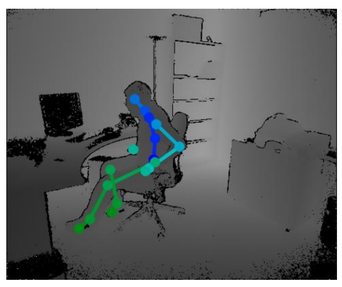
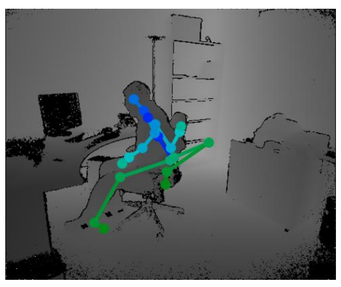
Manual Annotations on Depth Maps for Human Pose Estimation

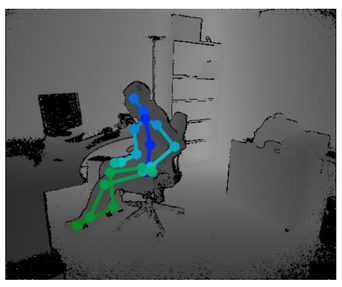
Andrea D'Eusanio, Stefano Pini, Guido Borghi,
Roberto Vezzani, Rita Cucchiara
University of Modena and Reggio Emilia, Italy
name.surname@unimore.it

Abstract

- The estimation of **human body pose** is a common task nowadays, however very few of the existing works use **depth maps** as input:
  - Just few depth map-based datasets are available
  - Body joint annotations are often automatically estimated using the method proposed by Shotton *et al.* [2], which is not reliable in many scenario
- We propose a **new set of annotations** for the **Watch-n-Patch** dataset, obtained by hand with a simple yet effective **annotation tool**
- We present a deep learning-based architecture that performs the body pose estimation directly on depth maps


Manual annotations

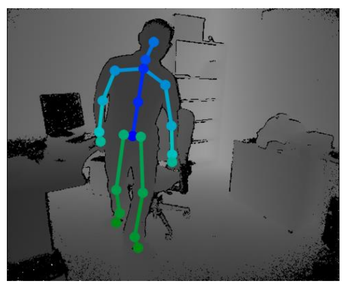

Shotton *et al.* [2]


Predicted

Motivations

We introduce a new dataset with:

- Manual Body Pose annotations
  - Training (or fine-tuning) of new models
  - Effective estimation of existing body pose methods
- Depth data



A baseline is proposed for the evaluation of future works

Benefits of using depth maps:

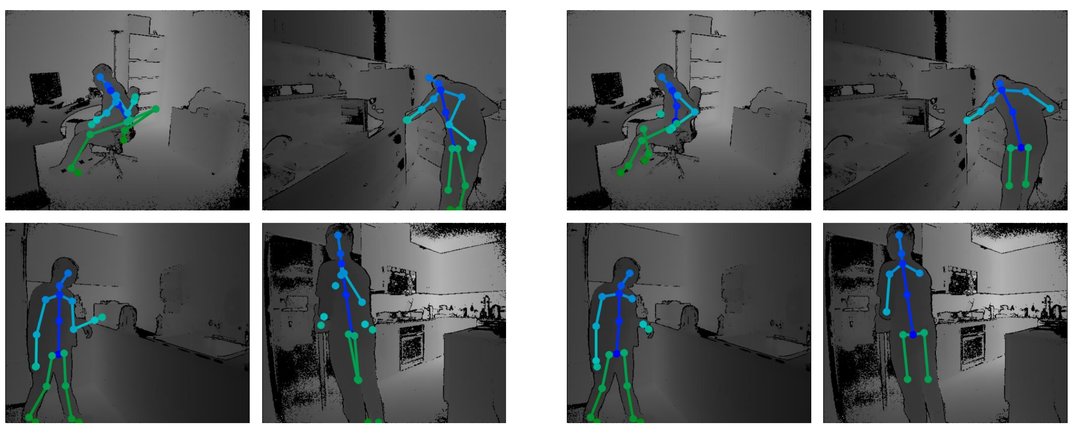
- Light invariance
- New depth sensors:
  - Cheap
  - Small Form Factor
  - High framerate
  - Good Accuracy

Dataset and Annotation Tool

Watch-R(efined)-Patch annotations

We released new set of annotations based on the public dataset *Watch-N-Patch* [4].

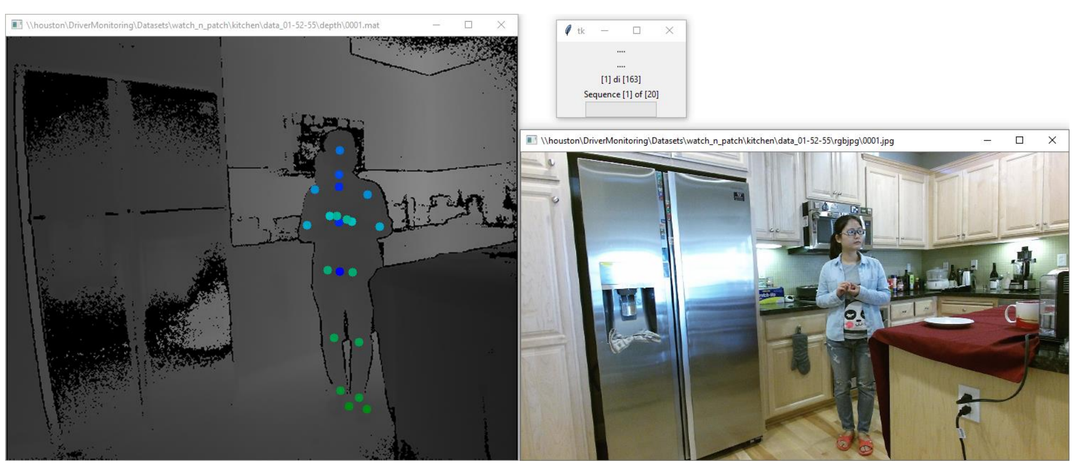
- **Refined Annotations**
  - Train: 1135 frames
  - Validation: 766 frames
  - Test: 1428 frames
  - ~70k refined joints
- **Kinect One** (second version)
  - Depth Maps: 512 x 424
  - RGB images: 1920 x 1080
  - *Time-of-Flight* technology
  - Range: 0.5 – 7 m


