View-Invariant Probabilistic Embedding for Human Pose Supplementary Materials

Jennifer J. Sun[1], Jiaping Zhao[2], Liang-Chieh Chen[2], Florian Schroff[2], Hartwig Adam[2], and Ting Liu[2]

[1] California Institute of Technology
jjsun@caltech.edu
[2] Google Research
{jiapingz,lcchen,fschroff,hadam,liuti}@google.com

In this document, we cover the details of the implementation and experiments for our work. We also provide additional ablation studies and analysis. Specifically, we have:

- **Section 1** describes how we decide the **NP-MPJPE threshold** based on its effect on visual pose similarity.
- **Section 2** provides additional **implementation details** on model training, keypoint definition and normalization, downstream task experiment setup, etc.
- **Section 3** provides additional **ablation studies**, including the effect of key hyperparameters, ordered embedding variance visualizations, and embedding space visualization.
- **Section 4** provides additional **quantitative pose retrieval** result comparisons with image-based EpipolarPose model [5] for view-invariant pose retrieval.
- **Section 5** provides additional **qualitative pose retrieval** results.
- **Section 6** describes the **qualitative video alignment** experiment. Please refer to `https://drive.google.com/open?id=1kTc_UT0Eq0H2ZBgfEoh8qEJMFBouC-Wv` for the video synchronization results.

1 Visualization of 3D Visual Similarity

The 3D pose space is continuous, and we use the NP-MPJPE as a proxy to quantify visual similarity between pose pairs. Fig. 1 shows pairs of 3D pose keypoints with their corresponding NP-MPJPE, where each row depicts a different NP-MPJPE range. This plot demonstrates the effect of choosing different κ, which controls matching threshold between 3D poses. If we choose $\kappa = 0.05$, then only the first row in Fig. 1 would be considered matching, and the rest of the rows are non-matching. Our current value of $\kappa = 0.10$ corresponds to using the first two rows as matching pairs and the rest of the rows as non-matching ones. By loosening κ, poses with greater differences will be considered as matching, as shown by different rows in Fig. 1. We note that pairs in rows 3 and 4 shows significant visual differences compared with the first two rows. We further investigate the effects of different κ during training and evaluation in Section 3.

