Normalized Human Pose Features for Human Action Video Alignment

Jingyuan Liu[1] Mingyi Shi[2] Qifeng Chen[1] Hongbo Fu[3] Chiew-Lan Tai[1]
[1]The Hong Kong University of Science and Technology
[2]The University of Hong Kong [3]The City University of Hong Kong
[1]{jliucb,cqf,taicl}@cse.ust.hk [2]myshi@cs.hku.hk [3]hongbofu@cityu.edu.hk

(a) (b) (c)

Figure 1. We propose to normalize human poses in video frames for computing pose similarity. People with different anthropometry (b) lead to differences in joint positions when performing the same pose (a), but their joint rotations are similar. The same pose performed by different subjects in (a) and (b) can be represented by the same normalized pose (c) to retain only pose information.

Abstract

We present a novel approach for extracting human pose features from human action videos. The goal is to let the pose features capture only the poses of the action while being invariant to other factors, including video backgrounds, the video subjects' anthropometric characteristics and viewpoints. Such human pose features facilitate the comparison of pose similarity and can be used for downstream tasks, such as human action video alignment and pose retrieval. The key to our approach is to first normalize the poses in the video frames by mapping the poses onto a pre-defined 3D skeleton to not only disentangle subject physical features, such as bone lengths and ratios, but also to unify global orientations of the poses. Then the normalized poses are mapped to a pose embedding space of high-level features, learned via unsupervised metric learning. We evaluate the effectiveness of our normalized features both qualitatively by visualizations, and quantitatively by a video alignment task on the Human3.6M dataset and an action recognition task on the Penn Action dataset.

1. Introduction

Video alignment aims to find dense temporal correspondences between a pair of videos. Finding alignments between two natural human action videos is especially challenging because the two videos to be aligned can have large variations in many factors, such as scales and orientations of the video subjects, camera viewpoints, action speeds and orientations, etc. A feature that is robust to these variations is desirable in finding the alignments.

A common approach to the problem of human action video alignment is to first estimate 2D or 3D human poses from two input videos, and then find the alignments by matching with features extracted from joint positions [48, 11], so as to reduce certain interference in video backgrounds and subjects' clothing. However, human poses still contain large variations in scale, bone length ratios, orientations, etc. Since existing 3D pose estimation methods [28, 37] recover 3D poses in camera coordinate systems, the joint positions relative to a root joint are dependent on viewpoints (as illustrated by a toy example in the supplementary material). Global orientation normalization by Procrustes alignment is hard to be applied to in-the-wild videos when the ground-truth 3D poses are not available. Besides viewpoint, the joint positions computed by existing 3D pose estimation methods are also dependent on the video subjects' anthropometric characteristics, such as bone lengths and ratios. Such anthropometric variation would cause a difference in distance measurements (e.g., L2 distance after Procrustes alignment) even when the subjects in the videos perform exactly the same poses, as illustrated in Figure 1.

Given the above limitations of using joint position-based pose representations for video alignment, an important observation is that pose similarity is better described by joint angular representations than relative joint positions. The poses of two subjects performing an identical pose should have the same joint angles or joint rotations, but could produce a difference in joint positions due to the difference in relative bone lengths, as illustrated in Figure 1. In addition, relative joint angles or rotations of physically connected joints are consistent among cameras and invariant to viewpoints. Thus the key to extract subject- and scene-invariant features for comparison is to extract features with respect to joint angular representations rather than joint position-based representations.

At first sight, a straightforward solution might be to compute joint angles from joint positions, and use raw joint angles [10] or their aggregations [35, 52] as features for matching. However, the joint angle features suffer from information loss by dropping the skeleton's relational context, which has a proven significance in capturing pose discrimination [8, 40]. Joint rotations also have limitations in that either directly regressing 3D joint rotations from 2D poses, or computing joint rotations from 3D poses by inverse kinematics (IK) is an ill-posed problem, where multiple possible sets of joint rotations can be mapped to the same set of joint positions [17, 53]. Even though existing works have attempted to add kinematic constraints to reduce the IK ambiguity [17, 46], it is still impractical to compare pose similarities directly in the joint rotation space using the joint rotations computed from joint positions [57].

To address the limitations in position-based and angular-based pose representations, we propose to use a normalized human pose, an intermediate pose representation that reflects the pose information with respect to joint rotations, and is parameterized by joint positions to preserve the relational context of body configurations, as shown in Figure 1(c). This normalized pose representation is enlightened by the recent works that use joint rotations as pose parameterizations for motion reconstruction [45] and pose sequence generation [53, 38]. They incorporate a deterministic forward kinematics (FK) layer in neural networks to convert the joint rotations into joint positions to avoid the joint rotation ambiguity problem in IK. FK recursively rotates the bones in a skeleton from a root joint to the leaf joints according to the joint rotations, resulting in joint positions that can be supervised by ground-truth joint positions. We adopt an FK layer to perform pose normalization. Our normalized pose representation retains the joint rotations of the subjects in video frames, such that it captures the pose information and is invariant to all other factors related to the original scene and subject in the video; and is parameterized by the joint positions of a pre-defined skeleton to reduce ambiguity in comparing pose similarities.

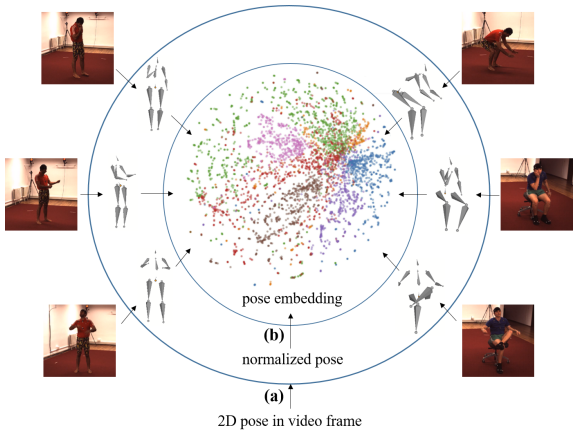

Figure 2. The pipeline of our proposed method. (a) Pose normalization: the 2D pose in each video frame is mapped onto a 3D condition skeleton; (b) pose embedding: the 3D condition skeleton poses are mapped to a pose embedding space.

