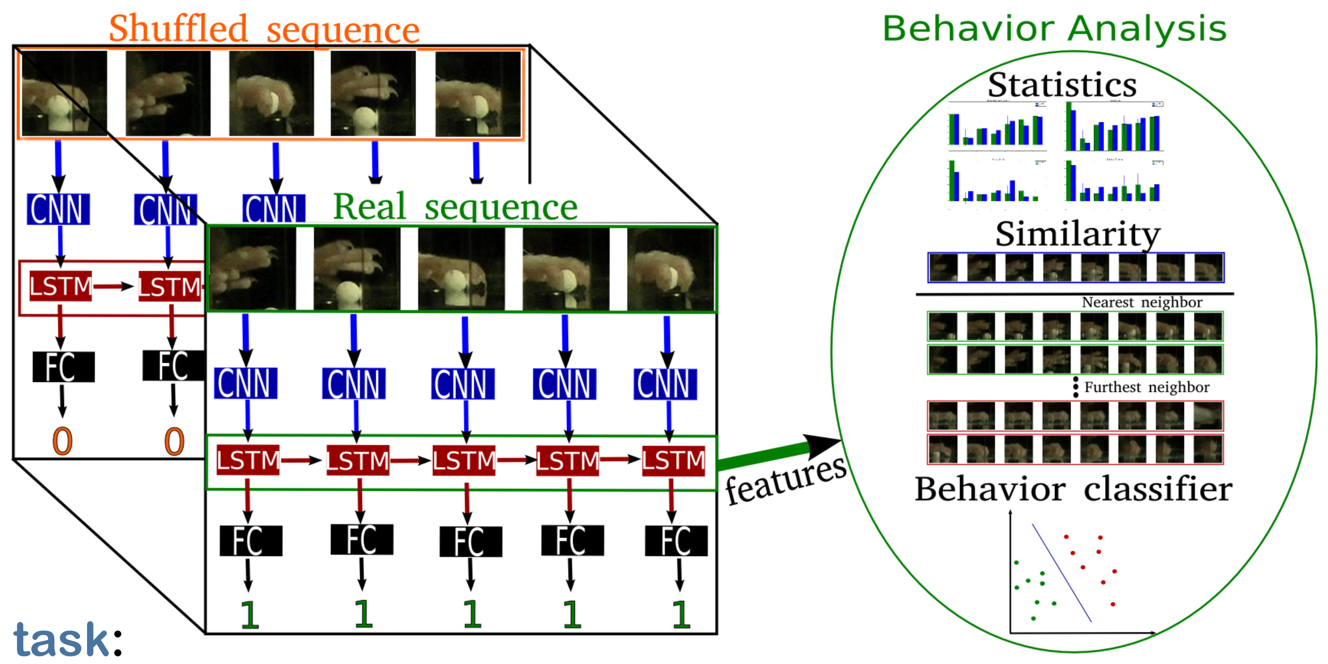
3 Self-supervised Learning

Goal: learn a representation of posture and behavior

⇨A CNN-LSTM network is trained using a surrogate task without labels

- LSTM learns the sequence representation from permuted sequences
- Fine-grained posture details determine behavior & sequence ordering

⇨Back-prop to individual frame improves the posture representation



⇨Self-supervised task:

- Input:real and permuted sequences $s_t = [d_t, d_{t+1}, \cdots, d_{t+l-1}]$ where d_t is the cropped detection of frame t and l the flexible length of sequence s_t
- The Mini-batches D_i are composed as

$$D_i = [s_t^{i_1}, \pi(s_t^{i_1}), s_t^{i_2}, \pi(s_t^{i_2}), \cdots, s_t^{i_n}, \pi(s_t^{i_n})]$$

$$L_i = [y_1, y_2, \cdots, y_{2n}], \text{ with } y_k = k \bmod 2$$

where y_k is the label of sequence k in D_i and $\pi(\cdot)$ a random permutation

⇨Bootstrap retraining: The CNN with improved posture representation is fine-tuned on the initial detection task providing more samples. This, in turn, enhances the learning of posture and behavior.