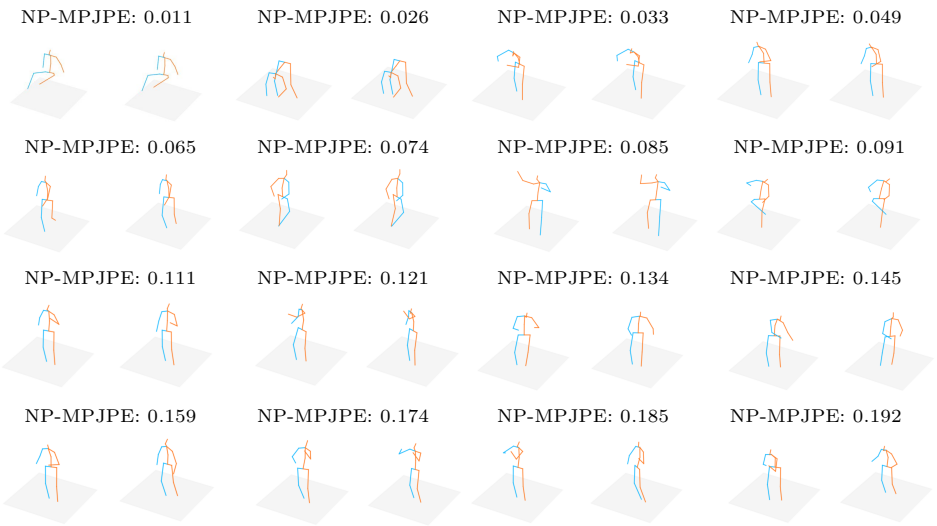


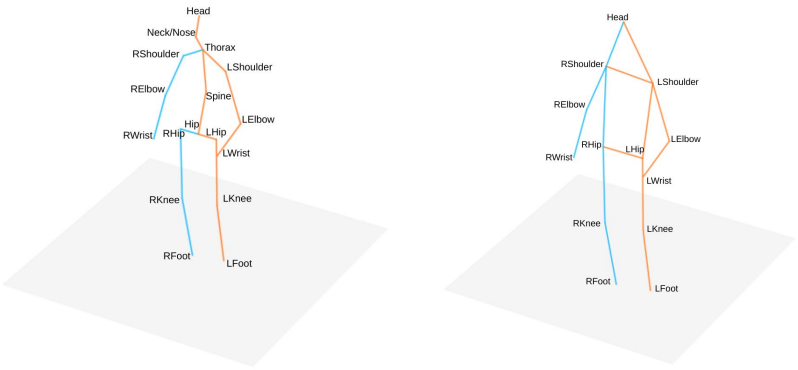
Fig. 1: 3D pose pairs with different NP-MPJPE, where the NP-MPJPE increases with each row. The poses are randomly sampled from the hold-out set of H3.6M. Row 1 shows pairs with 0.00 to 0.05 NP-MPJPE, row 2 shows pairs with 0.05 to 0.10 NP-MPJPE, row 3 shows pairs with 0.10 to 0.15 NP-MPJPE, and row 4 shows pairs with 0.15 to 0.20 NP-MPJPE.

2 Additional Implementation Details

The backbone network architecture for our model is based on [7]. We use two residual blocks, batch normalization, 0.3 dropout, and no maximum weight norm constraint [7]. During training, we use exponential moving average with 0.9999 decay rate and normalize matching probabilities to within $[0.05, 0.95]$ for numerical stability. We use Adagrad optimizer [2] with fixed learning rate 0.02 and batch size $N = 256$.

Keypoint Definition Fig. 2 illustrates the keypoints that we use in our experiments. The 3D poses used in our experiments are the 17 keypoints corresponding to the H3.6M [4] skeleton used in [7], shown in Fig. 2a. We use this keypoint definition to compute NP-MPJPE for 3D poses and evaluate retrieval accuracy. The Pr-VIPE training and inference process do not depend on a particular 2D keypoint detector. Here, we use PersonLab (ResNet152 single-scale) [8] in our experiments. Our 2D keypoints are selected from the keypoints in COCO [6], which is the set of keypoints detected by PersonLab [8]. We use the 12 body keypoints from COCO and select the "Nose" keypoint as the head, shown in Fig. 2b.

Pose Normalization We normalize our 2D and 3D poses such that camera parameters are not needed during training and inference. For 3D poses, our normalization procedure is similar to that in [1]. We translate a 3D pose so that the hip located at the origin. We then scale the hip to spine to thorax



(a) 17 keypoints based on H3.6M. (b) 13 keypoints based on COCO.

Fig. 2: Visualization of pose keypoints used in our experiments.

distance to a unit scale. For 2D poses, we translate the keypoints so that the center between LHip and RHip is at the origin. Then we normalize the pose such that the maximum distance between shoulder and hip joints is 0.5. This maximum distance is computed between all pairwise distances among RShoulder, LShoulder, RHip, and LHip.

Downstream Task Experiments For the action recognition experiment, we follow the standard evaluation protocol [10] and remove action "strum guitar" and several videos in which less than one third of the target person is visible. We use the official train/test split and report the averaged per-class accuracy. For the view-invariant action recognition experiments in which the index set only contains videos from a single view, we exclude the actions that have zero or only one sample under a particular view. We take the bounding boxes provided with the dataset and use [9] (ResNet101) for 2D pose keypoint estimation. For frames of which the bounding box is missing, we copy the bounding box from the nearest frame. Finally, since our embedding is chiral, but certain actions can be done with either body side (pitching a baseball with left or right hand), when we compare two frames, we extract our embeddings from both the original and the mirrored version of each frame, and use the minimum distance between all the pairwise combinations as the frame distance.

For the video alignment experiment, we follow the protocol in [3], excluding "jump rope" and "strum guitar" from our evaluation. For the evaluations between videos under only the same or different views, we exclude actions that have zero videos under a particular view from the average Kendall's Tau computation. Since certain actions can be done with either body side, for a video pair (v_1, v_2), we compute the Kendall's Taus between (v_1, v_2) and $(v_1, \text{mirror}(v_2))$, and use the larger number.

