An Efficient Convolutional Network for Human Pose Estimation

Umer Rafi[1]
rafi@vision.rwth-aachen.de
Ilya Kostrikov[1]
ilya.kostrikov@rwth-aachen.de
Juergen Gall[2]
gall@iai.uni-bonn.de
Bastian Leibe[1]
leibe@vision.rwth-aachen.de

[1] Computer Vision Group, RWTH Aachen University, Germany

[2] Computer Vision Group, University of Bonn, Germany

In recent years, human pose estimation has greatly benefited from deep learning and huge gains in performance have been achieved on popular benchmarks [1, 3, 4]. The trend to maximise the accuracy on benchmarks, however, resulted in computationally expensive deep network architectures that require expensive hardware and pre-training on large datasets. In this work, we propose an efficient deep network architecture that can be efficiently trained on mid-range GPUs without the need of any pre-training and that is on par with much more complex models on the benchmarks [1, 3, 4].

Our proposed Fully Convolutional GoogLeNet (FCGN) network (see Figure 1) is based on the network architecture from [2]. We take the first 17 layers of [2] and add a deconvolution layer to make it fully convolutional. In addition, we introduce a skip layer and combine two FCGNs with shared weights to obtain a multi-resolution network. Belief maps for each joint are then obtained by a deconvolution layer with large kernel size in combination with a sigmoid function for normalisation and spatial drop out for regularisation.

We compare the performance of the proposed architecture against convolutional pose machines [5] on the well-known FLIC, LSP, and MPII benchmarks [1, 3, 4]. Our proposed network outperforms most previous approaches and achieves competitive performance to the more complex model of [5], while requiring only 3GB of memory and far less training time.

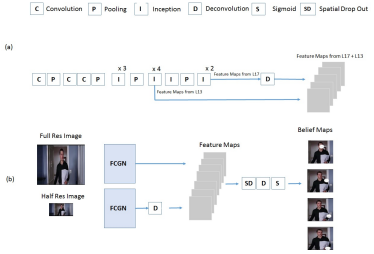

Figure 1: (a) Proposed fully convolutional GoogLeNet (FCGN) (b) The proposed multi-resolution network combines two FCGNs.

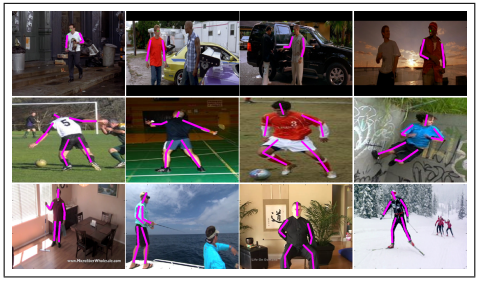

Figure 2: Our Qualitative results on FLIC [4], LSP [3] and MPII [1].

[1] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2D Human Pose Estimation: New BenchMark. In *CVPR*, 2014.

[2] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. 2015.

[3] S. Johnson and M. Everingham. Clustered Pose and Nonlinear Appearance Models for Human Pose Estimation. In *BMVC*, 2010.

[4] B. Sapp and B. Taskar. MODEC : Multimodel Decomposable Models for Human Pose Estimation. In *CVPR*, 2013.

[5] S. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional Pose Machines. In *CVPR*, 2016.