Fusing Wearable IMUs with Multi-View Images for Human Pose Estimation: A Geometric Approach

Zhe Zhang[1]*   Chunyu Wang[2]   Wenhu Qin[1]   Wenjun Zeng[2]

[1]Southeast University, Nanjing, China   [2]Microsoft Research Asia, Beijing, China

Abstract

We propose to estimate 3D human pose from multi-view images and a few IMUs attached at person's limbs. It operates by firstly detecting 2D poses from the two signals, and then lifting them to the 3D space. We present a geometric approach to reinforce the visual features of each pair of joints based on the IMUs. This notably improves 2D pose estimation accuracy especially when one joint is occluded. We call this approach Orientation Regularized Network (ORN). Then we lift the multi-view 2D poses to the 3D space by an Orientation Regularized Pictorial Structure Model (ORPSM) which jointly minimizes the projection error between the 3D and 2D poses, along with the discrepancy between the 3D pose and IMU orientations. The simple two-step approach reduces the error of the state-of-the-art by a large margin on a public dataset. Our code will be released at https://github.com/microsoft/imu-human-pose-estimation-pytorch.

1. Introduction

Estimating 3D poses from images has been a longstanding goal in computer vision. With the development of deep learning models, the recent approaches [5, 17, 2, 20, 21, 26] have achieved promising results on the public datasets. One limitation of the vision-based methods is that they cannot robustly solve the occlusion problem.

A number of works are devoted to estimating poses from wearable sensors such as IMUs [27, 22, 29, 30]. They suffer less from occlusion since IMUs can provide direct 3D measurements. For example, Roetenberg *et al.* [22] place 17 IMUs with 3D accelerometers, gyroscopes and magnetometers at the rigid bones. If the measurements are accurate, the 3D pose is fully determined. In practice, however, the accuracy is limited by a number of factors such as calibration errors and the drifting problem.

Recently, fusing images and IMUs to achieve more robust pose estimation has attracted much attention [27, 28,
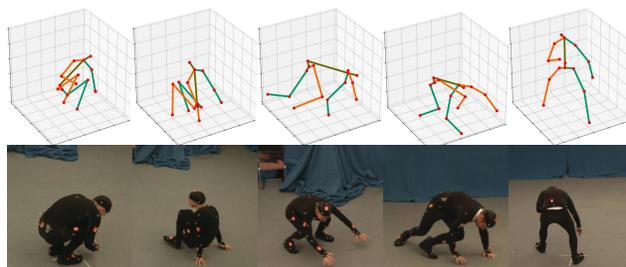

Figure 1. Our approach gets accurate 3D pose estimations even when severe *self-occlusion* occurs in the images.

6, 15]. They mainly follow a similar framework of building a parametric 3D human model and optimizing its parameters to minimize its discrepancy with the images and IMUs. The accuracy of these approaches is limited mainly due to the hard optimization problem.

We present an approach to fuse IMUs with images for robust pose estimation. It gets accurate estimations even when occlusion occurs (see Figure 1). In addition, it outperforms the previous methods [15, 28] by a notable margin on the public dataset. We first introduce Orientation Regularized Network (ORN) to *jointly* estimate 2D poses for multi-view images as shown in Figure 2. ORN differs from the previous multiview methods [19] in that it uses IMU orientations as a structural prior to *mutually* fuse the image features of each pair of joints linked by IMUs. For example, it uses the features of the elbow to reinforce those of the wrist based on the IMU at the lower-arm.

The cross-joint-fusion allows to accurately localize the occluded joints based on their neighbors. The main challenge is to determine the relative positions between each pair of joints in the images, which we solve elegantly in the 3D space with the help of IMU orientations. The approach significantly improves the 2D pose estimation accuracy especially when occlusion occurs.

In the second step, we estimate 3D pose from multi-view 2D poses (heatmaps) by a Pictorial Structure Model (PSM) [12, 17, 2]. It jointly minimizes the projection error between the 3D and 2D poses, along with the discrepancy be-

---

*Work done when Zhe Zhang is an intern at Microsoft Research Asia.

tween the 3D pose and the prior. The previous works such as [17, 19] often use the limb length prior to prevent from generating abnormal 3D poses. This prior is fixed for the same person and does not change over time. In contrast, we introduce an orientation prior that requires the limb orientations of the 3D pose to be consistent with the IMUs. The prior is complementary to the limb length and can reduce the negative impact caused by inaccurate 2D poses. We call this approach Orientation Regularized Pictorial Structure Model (ORPSM).