3 Additional Ablation Studies

Effect of Number of Samples K and Margin Parameter β Table 1 shows the effect of the number of samples K and the margin parameter β (actual triplet margin $\alpha = \log \beta$) on Pr-VIPE. The number of samples control how many points we sample from the embedding distribution to compute matching probability and β controls the ratio of matching probability between matching and non-matching pairs. Our model is robust to the choice of β in terms of retrieval accuracy as shown by Table 1. The main effect of β is on retrieval confidence, as non-matching pairs are scaled to a smaller matching probability for larger β. Pr-VIPE performance with 10 samples is competitive with the baselines in the main paper, but we do better with 20 samples. Increasing the number of samples further has similar performance. For our experiments, we use 20 samples and $\beta = 2$.

Table 1: Additional ablation study results of Pr-VIPE on H3.6M with the number of samples K and margin parameter β.

Hyperparameter	Value	Hit@1	Hit@10	Hit@20
K	10	0.744	0.948	0.971
	20	0.762	0.956	0.977
	30	0.755	0.955	0.975
β	1.25	0.758	0.956	0.977
	1.5	0.759	0.956	0.977
	2	0.762	0.956	0.977
	3	0.760	0.955	0.976

Effect of Camera Augmentation We explore the effect of different random rotations during camera augmentation on pose retrieval results in Table 2. All models are trained on the 4 chest-level cameras on H3.6M but the models with camera augmentation also use projected 2D keypoints from randomly rotated 3D poses. For the random rotation, we always use azimuth range of $\pm 180°$, and we test performance with different angle limits for elevation and roll. We see that the model with no augmentation does the best on the H3.6M, which has the same 4 camera views as training. With increase in rotation angles during mixing, the performance on chest-level cameras drop while performance on new camera views generally increases. The results demonstrate that mixing detected and projected keypoints reduces model overfitting on camera views used during training. Training using randomly rotated keypoints enables our model to generalize much better to new views.

Effect of NP-MPJPE threshold κ We train and evaluate with different values of the NP-MPJPE threshold κ in Table 3. κ controls the NP-MPJPE threshold for a matching pose pair and visualizations of pose pairs with different NP-MPJPE are in Fig. 1. Table 3 shows that Pr-VIPE generally achieves the best

Table 2: Additional ablation study results of Pr-VIPE on H3.6M and 3DHP using different rotation thresholds for camera augmentation. The angle threshold for azimuth is always $\pm 180^\circ$ and the angle thresholds in the table are for elevation and roll. The row for w/o aug. corresponds to Pr-VIPE without augmentation.

Hyperparameter	Range	Hit@1 on evaluation dataset H3.6M	3DHP (all)	3DHP (chest)
Elevation and Roll Angle	w/o aug.	0.762	0.199	0.255
	$\pm 15^\circ$	0.747	0.252	0.289
	$\pm 30^\circ$	0.737	0.264	0.283
	$\pm 45^\circ$	0.737	0.262	0.273

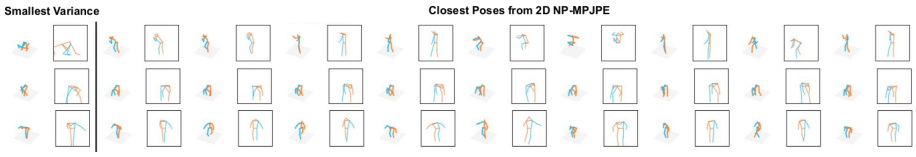
Table 3: Additional ablation study results of Pr-VIPE on H3.6M with different NP-MPJPE threshold κ for training and evaluation.

Training κ	Hit@1 with evaluation κ 0.05	0.10	0.15	0.20
0.05	**0.495**	0.761	0.908	0.962
0.10	0.489	**0.762**	0.909	0.963
0.15	0.462	0.753	**0.910**	**0.965**
0.20	0.429	0.731	0.906	**0.965**

