First-Person Pose Recognition using Egocentric Workspaces

Grégory Rogez[1,2], James S. Supančič III[1], Deva Ramanan[1]
[1]Dept of Computer Science, University of California, Irvine, CA, USA. [2]Universidad de Zaragoza, Zaragoza, Spain.

We tackle the problem of estimating the 3D pose of an individual's upper limbs (arms+hands) from a chest mounted depth-camera. Importantly, we consider pose estimation during everyday interactions with objects. Previous work for egocentric hand analysis tends to rely on local 2D features, such as pixel-level skin classification [1, 2] or gradient-based processing of depth maps with scanning-window templates [3]. Our approach follows in the tradition of [3], who argue that near-field depth measures obtained from a egocentric-depth sensor considerably simplifies hand analysis. In egocentric views, hands and arms are observable within a well defined volume in front of the camera that we call an *egocentric workspace*.

Contributions: We describe a new computational architecture that uses **global** egocentric views, **volumetric** representations, and **contextual** models of interacting objects and human-bodies. Rather than detecting hands with a local (translation-invariant) scanning-window classifier, we process the entire global egocentric view (*workspace*) in front of the observer (Fig. 1). Hand appearance is not translation-invariant due to perspective effects and kinematic constraints with the arm. To capture such effects, we build a library of synthetic 3D egocentric workspaces generated using real capture conditions. We animate a 3D human character inside virtual scenes with objects, and render such animations with a chest-mounted camera whose intrinsics match our physical camera. We simultaneously recognize arm and hand poses while interacting with objects by classifying the whole 3D volume using a multi-class Support Vector Machine (SVM) classifier. Recognition is simple and fast enough to be implemented in 4 lines of code.

Synthetic exemplars. Let θ be a vector of arm joint angles, and let ϕ be a vector of grasp-specific hand joint angles, obtained from a set of Poser models covering different grasping hand postures (with/without objects). To enrich the core set of posed hands with additional translations and viewpoints, we take a *rejection sampling* approach: we fix ϕ parameters to respect the hand grasps and add small perturbations to arm joint angles:

$$\theta_i' = \theta_i + \varepsilon \quad \text{where} \quad \varepsilon \sim N(0, \sigma^2).$$

Importantly, this generates hand joints $\mathbf{p}$ at different translations and viewpoints, correctly modeling the dependencies between both. For each perturbed pose, we render hand joints using a forward kinematic chain and keep visible poses (keypoint (u,v) coordinates lie within the image boundaries).

Associated with each pose, we construct a depth map by representing each rigid limb with a dense cloud of 3D vertices $\{\mathbf{u}_i\}$, written in a egocentric (camera) coordinate frame (Fig. 1.a). We render this dense cloud using forward kinematics, producing a set of points $\{\mathbf{p}_i\} = \{(p_{x,i}, p_{y,i}, p_{z,i})\}$. We define a 2D depth map $z[u,v]$ by ray-tracing:

$$z[u,v] = \min_{k \in \text{Ray}(u,v)} ||\mathbf{p}_k|| \tag{1}$$

where $\text{Ray}(u,v)$ denotes the points on the ray passing through pixel (u,v).

Perspective-aware binary depth features: Let us choose spherical bins $F(u,v,w)$ such that they project to a single pixel (u,v) in the depth map. This allows one to compute the binary voxel grid $b[u,v,w]$ by simply "reading off" the depth value for each $z(u,v)$ coordinates, quantizing it to z', and assigning 1 to the corresponding voxel:

$$b[u,v,w] = \begin{cases} 1 & \text{if} \quad w = z'[u,v] \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

This results in a sparse volumetric voxel feature visualized in Fig. 1.b. Once a depth measurement is observed at position $b[u',v',w'] = 1$, all voxels behind it are occluded for $w \geq w'$. We define such occluded voxels to be "1".