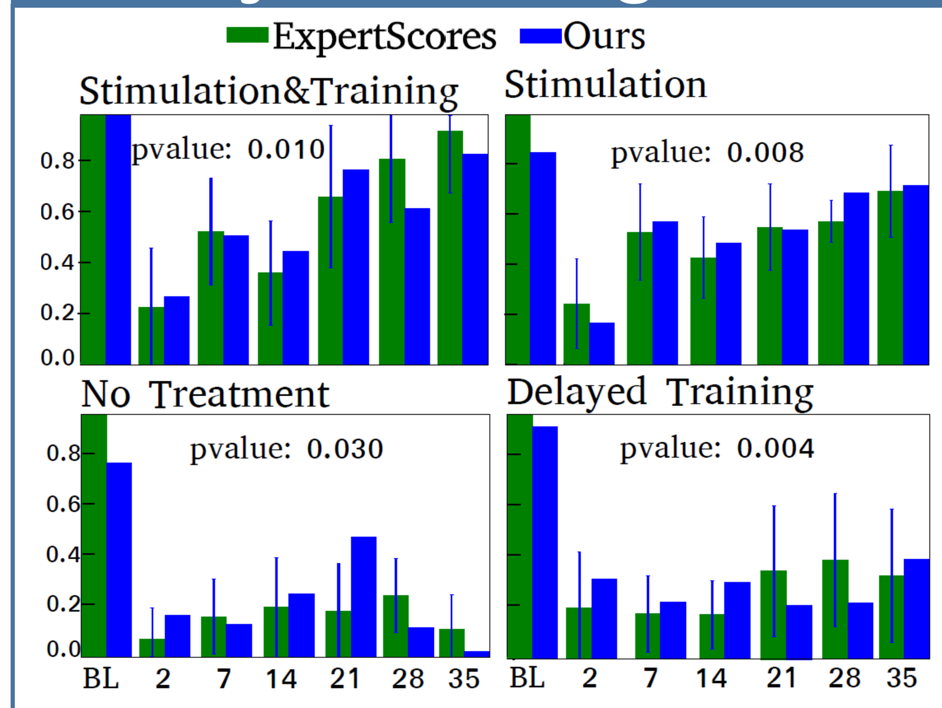
Models	Accuracy(%)
AlexNet	65.3
$PostureCNN_0$	72
$PostureCNN_2$	**85.6**

Table 2: Evaluation of Posture representation

Models	Accuracy(%)
Max frame similarity	74.1
Avg frame similarity	75.9
DTW	76.8
ClusterLSTM	64.0
$CNN\text{-}LSTM_2$	**80.5**