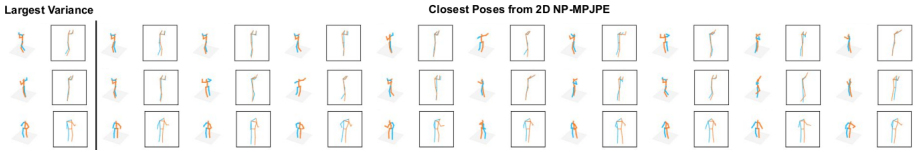
accuracy for a given NP-MPJPE threshold when the model is trained with the same matching threshold. Additionally, when we train with a tight threshold, e.g., $\kappa = 0.05$, we do comparatively well on accuracy at looser thresholds. In contrast, when we train with a loose threshold, e.g., $\kappa = 0.20$, we do not do as well given a tighter accuracy threshold at evaluation. This is because when we push non-matching poses using the triplet ratio loss, $\kappa = 0.20$ only pushes poses that are more than 0.20 NP-MPJPE apart, and does not explicitly push poses less than the NP-MPJPE threshold. The closest retrieved pose will then be within 0.20 NP-MPJPE but it is not guaranteed to be within any threshold < 0.20 NP-MPJPE. But when we use $\kappa = 0.05$ for training, poses that are more than 0.05 NP-MPJPE are pushed apart, which also satisfies $\kappa = 0.20$ threshold.

In the main paper, we use $\kappa = 0.1$. For future applications with other matching definitions, the Pr-VIPE framework is flexible and can be trained with different κ to satisfy different accuracy requirements.

Additional Plots for Ordered Variances Similar to the main paper, we retrieve poses using 2D NP-MPJPE for the top-3 2D poses with smallest and largest variances in Fig. 3. Fig. 3a shows that for the poses with the top-3 smallest variances, the nearest 2D pose neighbors are visually distinct, which means that these 2D poses are less ambiguous. On the other hand, the nearest 2D pose neighbors of the poses with the largest variances in Fig. 3b are visually similar, which means that these 2D poses are more ambiguous.



(a) Poses with top-3 smallest variance and their nearest neighbors in terms of 2D NP-MPJPE.



(b) Poses with top-3 largest variance and their nearest neighbors in terms of 2D NP-MPJPE.

Fig. 3: Top retrievals by 2D NP-MPJPE from H3.6M hold-out subset for queries with top-3 largest and smallest variances. 2D poses are shown in the boxes.

Embedding Space Visualization We run Principal Component Analysis (PCA) on the 16-dimensional embeddings using the Pr-VIPE model. Fig. 4 visualizes the first two principal dimensions. To visualize more unique poses, we randomly subsample the H3.6M hold-out set and select 3D poses at least 0.1 NP-MPJPE apart. Fig. 4 demonstrates that 2D poses from similar 3D poses are close together, while non-matching poses are further apart. Standing and sitting poses seem well separated from the two principle dimensions. Additionally, there are leaning poses between sitting and standing. Poses near the top of the figure have arms raised, and there is generally a gradual transition to the bottom of the figure, where arms are lowered. These results show that from 2D joint keypoints only, we are able to learn view-invariant properties with compact embeddings.

4 Additional Quantitative Pose Retrieval Results

We show an additional view-invariant pose retrieval evaluation comparing Pr-VIPE (with camera augmentation) to EpipolarPose [5], a recent multi-view image based model, on cross-view pose retrieval. For Human3.6M, EpipolarPose is trained with the same training split as Pr-VIPE. The evaluation split we use is a frame subset provided by [5] for which the authors provided cropping boxes based on groundtruth 3D keypoints. The input images are cropped using these bounding boxes, and the trained models provided by the authors are then ran on the cropped images. In this way, we evaluate EpipolarPose using all the information provided by the authors. In comparison, Pr-VIPE uses detected keypoints and no groundtruth information for inference.

We show retrieval results on Human3.6M since [5] is based on images and requires a different model to be trained for 3DHP. We emphasize that this is a different evaluation split from our main paper, since we use the evaluation subset of Human3.6M for which [5] provides bounding boxes. On this subset, Pr-VIPE

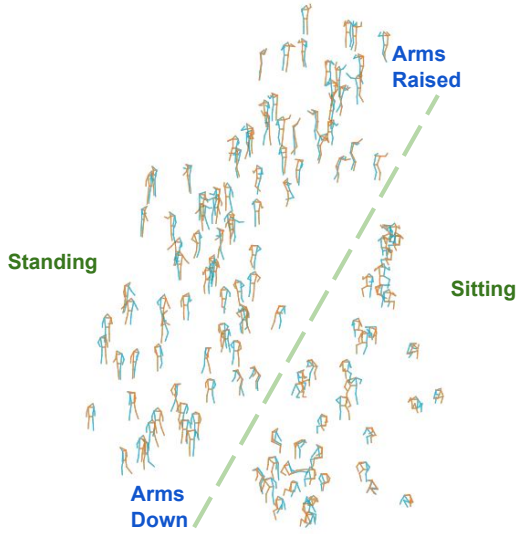


Fig. 4: Visualization of Pr-VIPE space with 2D poses in the H3.6M hold-out subset using the first two PCA dimensions.

