Hypothesis 1

> Action Information improves HPE.

Test to see whether action information actually improves HPE, if that's not the case then its an early indication to discuss another direction.

Steps

a. Train on hand-action-data-set (FHAD) + HPE (deep-prior) using depth as input, pose as output; measure performance.
b. Train on FHAD+HPE using depth as input and concatenated* action information and pose as output; measure performance.

Expected results

Strictly we wish to see perf(b) > perf(a) if yes, continue if not discuss/investigate further.

Note: For concatenation, because it involves concatenating a 2D image with a 1D one-hot class vector, there are several ways to do it. One paper on cGANs does it using spatial repetition. DCGAN does it using transposed_conv. I can probably try both methods:

- Src: https://arxiv.org/pdf/1611.06355.pdf (for spatial repetition, paper says concat after 1st conv layer is best)
- Src2: https://arxiv.org/pdf/1511.06434.pdf (latent z (1D) is converted to 2D using transposed_conv)
- Src3 (extra/conv-arithmetic): https://arxiv.org/pdf/1603.07285.pdf & https://pytorch.org/docs/stable/nn.html#convtranspose2d

Hypothesis 2

> For skeleton-based action recognition accurate pose (skeleton) estimates per frame is not very important.

This is a test to say temporal information (the transition of poses) is more important than per frame pose estimation. That is, if we use estimated poses then our performance doesn't drop significantly. This is already shown in FHAD paper, that under a specifc HPE arch and training data, perf difference is ~6% in action recognition

Steps

a. Train HAR on gt pose and gt action -- done, ~acc 71%
b. Train HAR based on pose estimates from pretrained_HPE and gt action.
c. Train HAR based on pose estimates on model simpler than deep-prior++ (e.g. Resnet-18 instead of 50) on FHAD and gt action.

Expected Results

Perf_HAR(b) < Perf_HAR(a) but not by significant amount. Also %error increase in Perf_HPE(c) vs Perf_HPE(b) should be ideally lower than Perf_HAR(c) vs Perf_HAR(b). If that's not the case then we know that we need a

very good HPE for HAR based on pose estimates.

From FHAD paper it is shown that a significant drop in HPE e.g. 15mm vs 30mm gives a much higher drop in HAR i.e. 30% vs 78%. What I want to investigate is if the same holds true for lower drops e.g 15mm vs 20mm or at what point (in mm) it gets 'really bad'. This will give a buffer threshold for all future HPE extensions. If this is already found then please let me know.

Next Steps

> For now for Hypothesis 1, I wish to know how exactly (and what arch) was used to train HPE in FHAD. So far details I know are 300k egocentric samples from Hands2.2m + train samples from FHAD 1:1 split? Can I get a link to download Hands2.2m samples please?