Efficient Object Localization Using Convolutional Networks

Jonathan Tompson, Ross Goroshin, Arjun Jain, Yann LeCun, Christoph Bregler
New York University
tompson/goroshin/ajain/lecun/bregler@cims.nyu.edu

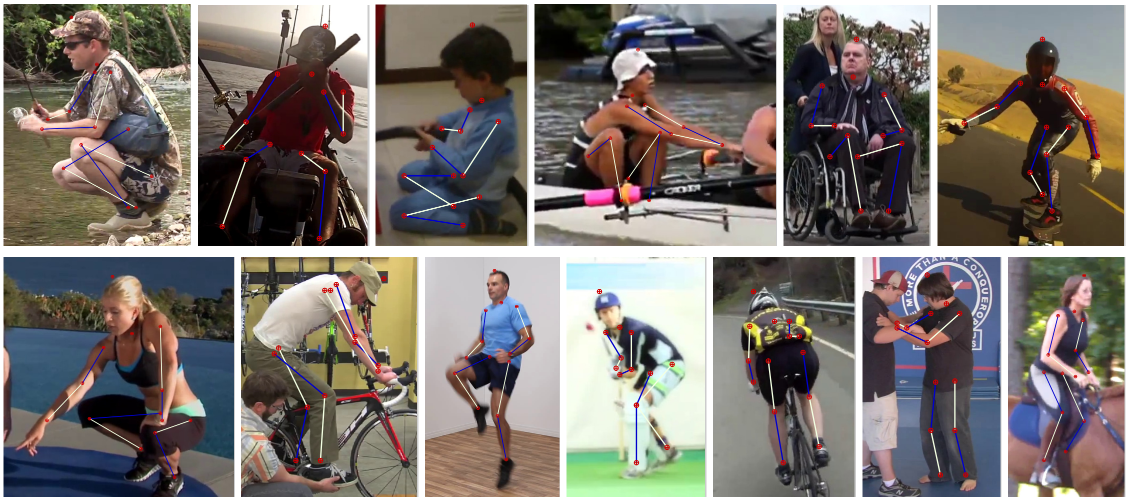

Figure 1: Our Model‘s Predicted Joint Positions on the MPII-human-pose database test-set[1]

Abstract

Recent state-of-the-art performance on human-body pose estimation has been achieved with Deep Convolutional Networks (ConvNets). Traditional ConvNet architectures include pooling and sub-sampling layers which reduce computational requirements, introduce invariance and prevent over-training. These benefits of pooling come at the cost of reduced localization accuracy. We introduce a novel architecture which includes an efficient 'position refinement' model that is trained to estimate the joint offset location within a small region of the image. This refinement model is jointly trained in cascade with a state-of-the-art ConvNet model [21] to achieve improved accuracy in human joint location estimation. We show that the variance of our detector approaches the variance of human annotations on the FLIC [20] dataset and outperforms all existing approaches on the MPII-human-pose dataset [1].

1. Introduction

State-of-the-art performance on the task of human-body part localization has made significant progress in recent years. This has been in part due to the success of Deep-Learning architectures - specifically Convolutional Networks (ConvNets) [21, 14, 22, 5] - but also due to the availability of ever larger and more comprehensive datasets [1, 16, 20] (our model's predictions for difficult examples from [1] are shown in Figure 1).

A common characteristic of all ConvNet architectures used for human body pose detection to date is that they make use of internal strided-pooling layers. These layers reduce the spatial resolution by computing a summary statistic over a local spatial region (typically a max operation in the case of the commonly used Max-Pooling layer). The main motivation behind the use of these layers is to promote invariance to local input transformations (particularly translations) since their outputs are invariant to spatial location within the pooling region. This is particularly important for image classification where local image transformations obfuscate object identity. Therefore pooling plays a vital role in preventing over-training while reducing computational complexity for classification tasks.

The spatial invariance achieved by pooling layers comes at the price of limiting spatial localization accuracy. As such, by adjusting the amount of pooling in the network,