with augmentation achieves 75.2% Hit@1, fully supervised EpipolarPose achieves 72.7% Hit@1 and self-supervised EpipolarPose achieves 67.8% Hit@1.

These results show the effectiveness of Pr-VIPE for pose retrieval. Our model, using detected 2D keypoints and no groundtruth information, can retrieve poses more accurately compared with [5]. We further note that 3D pose estimation models require rigid alignment between every query-index pairs to achieve their best performance for retrieval, while Pr-VIPE does not require post-processing.

5 Additional Qualitative Pose Retrieval Results

We present more view-invariant pose retrieval qualitative results for Pr-VIPE on all the relevant datasets in Fig. 5. The first two rows show results on H3.6M, the next three rows are on 3DHP and the last two rows shows results using the hold-out set in H3.6M to retrieve from 2DHP. We are able to retrieve across camera views and subjects on all datasets.

On H3.6M, retrieval confidence is generally high and retrievals are visually accurate. NP-MPJPE is in general smaller on H3.6M compared to 3DHP, since 3DHP has more diverse poses and camera views. The model works reasonably well on 3DHP despite additional variations on pose, viewpoints and subjects. For the pairs R4C3 and R5C3, the subjects are occluded by the chair and the pose inferred by the 2D keypoint detector may not be accurate. Our model is dependent on the result of the 2D keypoint detector. Interestingly, R3C2 and R4C3 show retrievals with large rolls, which is unseen during training. The results on 3DHP demonstrate the generalization capability of our model to unseen poses and views. To test on in-the-wild images, we use the hold-out set of H3.6M