We evaluate our approach on two public datasets including Total Capture [27] and H36M [9]. On both datasets, *ORN* notably improves the 2D estimation accuracy especially for the frequently occluded joints such as ankle and wrist, which in turn decreases the 3D pose error. Take the Total Capture dataset as an example, on top of the 2D poses estimated by *ORN*, *ORPSM* obtains a 3D position error of 24.6mm which is much smaller than the previous state-of-the-art [19] (29mm) on this dataset. This result demonstrates the effectiveness of our visual-inertial fusion strategy. To validate the general applicability of our approach, we also experiment on the H36M dataset which has different poses from the Total Capture dataset. Since it does not provide IMUs, we synthesize virtual limb orientations and only show proof-of-concept results.

2. Related Work

Images-based We classify the existing image-based 3D pose estimation methods into three classes. The first class is model/optimization based [5, 13] which defines a 3D parametric human body model and optimizes its parameters to minimize the discrepancy between model projections and extracted image features. These approaches mainly differ in terms of the used image features and optimization algorithms. These methods generally suffer from the difficult non-convex optimization which limits the 3D estimation accuracy to a large extent in practice.

With the development of deep learning, some approaches such as [20, 21, 16, 26, 10, 18] propose to learn a mapping from images to 3D pose in a supervised way. The lack of abundant ground truth 3D poses is their biggest challenge for achieving desired performance on wild images. Zhou *et al*. [33] propose a multi-task solution to leverage the abundant 2D pose datasets for training. Yang *et al*. [32] use adversarial training to improve the robustness of the learned model. Another limitation of this type of methods is that the predicted 3D poses by these methods are relative to their pelvis joints. So they are not aware of their absolute locations in the world coordinate system.

The third class of methods such as [1, 3, 17, 2, 7, 11, 4, 19] adopt a two-step framework. It first estimates 2D poses in each camera view and then recovers the 3D pose in a world coordinate system with the help of camera pa-

rameters. For example, Tome *et al*. [25] build a 3D pictorial model and optimize the 3D locations of the joints such that their projections match the detected 2D pose heatmaps and meanwhile the spatial configuration of the 3D joints matches the prior pose structure. Qiu *et al*. [19] propose to first estimate 2D poses for every camera view, and then estimate the 3D pose by triangulation or by pictorial structure model. This type of approaches has achieved the state-of-the-art accuracy due to the significantly improved 2D pose estimation accuracy.

IMUs-based There are a small number of works which attempt to recover 3D poses using only IMUs. For example, Slyper *et al*. [23] and Tautges *et al*. [24] propose to reconstruct human pose from 5 accelerometers by retrieving pre-recorded poses with similar accelerations from a database. They get good results when the test sequences are present in the training dataset. Roetenberg *et al*. [22] use 17 IMUs equipped with 3D accelerometers, gyroscopes and magnetometers and all the measurements are fused using a Kalman Filter. By achieving stable orientation measurements, the 17 IMUs can fully define the pose of the subject. Marcard *et al*. [30] propose to exploit a statistical body model and jointly optimize the poses over multiple frames to fit orientation and acceleration data. One disadvantage of the IMUs-only methods is that they suffer from drifting over time, and need a large amount of careful engineering work in order to make it work robustly in practice.

"Images+IMUs"-based Some works such as [29, 27, 28, 6, 15] propose to combine images and IMUs for robust 3D human pose estimation. The methods can be categorized into two classes according to how image-inertial fusion is performed. The first class [15, 28, 29] estimate 3D human pose by minimizing an energy function which is related to both IMUs and image features. The second class [27, 6] estimate 3D poses separately from the images and IMUs, and then combine them to get the final estimation. For example, Trumble *et al*. [27, 6] propose a two stream network to concatenate the pose embeddings separately derived from images and IMUs for regressing the final pose.

Although the simple two-step framework has achieved the state-of-the-art performance in the image only setting, it is *barely* studied for "IMU+images"-based pose estimation because it is nontrivial to leverage IMUs in the two steps. Our main contribution lies in proposing two novel ways of exploiting IMUs in the framework. More importantly, we empirically show that this simple two-step approach can significantly outperform the previous state-of-the-arts.