We design a neural network that learns to normalize human poses in videos. Specifically, the pose normalization network takes in 2D poses and estimate joint rotations, which are then applied by FK on a pre-defined 3D skeleton with unified fixed bone lengths (called *condition skeleton* as in [53]). The joint rotations of the subjects' poses in videos are thus converted into the joint positions of the condition skeleton to normalize the poses. In this way, the difference in joint positions of the condition skeleton is caused only by the difference in joint rotations. Since the normalized poses are not paired with ground-truth poses for training, our network adopted a cycle consistency training strategy (Section 3.2). With joint rotations, the poses can also be easily unified to the same global orientation by specifying the root joint rotation. Finally, the pose features are learned from the normalized 3D poses by metric learning. The resulting pose features are high-level human pose representations and can be directly compared by the Euclidean distance.

In this paper we mainly focus on the video alignment task, but the proposed feature could also be used for other pose similarity tasks, such as pose retrieval, action detection, etc. Experiments show that our proposed normalized pose is robust to variations in viewpoint and subjects' anthropometry. Pose features learned from the normalized poses have shown proven performances on a dense correspondence task on the Human3.6M dataset, and an action recognition task on the Penn Action dataset.

2. Related Work

Human Action Video Alignment. Alignment of human action videos has been actively explored in recent years for many video analysis tasks, such as action detection in unconstrained videos [16], human reconstruction from

uncalibrated multiview videos [11], action synchronization [12], few-shot video classification [6], etc. Since there is no large-scale dataset with frame-by-frame labeled alignments, existing works on human action video alignment have focused on exploiting natural spatio-temporal relations in videos and designed self-supervised tasks to bypass the requirement on datasets. For example, Dwibedi et al. [12] adopted cycle-consistency learning to maximize the number of corresponding frames between videos; Sermanet et al. [44] utilized multiview videos for cross-view correspondence; Misra et al. [33] and Sumer et al. [47] proposed to learn visual representations by sequence verification tasks that penalize the temporal order of action subsequences. While existing methods design temporal modeling techniques specific to video alignment to handle variations in videos, our method aims at obtaining normalize human poses that can also generalize to other pose similarity-related applications.

Human Pose Parameterizations. A common human pose parameterization is via joint positions. Approaches for regressing 3D joint positions from video frames or 2D poses have been extensively studied in compute vision [28, 49, 50]. However, regressing 3D poses from 2D information often suffers from artifacts due to the projective ambiguity. To reduce the ambiguity, a 3D articulated body model can be introduced to provide physical constraints. In this case, the poses in videos are often reconstructed by fitting the body model's projection to the 2D poses [30, 29] or video frames [23, 24], and the pose features can be represented by the body model parameters [23, 29]. Another parameterization of human poses is by joint rotations such that kinematic techniques could apply. Among the rotation parameterizations, Euler angles and exponential maps would cause exploding gradients due to their discontinuities and singularities and thus are not suitable for neural networks [38]. More optimized joint rotation parameterizations, such as quaternions [53, 45] and 6Ds [58], have been adopted in neural networks to ensure continuity. In this paper, we adopt quaternions to represent joint rotations, since they are compatible with major animation software, such as Blender and Unity.

Human Pose Features. The features (representations) of human pose or motion in videos have been extensively studied for down-stream tasks, such as video frame retrieval [48, 10], human pose estimation [18, 9, 41], motion graph transition detection [25, 2], etc. Hand-crafted low-level features, such as Euler angles computed from joint positions [8, 10] and pixel intensity-based features [13], are not comprehensive descriptions that capture contextual latent representations of human poses. Several works [34, 12] also proposed to learn features directly from video frames, without detecting body parts. These features extracted from general video frames are not tailored for human action videos and are subjected to variations in camera motions and background motions. Deep neural networks designed for human pose related tasks, such as human pose estimation [41], poses synthesis [40] and motion retargeting [1], capture certain pose features in network latent layers. But such latent representations are task-specific, and their distances might not directly reflect pose similarities. A few existing works also proposed to learn high-level human pose features by metric learning techniques from 2D poses or images [34, 47, 48]. Please see the following "Deep Representation Learning" subsection.

Deep Representation Learning. Representation learning predicts relative distances between different instances within a category, such as human faces [43], texts [31], graphs [5], motions [2], as well as human poses [34, 47, 48]. Contrary to these pose embedding methods that adopt assumptions on 2D poses [34, 48] and video temporal ordering [47] to indicate pose similarity, and let the neural networks learn to cover the variations in poses, our pose normalization is equivalent to a pre-processing step that explicitly factors out the variations in dataset entities to avoid introducing assumptions in the metric learning. Various losses have been proposed for the learning of relative distances. Besides the commonly used contrastive loss [19, 20] and triplet loss [43, 54], soft contrastive loss [5] and triplet ratio loss [48] enable probabilistic embedding that models the input uncertainties, and circle loss [51] re-weights similarities for more flexible optimizations. To automatically define similar and dissimilar instances, various mining strategies have also been proposed, such as semi-hard triplet mining [43], online triplet mining [21], batch-hard strategy [32], etc. In this paper, we also propose an adaptive triplet sampling strategy for human action videos (Section 3.3).

3. Methodology

3.1. Overview

Figure 2 shows the pipeline of our method. The input is 2D poses detected from video frames using off-the-shelf 2D pose detectors [7, 14]. The pipeline contains two steps: (1) pose normalization, which maps the poses of the subject in the video onto a 3D condition skeleton, such that the poses are disentangled from the video subject's anthropometry and unified to the same global orientation; (2) pose embedding, which extracts features of the normalized 3D poses by mapping them to a pose embedding space.