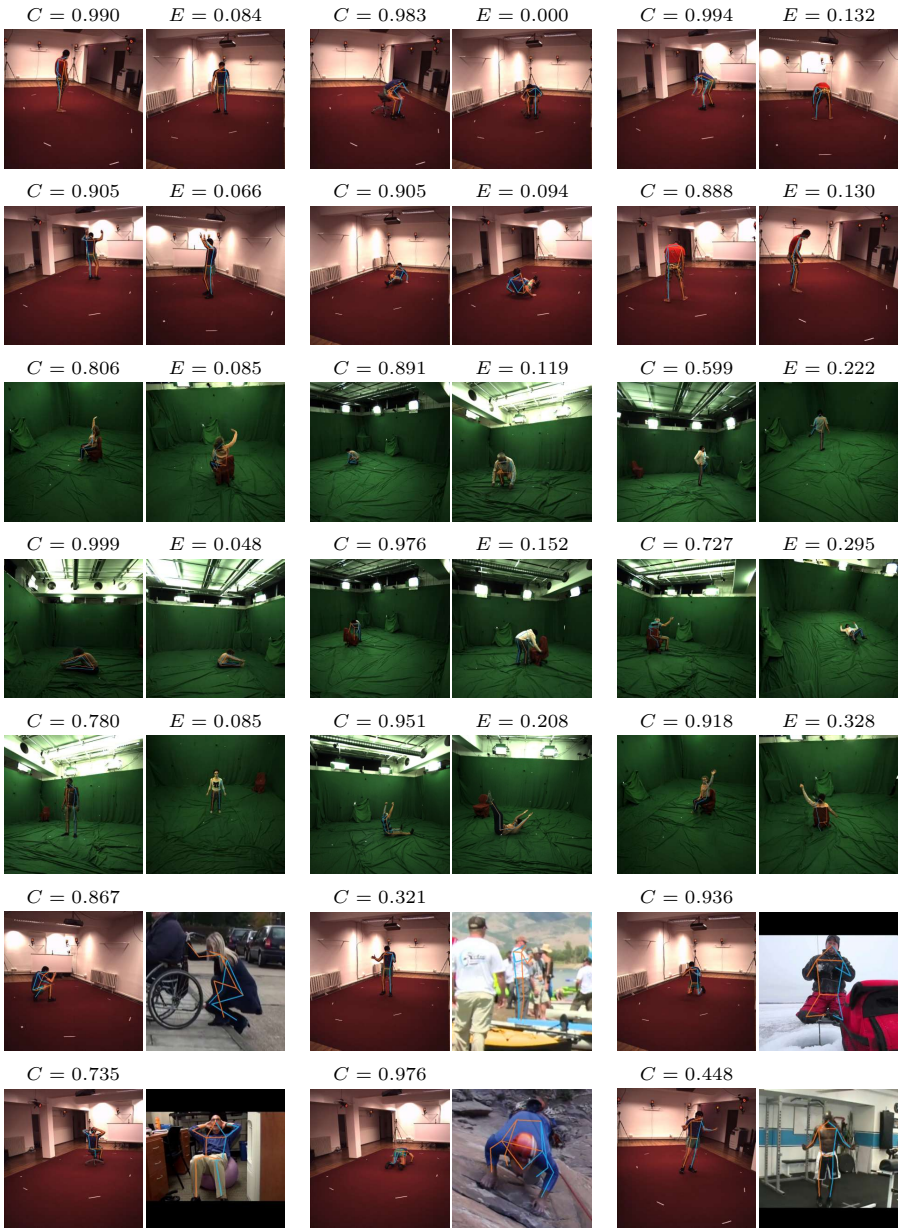


Fig. 5: Visualization of pose retrieval results. On each row, we show the query pose on the left for each image pair and the top-1 retrieval using the Pr-VIPE model with camera augmentation on the right. We display the retrieval confidences ("C") and top-1 NP-MPJPEs ("E", if 3D pose groundtruth is available).

to retrieve from 2DHP. The retrieval results demonstrate that Pr-VIPE embeddings can retrieve visually accurate poses from detected 2D keypoints. R7C2 is particularly interesting, as the retrieval has a large change in viewpoint. For the low confidence pairs R6C2 and R7C3, we can see that the arms of the subjects seems to be bent slightly differently. In contrast, the higher confidence retrieval pairs looks visually similar. The results suggest that performance of existing 2D keypoint detectors, such as [8], is sufficient to train pose embedding models to achieve the view-invariant property in diverse images.

6 Qualitative Video Alignment Results

We show that Pr-VIPE can be applied to synchronize action videos from different views from the Penn Action dataset (test set). The videos are synchronized to the pace of a target video (placed in the center of each video array). This allows us to play different videos of the same action at the same pace. The results for different aligned actions are located at `https://drive.google.com/open?id=1kTc_UT0Eq0H2ZBgfEoh8qEJMFBouC-Wv`. The alignment procedure for Pr-VIPE is described in Section 4.3.2 in the main paper.

References

1. Chen, C.H., Tyagi, A., Agrawal, A., Drover, D., Stojanov, S., Rehg, J.M.: Unsupervised 3D pose estimation with geometric self-supervision. In: CVPR (2019)
2. Duchi, J., Hazan, E., Singer, Y.: Adaptive subgradient methods for online learning and stochastic optimization. JMLR (2011)
3. Dwibedi, D., Aytar, Y., Tompson, J., Sermanet, P., Zisserman, A.: Temporal cycle-consistency learning. In: CVPR (2019)
4. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. IEEE TPAMI (2013)
5. Kocabas, M., Karagoz, S., Akbas, E.: Self-supervised learning of 3D human pose using multi-view geometry (2019)
6. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: ECCV (2014)
7. Martinez, J., Hossain, R., Romero, J., Little, J.J.: A simple yet effective baseline for 3D human pose estimation. In: ICCV (2017)
8. Papandreou, G., Zhu, T., Chen, L.C., Gidaris, S., Tompson, J., Murphy, K.: PersonLab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model. In: ECCV (2018)
9. Papandreou, G., Zhu, T., Kanazawa, N., Toshev, A., Tompson, J., Bregler, C., Murphy, K.: Towards accurate multi-person pose estimation in the wild. In: CVPR (2017)
10. Xia, L., Chen, C.C., Aggarwal, J.K.: View invariant human action recognition using histograms of 3D joints. In: CVPRW (2012)