LSTM Self-Supervision for Detailed Behavior Analysis

Uta Büchler[1]*, Biagio Brattoli[1]*, Anna-Sophia Wahl[2], Martin E. Schwab[2], Björn Ommer[1]
[1] HCI / IWR, Heidelberg University, Germany
[2] Department of HST, ETH Zurich, Switzerland
{uta.buechler,biagio.brattoli,bjoern.ommer}@iwr.uni-heidelberg.de,
{wahl,schwab}@hifo.uzh.ch

Abstract

Behavior analysis provides a crucial non-invasive and easily accessible diagnostic tool for biomedical research. A detailed analysis of posture changes during skilled motor tasks can reveal distinct functional deficits and their restoration during recovery. Our specific scenario is based on a neuroscientific study of rodents recovering from a large sensorimotor cortex stroke (the second leading source of disability worldwide) and skilled forelimb grasping is being recorded. Videos of behavior recorded during long-term studies on the recovery after neurological diseases provide an easily available, rich source of information to evaluate and adjust drug application and rehabilitative paradigm. The main bottleneck is presently that all analysis of skilled motor function depends on time-intensive, error-prone, and costly manual evaluation of behavior, e.g. by aggregating a large set of subtle characteristics of limb posture and its deformation over time [1]. Consequently, this detailed behavior representation required for studying skilled motor functions goes far beyond a trajectory analysis [8] which does not suffice to capture impairment of behavior. Thus, there is a dire need for an automatic evaluation of subtle differences in behavior and the underlying postures without costly manual supervision. The only available information for training are videos recorded before and after stroke, where even the healthy animals show a substantial number of failed grasps due to the complexity of the task. Therefore, we utilize self-supervision to automatically learn accurate posture and behavior representations for analyzing motor function. Learning our model depends on the following fundamental elements: *(i)* limb detection based on a fully convolutional network is initialized solely using motion information, *(ii)* a novel self-supervised training of LSTMs using only temporal permutation yields a detailed representation of behavior, and *(iii)* back-propagation of this sequence representation also improves the description of individual postures. Given weak initial candidate detections of grasping paws obtained using motion information [10], a CNN is trained to separate paws from clutter. Unrolling the fully convolutional layers of this model, we obtain a fully convolutional network (FCN [15]) for detecting paws. Moreover, due to the absence of posture annotations we will also utilize this CNN model as an implicitly learned, initial representation of posture. To further improve this representation we move from posture to behavior sequences. Therefore, the CNN for individual postures is directly linked to a recurrent network (LSTM) for behavior, indirectly optimizing the posture representation using the surrogate task of behavior learning through sequence ordering. Although this task of training an LSTM on original sequences against permuted ones sounds more difficult, we can now tap the large amounts of unlabeled videos by self-supervision. Bootstrap retraining then improves detections which in turn enhance the learning of behavior and as a result the individual posture representation, cf. Fig 1. Finally, we use multiple instance learning (MIL)[2, 3] to train a classifier to discover the subtle differences between healthy and impaired grasping behavior. We establish a novel test dataset with expert annotations for

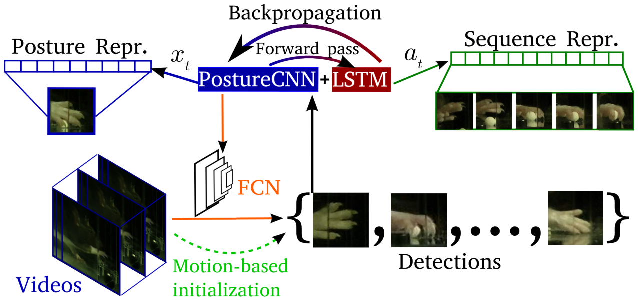

Figure 1: Overview of our self-supervised approach for posture and sequence representation learning using CNN-LSTM. After the initial training with motion-based detections we retrain our model for enhancing the learning of the representations.

[*] Indicates equal contribution

evaluation of fine-grained behavior analysis. Our approach compares favorably against expert manual evaluation of behavior that has been established in neuroscience. We measure the agreement between the manual annotations and our results using the p-value of the two-tailed t-statistic of a linear regression (null hypothesis is that our results does not predict the expert scores). We obtain a mean p-value of 0.01 indicating that the null hypothesis can be safely rejected. Moreover, we demonstrate the generality of our approach by successfully applying it to self-supervised learning of human posture on two standard benchmark datasets, cf. Table 1 and 2.