3.2. Pose Normalization

Figure 3 shows the pose normalization training pipeline. The core of this model is applying the 3D joint rotations

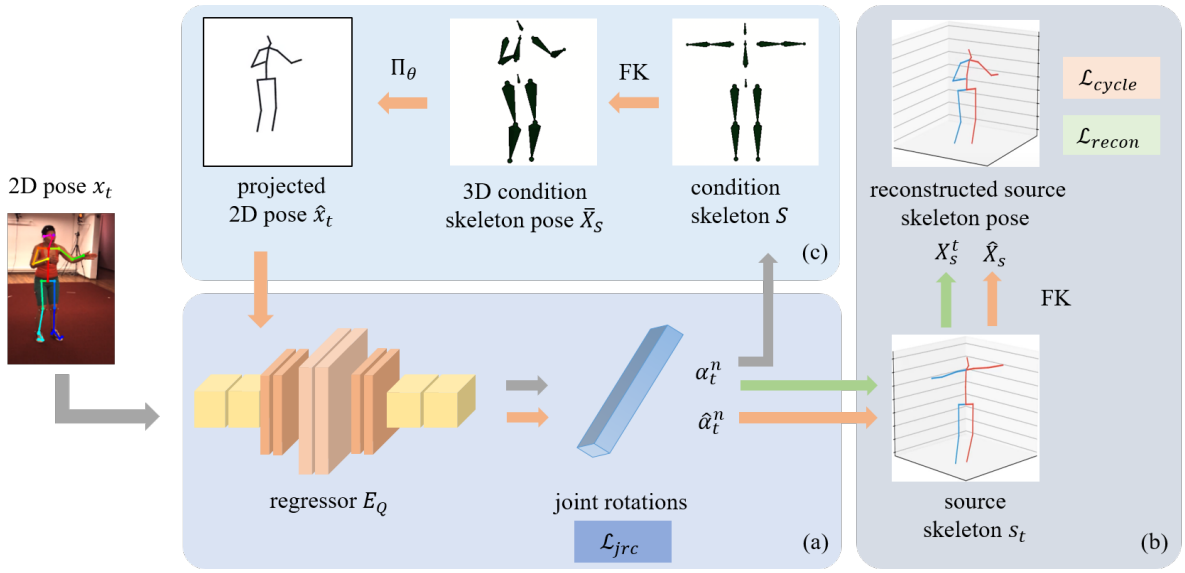

Figure 3. Our model for pose normalization. (a) The network E_Q regresses 3D joint rotations from input 2D poses; (b) reconstruction branch: apply joint rotations on source skeletons for 3D pose reconstruction for training; (c) cycle reconstruction branch: apply joint rotations on the condition skeleton, and then project to 2D as input for cycle consistency.

from input 2D poses computed by a convolutional neural network E_Q onto a condition skeleton to normalize the poses (as shown by the gray data path in Figure 3(a)(c)). For a video containing T frames, we denote the 2D position of joint n at frame t as $x_t^n \in \mathbf{R}^2$, $t = 1, 2, ..., T$, $n = 1, 2, ...N$, where N is the total number of joints. The joint rotations, computed from the input 2D poses by E_Q, are represented as a unit quaternion for each joint $\alpha_t^n \in \mathbf{R}^4$.

Denote the FK process as $X = FK(s, \alpha)$, where bones in a skeleton s are rotated according to a set of joint rotations α, resulting in the 3D joint positions X of the skeleton. In order to train E_Q, we apply the joint rotations α_t^n computed from 2D poses x_t^n on two types of skeletons: the condition skeleton S, to facilitate the learning of pose normalization; and the source 3D skeletons of the video subjects s_t computed from the ground-truth 3D poses, to assist the training of E_Q. In the following, we describe the two FK branches, namely reconstruction branch and cycle reconstruction branch, for training E_Q.

Reconstruction Branch. The reconstruction branch is shown as the green data path in Figure 3(a)(b). Applying α_t^n on the source 3D skeletons s_t results in reconstructed 3D poses X_s^t, which can be directly supervised by the ground-truth 3D poses X_t^n using the reconstruction loss:

$$\mathcal{L}_{recon} = \sum_{t,n} \|FK(s_t, \alpha_t^n) - X_t^n\|^2.$$

Cycle Reconstruction Branch. The cycle reconstruction branch is represented by the orange data path in Figure 3. The design of the cycle reconstruction branch is based on the observation that, since the condition skeleton has the same poses as the poses of the subject in the video frames, its projections should yield the same 3D joint rotations as the 3D joint rotations produced by the original input 2D poses. However, applying rotations α_t^n on the condition skeleton results in new 3D poses $\bar{X}_S$ without paired ground-truth for supervised training. Thus, we adopt this cycle reconstruction that projects the 3D condition skeleton pose into 2D pose, and then compute 3D joint rotations from the projected 2D poses.

Specifically, the 3D condition skeleton poses $\bar{X}_S$ are projected by ground-truth camera parameters into the 2D poses $\hat{x}_t^n$, which are then input to E_Q to compute the joint rotations $\hat{\alpha}_t^n$ of the projected poses of the condition skeleton. These rotations are applied again to the ground-truth skeleton by FK, resulting in cycle reconstructed 3D poses. This process produces two constraints for training. The joint rotation consistency loss is computed as the differences between joint rotations from the input 2D poses and from the projected skeleton poses:

$$\mathcal{L}_{jrc} = \sum_{t,n} \|\alpha_t^n - \hat{\alpha}_t^n\|^2.$$

The cycle reconstruction loss is:

$$\mathcal{L}_{cycle} = \sum_{t,n} \|FK(s_t, \hat{\alpha}_t^n) - X_t^n\|^2.$$

Besides the above losses, we also adopted the foot contact loss $\mathcal{L}_{fc}$, which is commonly used in 3D pose estimations to reduce the skating effect [46]. The total loss function for training is:

$$\mathcal{L} = \mathcal{L}_{recon} + \varphi\mathcal{L}_{cycle} + \beta\mathcal{L}_{jrc} + \lambda\mathcal{L}_{fc},$$