Table 3: Evaluation of sequence representation

4 Analysis of Long-term Behavior Changes

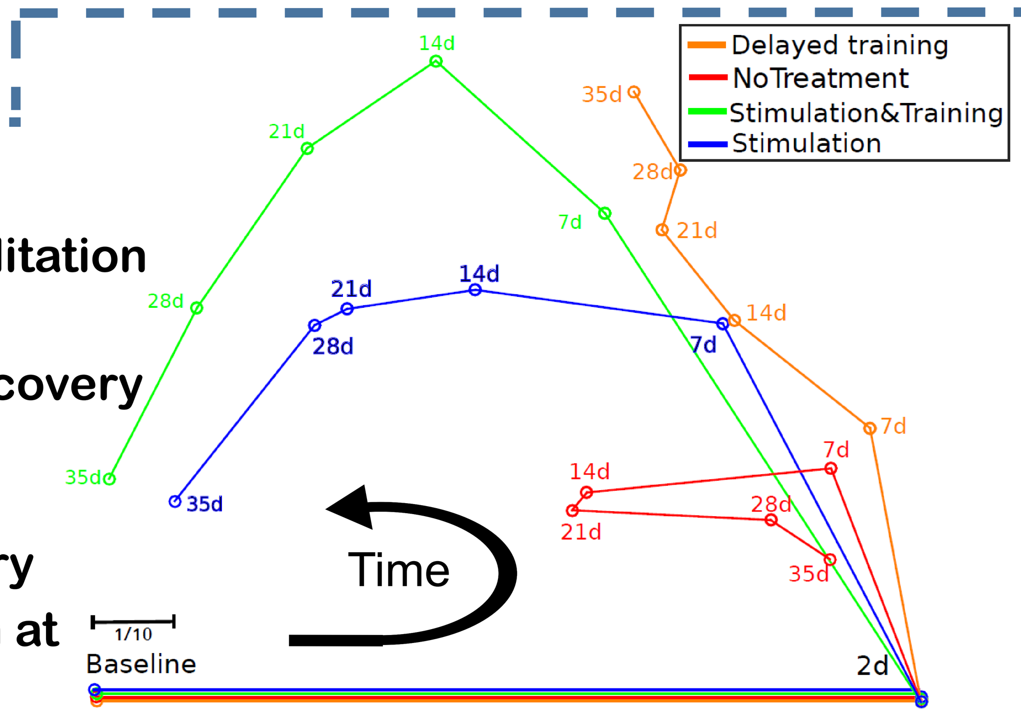


Predicting the fitness of Skilled Motor function

⇨Comparison between our prediction and Expert scores
⇨Experts assess grasping by scoring ten criteria
⇨We circumvent this tedious manual analysis by directly mapping sequences to a final fitness score

Rehabilitation Analysis

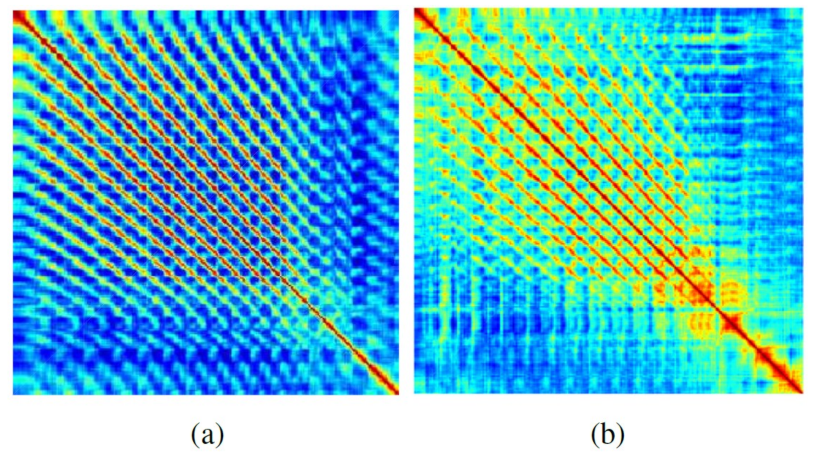
⇨ Long-term recordings during rehabilitation
⇨ Behavior analysis can reveal subtle changes in motor function during recovery
⇨ We relate the behavior to healthy kinematics (Baseline) and impaired samples from 2 days after the surgery
⇨ Provides similarity of motor function at any stage to Baseline and 2 days



5 Non-parametric Human Pose Analysis

Self-supervised Training and Evaluation on Olympic Sports

Averaging the 100NN of a query frame from category Long Jump and Hammer Throw



Similarity Matrices of the category Long Jump using (a) our approach and (b) CliqueCNN

Category	HOG-LDA	Ex. SVM	Ex. CNN	Alex net	Clique CNN	Ours
Mean	0.58	0.67	0.56	0.65	0.79	**0.83**

Table 4: Mean of the average AUC of all categories of the Olympic Sports Dataset

Transferring the Learned Representation to Leeds Sports

Parts	HOG LDA	Alex net	Clique CNN	Ours	Pose Mach	Deep Cut	GT
Mean	38.4	41.1	43.5	**46.6**	67.8	85.0	69.2

Table 5: Mean PCP-measure of the Leeds Sports Dataset