Our work differs from the previous works [27, 28, 6, 15, 14] in two-fold. First, instead of estimating 3D poses or pose embeddings from images and IMUs separately and then fusing them in a late stage, we propose to fuse IMUs
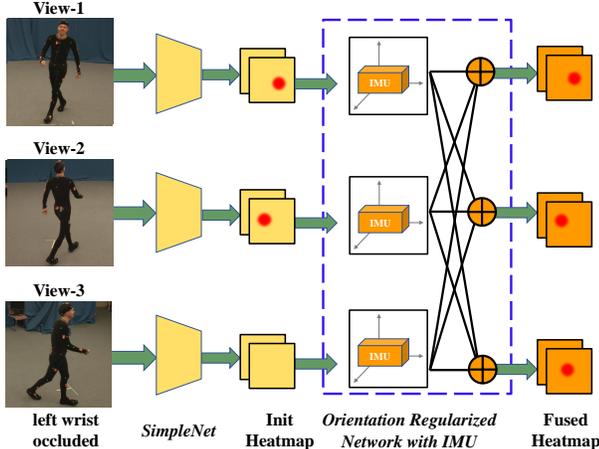
Figure 2. Overview of ORN. It firstly takes multi-view images as input and estimates initial heatmaps (based on SimpleNet [31]) independently for each camera view. Then with the aid of IMU orientations, it mutually fuses the heatmaps of the linked joints across all views. It enforces supervision on both initially estimated and fused heatmaps during the end-to-end training.

and image features in a very early stage with the aid of 3D geometry. This directly gives improved 2D poses rather than attempting to get accurate poses from two inaccurate ones as in late fusion. Second, in the 3D pose estimation step, we leverage IMUs in the pictorial structure model. Although pictorial model is not new, the effect of using IMUs has not been discussed. Finally, we hope this simple yet effective approach could promote more research in the two-step pose estimation direction.

3. ORN for 2D Pose Estimation

We represent a 2D pose by a graph which consists of M joints $\mathcal{J} = \{J_1, J_2, \cdots, J_M\}$ and N edges $\mathcal{E} = \{e_1, e_2, \cdots, e_N\}$ as shown in Figure 3 (c). Each J represents the state of a joint such as its 2D location in the image. Each edge e connects two joints, representing their conditional dependence. In this work, we attach IMUs to W limbs to obtain their 3D orientations $\mathcal{O} = \{o_1, o_2, \cdots, o_W\}$. This orientation information will be used to constrain relative positions between two joints. In the following, we will describe in detail how we estimate 2D poses with the help of orientations.

3.1. Methodology

We start by describing how 3D limb orientations can be used to mutually enhance the features between pairs of joints linked by IMUs in the *same* camera view. Then we extend it to handle multi-view features.

Same-View Fusion We explain the main idea of our approach with a pair of joints J_1 and J_2 as an example. The two joints are connected by the limb e whose 3D orientation is o. In practice, we will apply the fusion operation to all pairs of joints linked by IMUs. Figure 3 sketches the idea. Let the heatmaps of J_1 and J_2 be H_1 and H_2, respectively. For a location Y_P in H_1, its heatmap value represents the confidence that J_1 is at Y_P. We propose to enhance it by the confidence of the linked joint J_2 at K possible *corresponding* locations $Y_{Q_k}, k = 1, \cdots, K$ which are consistent with Y_P according to limb orientation o.

The main challenge is to determine the locations of Y_{Q_k}. From Figure 3 (a), it is clear that the corresponding 3D point P of Y_P has to lie on the line defined by the camera center C_1 and Y_P. Since the exact depth of P is unknown, we log-uniformly sample K locations $P_k, k = 1, \cdots, K$ on the line as its candidates [1]. In addition, we assume the limb length l between J_1 and J_2 is provided as a prior which is the average limb length computed on the training dataset. Together with the 3D orientation o between the two joints, we can compute the 3D locations of J_2 as follows:

$$Q_k = P_k + o * l \quad \forall k = 1, \cdots, K \tag{1}$$

Finally we project Q_k onto the image using the camera parameters and get the 2D locations as Y_{Q_k}. Intuitively, a high response at Y_{Q_k} in H_2 actually indicates J_1 has a high probability to be at Y_P. This observation is the core of our fusion approach. However, there is ambiguity because we do not know which of the K candidates Y_{Q_k} is the corresponding point due to the lack of depth.