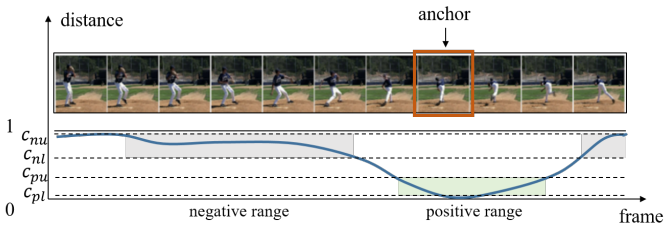

Figure 4. Adaptive triplet mining: determining the positive and negative candidate ranges based on primary pose similarities.

where φ, β and λ are weights for the cycle reconstruction loss, the joint rotation consistency loss and the foot contact loss, respectively.

The reconstruction branch and the cycle reconstruction branch are only used for training. For inference, only the regressor E_Q is used to compute the 3D joint rotations from 2D poses to be applied onto the condition skeleton. To unify the global orientations of the poses, the quaternion for the root joint is set to a specific rotation value (in our experiments $[1, 0, 0, 0]$) so as to rotate the poses to the same orientation.

3.3. Pose Embedding

After the poses in the video frames are normalized into unified bone lengths and viewpoints, we use metric learning to map the poses to a pose embedding space to extract high-level pose features. Specifically, we use another neural network E_P to extract features from the normalized 3D poses and train E_P with a triplet loss [43]. We experimented on three types of networks as the feature extractor, including fully connected [28], GCN [56], and PointNet [39], and have empirically found that fully connected performs better than the other two alternatives (see the ablation study in Section 4.4).

To train the feature extractor by metric learning, an important consideration is the definition of positive and negative pair for each anchor 3D pose. Existing works on human pose related metric learning often make use of action coherence in videos [47, 2, 44], and sample the positive pair at a fixed temporal offset of the anchor, while the negative pair outside a temporal window of the anchor or from another action. However, since actions in different videos are performed at various speeds, and some poses are repetitive throughout an action, a fixed temporal offset might not generalize well to all videos.

We thus propose to adaptively mine triplets within a video. As shown in Figure 4, for each frame in the video (anchor frame), we compute a primary similarity with all other frames, by measuring the Mean Per Joint Position Error (MPJPE) between the normalized poses in the anchor frame and in other frames. These primary similarities are linearly normalized to $[0, 1]$, such that setting thresholds $c. \in [0, 1]$ divides the primary similarities into positive ($[c_{pl}, c_{pu}]$) and negative ($[c_{nl}, c_{nu}]$) candidate ranges from which triplets are sampled. Adaptive mining also facilitates curriculum learning [3], where the training starts from easy negative pairs and gradually shifts to semi-hard pairs. The difficulty level of triplets can be easily modulated by setting the thresholds.

4. Experiments

We trained our networks on the training set of Human3.6M dataset [22], which provides ground-truth 2D poses, 3D poses, and camera parameters. The ground-truth source skeleton can be computed from the ground-truth 3D poses. The bone lengths of the condition skeleton is defined as the average bone lengths in the Human3.6M training set. Please refer to the supplementary material for more implementation details.

4.1. Robustness to Anthropometry

Dataset. To produce variations in subjects' anthropometry, we augment Human3.6M dataset with different skeletons. With raw joint angles and bone lengths in the dataset, for every 50 frames in each video, we multiply the original bone lengths by a random scale factor from the range $[0.75, 1.25]$ (skeleton symmetry preserved), and compute the new ground-truth 2D and 3D joint positions following the original construction process of Human3.6M dataset.

Baselines. We experimented on three state-of-the-art methods [28, 56, 37] that estimate 3D poses from 2D poses. Their models and our model were all re-trained using the respective original settings (e.g. pre-processings of 2D and 3D poses, number of epochs, etc.)

Metrics. The goal of this experiment is to measure the changes in the 3D poses when there are variations in the 2D poses caused by bone length variations. As illustrated in Figure 5, suppose x and Δx represents the input 2D poses and variations in 2D poses, and y and Δy are the corresponding output 3D poses and the variations in the output. The model is robust when Δy is close to zero at the presence of Δx. We thus define the metric as the mean errors in reconstructed 3D poses with and without bone length variations in input 2D poses, denoted as Δ_{MPJPE}.

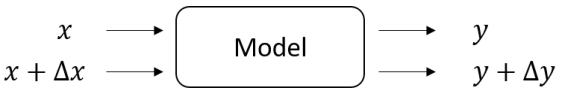

Figure 5. An illustration of variations in input and output of a model.

We also report the MPJPE under Protocol #1 [49]. The MPJPE for baseline methods are all from the original papers. The MPJPE for our method is computed by applying FK using the estimated joint rotations on the ground-truth source skeletons from the Human3.6M test set.

Method	MPJPE ↓	Δ_{MPJPE} ↓
Martinez [28]	45.50	58.27
SemGCN [56]	40.78	61.07
VideoPose3D [37]	**37.20**	58.40
Ours	52.61	**53.35**

Table 1. Results of pose reconstruction accuracy and robustness to variations in anthropometry. (unit:mm)

Results. The results are shown in Table 1. The MPJPE scores of the 3D pose estimation methods [28, 56, 37] outperforms our reconstructed poses, partly because their training is directly supervised by the joint position errors, while ours need to satisfy the FK constraint. However, the FK constraint in our method helps to distinguish the variations in anthropometry from the variations of joint rotations, and thus our methods is more robust to anthropometry, leading to the lowest value of Δ_{MPJPE}.