Shotton *et al.* annotations

Porposed refined annotations

Annotation tool

In the proposed tool both RGB and depth frame are shown simultaneously.
The RGB is used only as visual reference to help correcting joints in the depth frame.
The tool works with different types of images and we provide various examples.

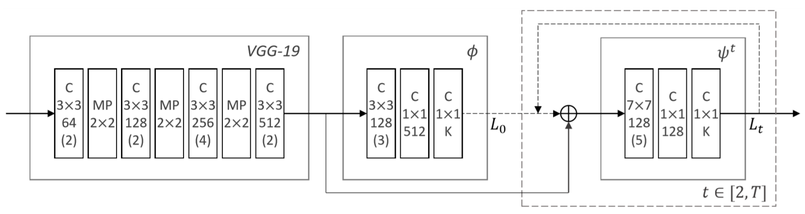


Refining (existing) automatic annotations moving erroneous joints is a faster and simpler way of annotation with respect to annotate the data from scratch.

To this end, our tool simplifies the refinement of existing landmark-based annotations.

Proposed method and Experimental Results

Architecture



Proposed architecture:

- C: Convolutional layer
- MP: Max-pooling layer
- VGG-19 [5]: architecture for feature extraction
- The output of each block $(\phi, \psi^1, \ldots, \psi^t)$ is:

$$\mathbf{P}^1 = \phi(\mathbf{F}, \theta^1)$$

$$\mathbf{P}^t = \psi^t(\mathbf{F} \oplus \mathbf{P}^{t-1}, \theta^t) \quad \forall t \in [2, T]$$

Loss Function

$$L^t = \sum_{k=1}^{K} \alpha_k \cdot \sum_{\mathbf{p}} \|\mathbf{P}_k^t(\mathbf{p}) - \mathbf{H}_k(\mathbf{p})\|_2^2$$

where $\mathbf{P}$ is the prediction at location $\mathbf{p}$ and $\mathbf{H}$ is the ground-truth heatmap.

Results

We use the *mean Average Precision* (mAP) as our main evaluation metrics, expressed by the formula:

$$\text{mAP} = \frac{1}{10} \sum_{i=1}^{10} \text{AP}^{\text{OKS}=0.45+0.05i}$$

where OKS stand for *Object Keypoints Similarity* [3].

mAP reached by different methods computed on the *Watch-R-Patch* dataset.

	Shotton *et al.* [1]	Ours$_{orig}$	Ours$_{last}$	Ours$_{blk}$	**Ours**
$\text{AP}^{\text{OKS}=0.50}$	0.669	0.845	0.834	0.894	**0.901**
$\text{AP}^{\text{OKS}=0.75}$	0.618	0.763	0.758	0.837	**0.839**
mAP	0.610	0.729	0.726	0.792	**0.797**

The model is able to run in real-time (5.37ms, 186fps) on a workstation equipped with an Intel Corei7-6850K and a GPU Nvidia GTX 1080 Ti.

mAP of each body joint present in the *Watch-R-Patch* dataset.

Joint	Shotton *et al.* [1]	Ours$_{orig}$	**Ours**
SpineBase	0.832	0.841	**0.905**
SpineMid	0.931	0.911	**0.935**
Neck	**0.981**	0.975	0.978
Head	**0.971**	0.961	0.962
ShoulderLeft	0.663	0.673	**0.819**
ElbowLeft	0.490	0.635	**0.772**
WristLeft	0.456	0.625	**0.677**
HandLeft	0.406	0.599	**0.680**
ShoulderRight	0.538	0.547	**0.782**
ElbowRight	0.454	0.618	**0.748**
WristRight	0.435	0.642	**0.727**
HandRight	0.412	0.641	**0.712**
HipLeft	0.646	0.766	**0.824**
KneeLeft	0.494	0.743	**0.788**
AnkleLeft	0.543	0.771	**0.800**
FootLeft	0.497	0.743	**0.801**
HipRight	0.696	0.778	**0.860**
KneeRight	0.493	0.670	**0.763**
AnkleRight	0.508	0.630	**0.648**
FootRight	0.388	**0.605**	0.605
SpineShoulder	**0.969**	0.942	0.955

References

[1] Cao, Z., Simon, T., Wei, S.E., Sheikh, Y. "Realtime multi-person 2d pose estimation using part affinity fields", CVPR, 2017
[2] Shotton, J., Fitzgibbon, A., Cook, M. "Real-time human pose recognition in parts from single depth images", CVPR, 2011
[3] Common Objects in Context (COCO) Keypoints evaluation: http://cocodataset.org/#keypoints-eval
[4] Wu, C. *et al.* "Watch-n-patch: Unsupervised learning of actions and relations", IEEE TPAMI, 2018
[5] Simonyan *et al.* "Very deep convolutional networks for large-scale image recognition", ARXIV 2014

Links

Watch-R-Patch

Annotations tool

Video demo