Forecasting human dynamics from static images

Shuang Sha

June 31,2018

Abstract

This paper proposed that forecasting human dynamics from static images at the first time. Researchers proposed the 3D Pose Forecasting Network (3D-PFNet). They train our 3D-PFNet using a three-step training strategy to leverage a diverse source of training data, including image and video based human pose datasets and 3D motion capture (MoCap) data.

1. Introduction

Human pose forecasting is the capability of predicting future human body dynamics from visual observations. For example, by looking at the left image of Figure 1, we can predict that the next step of the tennis player, namely a forehand swing. As shown in the right image of Figure 1, this is the sequence of upcoming pose.

Figure 1. Forecasting human dynamics from static images. Left: the input image. Right: the sequence of upcoming poses.

This paper presents the first study on human pose forecasting from static images [1]. The chief task is to take a single RGB image and output a sequence of future human body poses. Firstly, as opposed to other forecasting tasks that assume a multi-frame input (e.g. videos) [2], this work assumes a single-frame input. Secondly, like most forecasting problems [3], this work first represent the forecasted poses in the 2D image space, and then convert from 2D space to 3D space.

2. Network architecture

This paper proposed a deep recurrent network to predict human skeleton sequences, as shown in Figure 2. This network is divided into two components: (1) a 2D pose sequence generator that takes an input image and sequentially generate 2D body poses, where each pose is represented by heatmaps of keypoints; (2) a 3D skeleton converter that converts each 2D pose into a 3D skeleton.

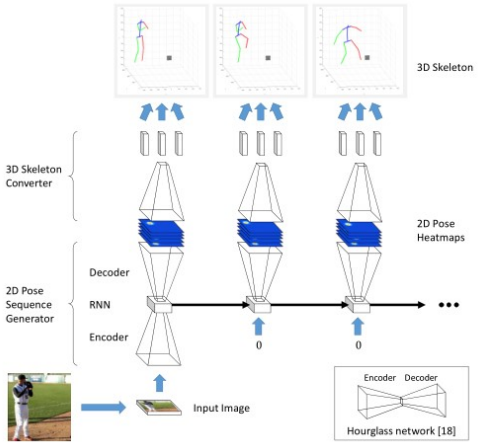

Figure 2. A schematic view of the unrolled 3D-PFNet.

References

[1] Y. W. Chao, J. Yang, B. Price, S. Cohen, and J. Deng. Forecasting human dynamics from static images. In *CVPR*, 2017. 1

[2] N. Srivastava, E. Mansimov, and R. Salakhudinov. Unsupervised learning of video representations using lstms. In *ICML*, pages 843–852, 2015. 1

[3] J. Yuen and A. Torralba. A data-driven approach for event prediction. In *ECCV*, pages 707–720, 2010. 1