4.2. Robustness to Viewpoint

To evaluate the effectiveness of normalization in viewpoints, we visualized the normalized poses in video frames, as shown in Figure 6. We also experimented on an alternative viewpoint normalization method, which first estimates the 3D poses given the detected 2D poses from the video frames [28], and then aligns the poses with a pre-defined T-pose in a fixed global orientation by Procrustes alignment. As shown in Figure 6(c), the estimated 3D poses transformed to the world coordinates by ground-truth camera parameters are close to the ground-truth 3D poses. However, when there is no ground-truth camera parameters, the joint positions of the 3D poses are dependent on viewpoints (Figure 6(d)). Applying rigid transformations can only roughly align them to the same global orientation (Figure 6(e)), which would still have a large impact on joint positions. In contrast, our method can both accurately capture the poses in the video frames (Figure 6(f)), and effectively transform the poses to a unified orientation (Figure 6(g)).

4.3. Dense Correspondence

To evaluate our pose feature in a video alignment task, we design an experiment on finding dense correspondence between pairs of human action videos.

Dataset. To the best of our knowledge, there is no existing dataset of videos with densely labeled ground-truth correspondences, since manual annotation of such labeling would be extremely laborious. We thus make use of the synchronized multiview videos in the Human3.6M test set to build a synthesized correspondence dataset. For each of the 59 actions in the Human3.6M test set, we take the two frontal viewpoint videos as a pair of source and target videos, which are originally strictly aligned temporally. To produce the difference in lengths between the source and target videos, each frame in videos is randomly retained or dropped with a probability of $p = 0.5$, while a temporal filter is applied to ensure that no five consecutive frames are dropped together, to ensure the realism of the reconstructed video dataset. Then new correspondences between source and target frames are constructed by applying a Dynamic Time Warping (DTW) [4] on the indexes of retained frames. These correspondences will be used as the ground-truth in the dense correspondence task. Please refer to the supplementary material for more details of dataset construction.

Metrics. We design the task of finding dense correspondence on this synthesized dataset as follows: for each frame in the target video, retrieve the index of the corresponding frame in the source video based on the similarity of pose features extracted at each frame. Two evaluation metrics are defined over the constructed dataset: (a) hit ratio: the percentage of frames that have the retrieved source frame indexed within a small temporal threshold ($\tau = 5$ in our implementation) of the ground-truth indexes; (b) mean square error (MSE): the mean square error on temporal distances in the source video between the retrieved and ground-truth correspondences.

Baselines. The baselines we compared with include two categories: (1) existing human pose features, such as pose features from 2D poses [10, 48], SMPL-based pose parameters [24], and latent representations captured by networks [56], and (2) other alternatives of our current methods, such as using 3D joint positions or angles from 3D poses. Except SMPL [24], which requires video frames as input, all the other methods we compared with take a 2D pose sequence as input and output a sequence of features.

Results. The results are shown in Table 2. Matching with Euler angles from 3D poses outperforms that from 2D poses by a large margin. Our pose feature outperforms the features learned from 2D poses [48] and SMPL-based pose features in terms of both metrics. In this experiment, using L2 distances of normalized poses to represent pose similarity achieves a significantly better performance than using learned pose features, because each pair of videos record the poses from the same subject, and thus the reconstructed poses are supposed to be the same.

Figure 7 shows a visualization of video dense correspondence on in-the-wild videos. For each pair of videos we selected eight representative keyframes of the action in the source video (upper row), and retrieve their corresponding keyframes in the target video (lower row) by applying the DTW with our pose features. Even though the subjects are in different appearances and orientations, the retrieved frames are correspondent with keyframes in terms of the poses to the action.

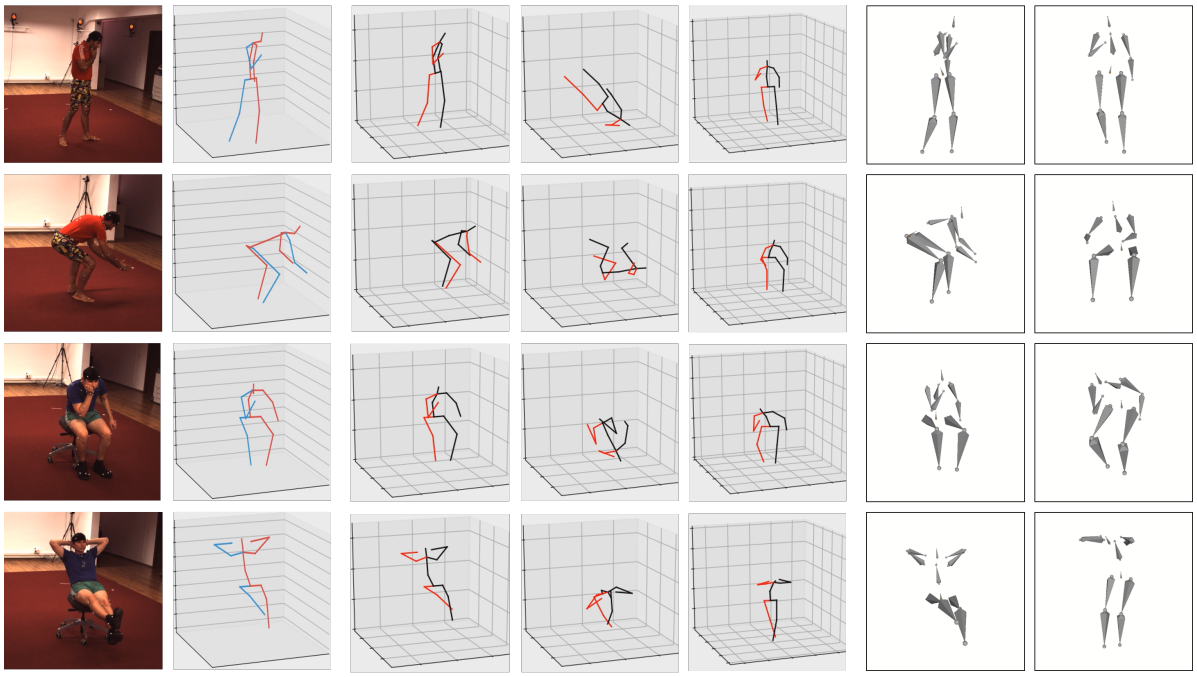

Figure 6. Visualization of pose normalization results. (a) The video frames from the Human3.6M test set; (b) ground-truth 3D poses in world coordinates; (c) estimated 3D poses by Martinez [28] in world coordinates; (d) 3D joint positions in camera coordinates; (e) unify 3D poses in camera coordinates by Procrustes alignment with a pre-defined T-pose; (f) 3D condition skeleton poses by our method; (g) our normalized 3D poses under a unified global orientation.