Our solution is to find the maximum response among all locations $Y_{Q_k}, k = 1, \cdots, K$:

$$H_1(Y_P) \leftarrow \lambda H_1(Y_P) + (1 - \lambda) \max_{k=1\cdots K} H_2(Y_{Q_k}) \tag{2}$$

Since fusion happens in the heatmap layer, ideally, Y_{Q_k} should have the largest response at the correct J_2 location and zeros at other locations. It means the non-corresponding locations will contribute no or little to the fusion. We set the balancing parameter λ to be 0.5 in our experiments. We sample 200 points whose depths range from zero to the maximum depth value, which is determined by the size of the room.

Cross-View Fusion One limitation of the *Same-View Fusion* is that the correct location $Y_{Q_{k*}}$ which has the maximum response among the K candidates in H_2, will contribute to multiple candidates like Y_P in H_1. These candidates also lie on a line. But most of such locations do not correspond to the joint type J_1. In other words, some non-corresponding locations are mistakenly enhanced. For example, there are blurred lines in the "Enhanced Heatmap" in Figure 4 with each from a different camera view.

---

[1]We use log-uniform instead of uniform sampling to prevent from generating redundant collapsed 2D projections
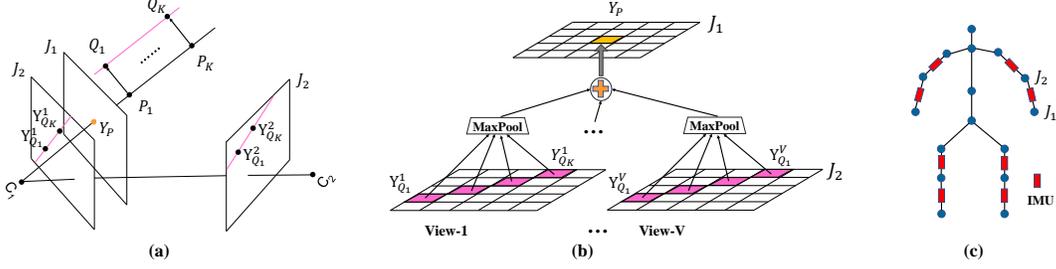
3

Figure 3. Illustration of the cross-joint-fusion idea in ORN. (a) For a location Y_P in H_1, we estimate its 3D points P_k lying on the line defined by the camera center C_1 and Y_P. Then based on the 3D limb orientation provided by IMU and the limb length, we get candidate 3D locations of J_2 which are denoted as Q_k. We project Q_k to the image as Y_{Q_k} and get the corresponding heatmap confidence. If the confidence is high, J_1 has high confidence being located at Y_P. (b) We enhance the initial confidence of J_1 at Y_P with the confidence of J_2 at Y_{Q_k} in all views. Similarly, we can fuse the heatmap of J_2 using that of J_1. (c) We show the skeleton model used in this work.

To resolve this problem, we propose to perform fusion across multiple views simultaneously:

$$H_1(Y_P) \leftarrow \lambda H_1(Y_P) + \frac{(1-\lambda)}{V} \sum_{v=1}^{V} \max_{k=1\cdots K} H_2^v(Y_{Q_k}^v),$$
(3)

where $Y_{Q_k}^v$ is the projection of Q_k in the camera view v and H_2^v is the heatmap of J_2 in view v. The result is that the lines from multiple views will intersect at the correct location. Consequently, the correct location will be enhanced most which resolves the ambiguity. See the fused heatmap in Figure 4 for illustration. Another desirable effect of cross-view fusion is that it helps solve the occlusion problem by fusing the features from multiple views because a joint occluded in one view may be visible in other views. This notably increases the joint detection rates.

3.2. Implementation

We use the network proposed in [31], referred to as SimpleNet (SN) to estimate initial pose heatmaps. It uses ResNet50 [8] as its backbone which was pre-trained on the ImageNet classification dataset. The image size is 256×256 and the heatmap size is 64×64. The orientation regularization module can either be trained end-to-end with SN, or added to a already trained SN as a plug-in since it has no learnable parameters. In this work, we train the whole ORN end-to-end. We generate ground-truth pose heatmaps as the regression targets and enforce l_2 loss on all views before and after feature fusion. In particular, we do not compute losses for background pixels of the fused heatmap since the background pixels may have been enhanced. The network is trained for 15 epochs. The parameter λ is 0.5 in all experiments. Other hyper-parameters such as learning rate and decay strategy are the same as in [31].