Category	HOG-LDA [6]	Ex. SVM [11]	Ex. CNN [5]	Alex net [9]	Clique CNN [4]	Ours
Basketball	0.51	0.63	0.58	0.55	0.70	**0.75**
Bowling	0.57	0.63	0.58	0.55	0.85	**0.87**
Clean&Jerk	0.61	0.71	0.58	0.62	0.81	**0.85**
Discus Thr.	0.42	0.76	0.56	0.59	0.65	**0.68**
Diving 10m	0.42	0.54	0.51	0.57	0.70	**0.76**
Diving 3m	0.50	0.57	0.52	0.66	0.76	**0.84**
HammerThr.	0.62	0.64	0.51	0.66	0.82	**0.88**
High Jump	0.64	0.76	0.59	0.62	0.82	**0.87**
Javelin Thr.	0.71	0.72	0.57	0.74	**0.85**	**0.85**
Long Jump	0.60	0.69	0.57	0.71	0.78	**0.85**
Pole Vault	0.59	0.64	0.60	0.64	0.81	**0.83**
Shot Put	0.51	0.67	0.52	0.70	0.75	**0.76**
Snatch	0.64	0.76	0.59	0.67	0.84	**0.89**
TennisServe	0.70	0.75	0.64	0.71	0.84	**0.87**
Triple Jump	0.63	0.65	0.58	0.65	0.80	**0.83**
Vault	0.59	0.71	0.63	0.68	0.81	**0.86**
Mean	0.58	0.67	0.56	0.65	0.79	**0.83**

Table 1: Average AUC of all categories of the Olympic Sports dataset [12] using the state-of-the-art and our approach.

Parts	HOG LDA [6]	Alex net [9]	Clique CNN [4]	Ours	Pose Machines [14]	Deep Cut [13]
Torso	73.7	76.9	80.1	**82.4**	88.1	96.0
U.legs	41.8	47.8	50.1	**53.3**	79.0	91.0
L.legs	39.2	41.8	45.7	**48.0**	73.6	83.5
U.arms	23.2	26.7	27.2	**30.9**	62.8	82.8
L.arms	10.3	11.2	12.6	**16.0**	39.5	71.8
Head	42.2	42.4	45.5	**48.9**	80.4	96.2
Mean	38.4	41.1	43.5	**46.6**	67.8	85.0

Table 2: PCP measure (observer-centric) of the Leeds Sport dataset [7] using our, state-of-the-art and two fully supervised approaches.

References

[1] M. Alaverdashvili and I. Q. Whishaw. A behavioral method for identifying recovery and compensation: hand use in a preclinical stroke model using the single pellet reaching task. *Neuroscience & Biobehavioral Reviews*, 37(5):950–967, 2013. 1

[2] S. Andrews, I. Tsochantaridis, and T. Hofmann. Support vector machines for multiple-instance learning. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*, pages 577–584. MIT Press, 2003. 1

[3] B. Antic and B. Ommer. Robust multiple-instance learning with superbags. In *ACCV*. Springer, Springer, 2012. 1

[4] M. A. Bautista, A. Sanakoyeu, E. Sutter, and B. Ommer. Cliquecnn: Deep unsupervised exemplar learning. *NIPS*, 2016. 2

[5] A. Dosovitskiy, J. T. Springenberg, M. Riedmiller, and T. Brox. Discriminative unsupervised feature learning with convolutional neural networks. In *Advances in Neural Information Processing Systems 27*, pages 766–774. Curran Associates, Inc., 2014. 2

[6] B. Hariharan, J. Malik, and D. Ramanan. *Discriminative Decorrelation for Clustering and Classification*, pages 459–472. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012. 2

[7] S. Johnson and M. Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *Proceedings of the British Machine Vision Conference*, 2010. doi:10.5244/C.24.12. 2

[8] M. Kabra, A. A. Robie, M. Rivera-Alba, S. Branson, and K. Branson. Jaaba: interactive machine learning for automatic annotation of animal behavior. *nature methods*, 10(1):64–67, 2013. 1

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012. 2

[10] C. Liu. *Beyond pixels: exploring new representations and applications for motion analysis*. PhD thesis, Citeseer, 2009. 1

[11] T. Malisiewicz, A. Gupta, and A. A. Efros. Ensemble of exemplar-svms for object detection and beyond. In *ICCV*, 2011. 2

[12] J. C. Niebles, C.-W. Chen, and L. Fei-Fei. *Modeling Temporal Structure of Decomposable Motion Segments for Activity Classification*, pages 392–405. Springer Berlin Heidelberg, 2010. 2

[13] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. V. Gehler, and B. Schiele. Deepcut: Joint subset partition and labeling for multi person pose estimation. In *CVPR*, pages 4929–4937, 2016. 2

[14] V. Ramakrishna, D. Munoz, M. Hebert, A. J. Bagnell, and Y. Sheikh. Pose machines: Articulated pose estimation via inference machines. In *ECCV*, 2014. 2

[15] E. Shelhamer, J. Long, and T. Darrell. Fully convolutional networks for semantic segmentation. 2016. 1