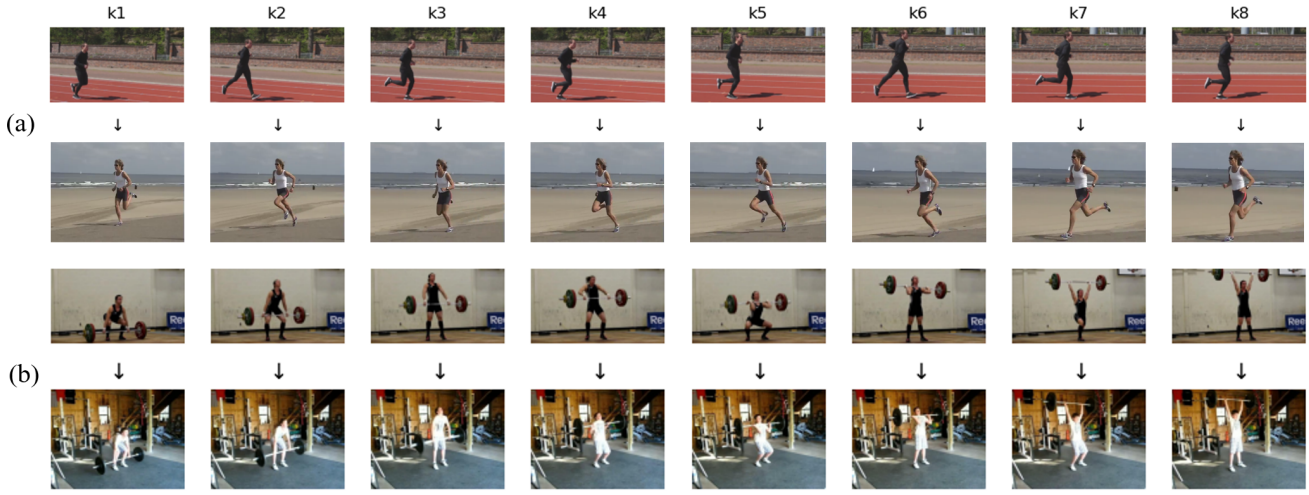

Figure 7. Visualization of dense correspondence on running (a) and weight-lifting (b) videos.

4.4. Ablation Study

We conducted ablative experiments on our dense correspondence dataset to verify the importance of individual system components and our choices as shown in Table 3. We compared three types of neural networks for pose embedding (the second step of our pipeline), including fully-connected layers [28], PointNet for global features [39], and graph convolutional network [56], to determine which best captures the pose features. We have found out that the performance when using fully-connect layers as the encoder network outperforms the other two. We also tested alternative configurations, including (1) joint training of E_Q and E_P; (2) adding a temporal convolution module to extract features from a small temporal window to include motion

Method	Hit Ratio (%) ↑	MSE ↓
Euler+2D Poses [10]	33.20	81.20
Euler+3D Poses [28]	66.92	16.63
PA 3D Poses [28]	53.18	16.55
SemGCN latent [56]	41.86	28.34
POEM [48]	59.88	16.90
SMPL [24]	69.04	13.99
Ours (Normalized Poses)	**94.25**	**2.35**
Ours (Pose Features)	74.95	12.71

Table 2. Accuracy of hit ratio and MSE on the finding dense correspondences task.

Model	Hit Ratio (%) ↑	MSE ↓
SemGCN	61.05	15.9751
PointNet	68.96	13.8574
FC	**74.95**	**12.7106**
joint training	25.76	30.9361
temporal window	44.81	25.8550
w/o normalization	53.69	17.6033
w/o adaptive mining	41.92	28.7598

Table 3. Ablation study on the dense correspondence task.

features instead of a single pose; (3) learn features from Procrustes-aligned 3D poses without normalizations, and (4) using fixed-sized temporal windows for triplet mining. All of these alternatives result in inferior performance to our current configurations.

4.5. Action Recognition

To verify how well our proposed pose features work in other human action video analysis tasks, we conduct experiments on the task of unsupervised human action recognition by matching [16]. Since the pose features are computed on a frame-by-frame basis, a temporal encoding is needed to aggregate the pose features to further describe actions. We adopted rank pooling [15] as the temporal encoder, as described in [16].

We experimented on the Penn Action dataset [55]. For each video in the dataset, we first computed the per-frame features using either our method or other baseline methods; then the feature sequence was encoded as a fixed-length vector by rank pooling; finally, the samples in the test set were classified using the vectors by a K-Nearest Neighbor (k-NN) matching with the vectors in the training set. We did not re-train on the Penn Action dataset; instead we re-trained our models on the Human3.6M dataset with augmented virtual cameras [27], and pre-process the input 2D poses according to the original implementations. Except TCC [12], which takes video frames as input and EnGAN [26], which takes 3D poses as input (computed by [28]), all the other methods take 2D poses as input, and thus can directly apply the pre-trained models to the Penn Action dataset. The results of action recognition accuracy by 1-NN and 5-NN are shown in Table 4.

Methods	Top-1 (%) ↑	Top-5 (%) ↑
Euler+2D Poses [10]	52.53	79.77
PA 3D Poses [28]	50.37	74.34
Pose Perceptual [40]	74.48	87.55
TCC [12]	15.07	43.82
EnGAN [26]	53.18	68.91
SemGCN latent [56]	49.34	72.19
Ours (Normalized Poses)	54.12	75.19
Ours (Pose Features)	**75.66**	**88.58**

Table 4. The accuracy of action recognition task on the Penn Action dataset.