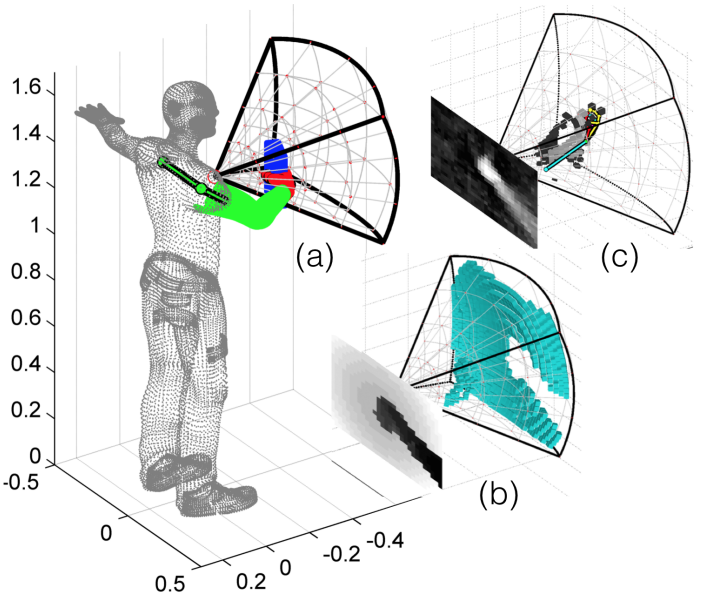

Figure 1: **Egocentric workspaces**. We directly model the observable egocentric workspace in front of a human with a 3D volumetric descriptor, extracted from a 2.5D egocentric depth sensor. We propose an efficient pipeline which 1) generates synthetic workspace exemplars for training using a virtual chest-mounted camera whose intrinsic parameters match our physical camera (a), 2) computes perspective-aware binary depth features on this entire volume (in the example visualized in (b), the volume is discretized into $24 \times 32 \times 35$ bins) and 3) recognizes discrete arm+hand pose classes through a sparse multi-class SVM (c). This computational architecture can be used to accurately predict shoulder, arm, hand poses, even when interacting with objects.

Global classification: We use a linear SVM for multi-class classification of upper-limb poses. However, instead of classifying local scanning-windows, we classify global depth maps quantized into our binarized depth feature $b[u,v,w]$ from (2). Global depth maps allow the classifier to exploit contextual interactions between multiple hands, arms and objects. In particular, we find that modeling arms is particularly helpful for detecting hands. We cluster our dataset and train a one-vs-all SVM classifier for each resulting class k. We obtain K weight vectors which can be re-arranged into $N_u \times N_v \times N_w$ tensors $\beta_k[u,v,w]$. The score for class k is then obtained by a simple dot product of its weights and our binarized feature $b[u,v,w]$:

$$\text{score}[k] = \sum_u \sum_v \sum_w \beta_k[u,v,w] \cdot b[u,v,w]. \tag{3}$$

In Fig. 1.c, we show the weight tensor $\beta_k[u,v,w]$ for a particular pose cluster. To increase run-time efficiency, we exploit the sparsity of our binarized volumetric feature and jointly implement feature extraction and SVM scoring. Since our binarized depth features do not require any normalization and the classification score is a simple dot product, we can readily extract the feature and update the score on the fly.

Results: We achieve state-of-the-art hand pose recognition performance from egocentric RGB-D images in real-time (275 fps).

[1] Cheng Li and Kris M. Kitani. Pixel-level hand detection in ego-centric videos. In *CVPR*, 2013.

[2] Cheng Li and Kris M. Kitani. Model recommendation with virtual probes for egocentric hand detection. In *ICCV*, 2013.

[3] Gregory Rogez, Maryam Khademi, J.S Supancic, J.M.M. Montiel, and Deva Ramanan. 3d hand pose detection in egocentric rgbd images. In *ECCV Workshop on Consuper Depth Camera for Vision (CDC4V)*, pages 1–11, 2014.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.