Our pose feature outperforms most of the baseline methods by a large margin on this task. The pose perceptual feature [40] also achieves a comparable performance with ours. It involves poses from neighboring frames and thus also capture motion features. While the pose perceptual features reserve most network output at hidden layers (on average 12,672 parameters for each pose), our feature is a more compact representation (64 parameters each).

5. Conclusion and Future Work

In this paper, we proposed a normalized human pose feature for video alignment. A novel pose normalization method has been proposed to obtain normalized poses from video frames invariant to viewpoints and subjects' physical structures. In addition, an adaptive triplet mining strategy has been proposed to make the metric learning using poses from videos more robust to action speeds. Experiments on the video dense correspondence task and the action recognition task show that our proposed feature outperforms human pose features by the state-of-the-art techniques.

Our current method has a limitation in that it extracts features from a complete normalized pose. Future work includes modeling the normalized pose with partial observations. A potential solution that worth exploring is to adopt a probabilistic modeling with kinematics constraints [46, 23] as priors for the FK layer in the network, such that the missing joint positions would be filled by satisfying both kinematics priors and non-missing joint positions. Another limitation of our method is that it requires ground-truth 3D joint positions in training. Adopting weakly-supervised settings, such as using ordinal depths of joint pairs [36, 42] might enable training on in-the-wild datasets.

Acknowledgment

We sincerely thank the anonymous reviewers for their insightful comments and suggestions.

References

[1] Kfir Aberman, Rundi Wu, Dani Lischinski, Baoquan Chen, and Daniel Cohen-Or. Learning character-agnostic motion for motion retargeting in 2d. *ACM Transactions on Graphics (TOG)*, 38(4):1–14, 2019. 3

[2] Andreas Aristidou, Daniel Cohen-Or, Jessica K Hodgins, Yiorgos Chrysanthou, and Ariel Shamir. Deep motifs and motion signatures. *ACM Transactions on Graphics (TOG)*, 37(6):1–13, 2018. 3, 5

[3] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48, 2009. 5

[4] Donald J Berndt and James Clifford. Using dynamic time warping to find patterns in time series. In *KDD workshop*, volume 10, pages 359–370. Seattle, WA, USA:, 1994. 6

[5] Aleksandar Bojchevski and Stephan Günnemann. Deep gaussian embedding of graphs: Unsupervised inductive learning via ranking. In *International Conference on Learning Representations*, pages 1–13, 2018. 3

[6] Kaidi Cao, Jingwei Ji, Zhangjie Cao, Chien-Yi Chang, and Juan Carlos Niebles. Few-shot video classification via temporal alignment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10618–10627, 2020. 3

[7] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017. 3

[8] Cheng Chen, Yueting Zhuang, Feiping Nie, Yi Yang, Fei Wu, and Jun Xiao. Learning a 3d human pose distance metric from geometric pose descriptor. *IEEE Transactions on Visualization and Computer Graphics*, 17(11):1676–1689, 2010. 2, 3

[9] Xipeng Chen, Kwan-Yee Lin, Wentao Liu, Chen Qian, and Liang Lin. Weakly-supervised discovery of geometry-aware representation for 3d human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10895–10904, 2019. 3

[10] Myung Geol Choi, Kyungyong Yang, Takeo Igarashi, Jun Mitani, and Jehee Lee. Retrieval and visualization of human motion data via stick figures. In *Computer Graphics Forum*, volume 31, pages 2057–2065. Wiley Online Library, 2012. 2, 3, 6, 8

[11] Junting Dong, Qing Shuai, Yuanqing Zhang, Xian Liu, Xiaowei Zhou, and Hujun Bao. Motion capture from internet videos. In *European Conference on Computer Vision*, pages 210–227. Springer, 2020. 1, 3

[12] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. Temporal cycle-consistency learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1801–1810, 2019. 3, 8

[13] Marcin Eichner, Manuel Marin-Jimenez, Andrew Zisserman, and Vittorio Ferrari. 2d articulated human pose estimation and retrieval in (almost) unconstrained still images. *International journal of computer vision*, 99(2):190–214, 2012. 3

[14] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. Rmpe: Regional multi-person pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2334–2343, 2017. 3

[15] Basura Fernando, Efstratios Gavves, José Oramas, Amir Ghodrati, and Tinne Tuytelaars. Rank pooling for action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 39(4):773–787, 2016. 8

[16] Basura Fernando, Sareh Shirazi, and Stephen Gould. Unsupervised human action detection by action matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–9, 2017. 2, 8

[17] Keith Grochow, Steven L Martin, Aaron Hertzmann, and Zoran Popović. Style-based inverse kinematics. In *ACM SIGGRAPH 2004 Papers*, pages 522–531. 2004. 2

[18] Ikhsanul Habibie, Weipeng Xu, Dushyant Mehta, Gerard Pons-Moll, and Christian Theobalt. In the wild human pose estimation using explicit 2d features and intermediate 3d representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10905–10914, 2019. 3

[19] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 1735–1742, 2006. 3

[20] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020. 3

[21] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017. 3

[22] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, 2014. 5

[23] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7122–7131, 2018. 3, 8

[24] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5253–5263, 2020. 3, 6, 8

[25] Lucas Kovar, Michael Gleicher, and Frédéric Pighin. Motion graphs. In *ACM SIGGRAPH 2008 classes*, pages 1–10. 2008. 3

[26] Jogendra Nath Kundu, Maharshi Gor, Phani Krishna Uppala, and R Venkatesh Babu. Unsupervised feature learning of human actions as trajectories in pose embedding manifold. *arXiv preprint arXiv:1812.02592*, 2018. 8

[27] Lei Li, Siyu Zhu, Hongbo Fu, Ping Tan, and Chiew-Lan Tai. End-to-end learning local multi-view descriptors for 3d point clouds. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 8

[28] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3d human pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2640–2649, 2017. 1, 3, 5, 6, 7, 8

[29] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Mohamed Elgharib, Pascal Fua, Hans-Peter Seidel, Helge Rhodin, Gerard Pons-Moll, and Christian Theobalt. Xnect: Real-time multi-person 3d motion capture with a single rgb camera. *ACM Transactions on Graphics (TOG)*, 39(4):82–1, 2020. 3

[30] Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt. Vnect: Real-time 3d human pose estimation with a single rgb camera. volume 36, 2017. 3

[31] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013. 3

[32] Anastasiia Mishchuk, Dmytro Mishkin, Filip Radenovic, and Jiri Matas. Working hard to know your neighbor's margins: Local descriptor learning loss. In *Advances in Neural Information Processing Systems*, pages 4826–4837, 2017. 3

[33] Ishan Misra, C Lawrence Zitnick, and Martial Hebert. Shuffle and learn: unsupervised learning using temporal order verification. In *European Conference on Computer Vision*, pages 527–544. Springer, 2016. 3

[34] Greg Mori, Caroline Pantofaru, Nisarg Kothari, Thomas Leung, George Toderici, Alexander Toshev, and Weilong Yang. Pose embeddings: A deep architecture for learning to match human poses. *arXiv preprint arXiv:1507.00302*, 2015. 3

[35] Eshed Ohn-Bar and Mohan Trivedi. Joint angles similarities and hog2 for action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 465–470, 2013. 2

[36] Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Ordinal depth supervision for 3d human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7307–7316, 2018. 8

[37] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7745–7754, 2019. 1, 5, 6

[38] Dario Pavllo, David Grangier, and Michael Auli. Quaternet: A quaternion-based recurrent model for human motion. *arXiv preprint arXiv:1805.06485*, 2018. 2, 3

[39] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. *arXiv preprint arXiv:1612.00593*, 2016. 5, 7

[40] Xuanchi Ren, Haoran Li, Zijian Huang, and Qifeng Chen. Self-supervised dance video synthesis conditioned on music. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 46–54, 2020. 2, 3, 8

[41] Helge Rhodin, Mathieu Salzmann, and Pascal Fua. Unsupervised geometry-aware representation for 3d human pose estimation. In *Proceedings of the European Conference on Computer Vision*, pages 750–767, 2018. 3

[42] Matteo Ruggero Ronchi, Oisin Mac Aodha, Robert Eng, and Pietro Perona. It's all relative: Monocular 3d human pose estimation from weakly supervised data. In *BMVC*, 2018. 8

[43] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. 3, 5

[44] Pierre Sermanet, Corey Lynch, Yevgen Chebotar, Jasmine Hsu, Eric Jang, Stefan Schaal, Sergey Levine, and Google Brain. Time-contrastive networks: Self-supervised learning from video. In *2018 IEEE International Conference on Robotics and Automation*, pages 1134–1141. IEEE, 2018. 3, 5

[45] Mingyi Shi, Kfir Aberman, Andreas Aristidou, Taku Komura, Dani Lischinski, Daniel Cohen-Or, and Baoquan Chen. Motionet: 3d human motion reconstruction from monocular video with skeleton consistency. *ACM Transactions on Graphics (TOG)*, 40(1):1–15, 2020. 2, 3

[46] Soshi Shimada, Vladislav Golyanik, Weipeng Xu, and Christian Theobalt. Physcap: Physically plausible monocular 3d motion capture in real time. *ACM Transactions on Graphics (TOG)*, 39(6), dec 2020. 2, 4, 8

[47] Omer Sumer, Tobias Dencker, and Bjorn Ommer. Self-supervised learning of pose embeddings from spatiotemporal relations in videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4298–4307, 2017. 3, 5

[48] Jennifer J Sun, Jiaping Zhao, Liang-Chieh Chen, Florian Schroff, Hartwig Adam, and Ting Liu. View-invariant probabilistic embedding for human pose. In *European Conference on Computer Vision*, pages 53–70. Springer, 2020. 1, 3, 6, 8

[49] Xiao Sun, Jiaxiang Shang, Shuang Liang, and Yichen Wei. Compositional human pose regression. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2602–2611, 2017. 3, 5

[50] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. Integral human pose regression. In *Proceedings of the European Conference on Computer Vision*, pages 529–545, 2018. 3

[51] Yifan Sun, Changmao Cheng, Yuhan Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang, and Yichen Wei. Circle loss: A unified perspective of pair similarity optimization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6398–6407, 2020. 3

[52] Raviteja Vemulapalli, Felipe Arrate, and Rama Chellappa. Human action recognition by representing 3d skeletons as points in a lie group. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 588–595, 2014. 2

[53] Ruben Villegas, Jimei Yang, Duygu Ceylan, and Honglak Lee. Neural kinematic networks for unsupervised motion

retargetting. In *The IEEE Conference on Computer Vision and Pattern Recognition*, June 2018. 2, 3

[54] Jiang Wang, Yang Song, Thomas Leung, Chuck Rosenberg, Jingbin Wang, James Philbin, Bo Chen, and Ying Wu. Learning fine-grained image similarity with deep ranking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1386–1393, 2014. 3

[55] Weiyu Zhang, Menglong Zhu, and Konstantinos G Derpanis. From actemes to action: A strongly-supervised representation for detailed action understanding. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2248–2255, 2013. 8

[56] Long Zhao, Xi Peng, Yu Tian, Mubbasir Kapadia, and Dimitris N. Metaxas. Semantic graph convolutional networks for 3d human pose regression. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3425–3435, 2019. 5, 6, 7, 8

[57] Feng Zhou, Fernando De la Torre, and Jessica K Hodgins. Hierarchical aligned cluster analysis for temporal clustering of human motion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(3):582–596, 2012. 2

[58] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5745–5753, 2019. 3