4. ORPSM for 3D Pose Estimation

A human is represented by a number of joints $\mathcal{J} = \{J_1, J_2, \cdots, J_M\}$. Each J represents its 3D position in a world coordinate system. Following the previous works [12, 17, 2, 19], we use the pictorial model to estimate 3D pose as it is more robust to inaccurate 2D poses. But different from the previous works, we also introduce and evaluate a novel limb orientation prior based on IMUs as will be described in detail later. Each J takes values from a discrete state space. An edge between two joints denotes their conditional dependence such as limb length. Given a 3D pose $\mathcal{J}$ and multi-view 2D pose heatmaps $\mathcal{F}$, we compute the posterior as follows

$$p(\mathcal{J}|\mathcal{F}) = \frac{1}{Z(\mathcal{F})} \prod_{i=1}^{M} \phi_i^{\text{conf}}(J_i, \mathcal{F}) \prod_{(m,n)\in\mathcal{E}_{limb}} \psi^{\text{limb}}(J_m, J_n)$$
$$\prod_{(m,n)\in\mathcal{E}_{IMU}} \psi^{\text{IMU}}(J_m, J_n),$$
(4)

where $Z(\mathcal{F})$ is the partition function, $\mathcal{E}_{limb}$ and $\mathcal{E}_{IMU}$ are sets of edges on which we enforce limb length and orientation constraints, respectively. The unary potential $\phi_i^{\text{conf}}(J_i, \mathcal{F})$ is computed based on 2D pose heatmaps $\mathcal{F}$. The pairwise potential $\psi^{\text{limb}}(J_m, J_n)$ and $\psi^{\text{IMU}}(J_m, J_n)$ encode the limb length and orientation constraints. We describe each term in detail as follows.

Discrete State Space We first estimate the 3D location of the root joint by triangulation based on its 2D locations detected in all views. Note that this step is usually very accurate because the root joint can be detected in most times. Then the state space of the 3D pose is within a 3D bounding volume centered at the root joint. The edge length of the volume is set to be 2000mm which is large enough to cover every body joint. The volume is discretized by an $N \times N \times N$ regular grid $\mathcal{G}$. Each joint can take one of the bins of the grid as its 3D location. Note that all body joints share the same state space $\mathcal{G}$ which consists of N^3 discrete locations (bins).

4

Table 1. The 2D pose estimation accuracy (PCKh@t) on the Total Capture Dataset. "SN" means SimpleNet which is the baseline. ORN^{same} and ORN, respectively, represent that the same-view and cross-view fusion are used. "Mean (six)" is the average result over the six joint types. "Others" is the average result over the rest of the joints. "Mean (All)" is the result over all joints.

Methods	PCKh@	Hip	Knee	Ankle	Shoulder	Elbow	Wrist	Mean (Six)	Others	Mean (All)
SN	1/2	99.3	98.3	98.5	98.4	96.2	95.3	97.7	99.5	98.1
ORN^{same}	1/2	99.4	99.0	98.8	98.5	97.7	96.7	98.3	99.5	98.6
ORN	1/2	**99.6**	**99.2**	**99.0**	**98.9**	**98.0**	**97.4**	**98.7**	99.5	98.9
SN	1/6	97.5	92.3	92.5	78.3	80.8	80.0	86.9	95.4	89.1
ORN^{same}	1/6	97.2	94.0	93.3	78.1	83.5	82.0	88.0	95.4	89.9
ORN	1/6	**97.7**	**94.8**	**94.2**	**81.1**	**84.7**	**83.6**	**89.3**	95.4	90.9
SN	1/12	**87.6**	67.0	68.6	47.4	50.0	49.3	61.7	78.1	65.8
ORN^{same}	1/12	81.2	70.1	68.0	43.9	51.6	50.1	60.8	78.1	65.2
ORN	1/12	85.3	**71.6**	**70.6**	**47.7**	**53.2**	**51.9**	**63.4**	78.1	67.1

Unary Potential Every body joint hypothesis, *i.e.*, a bin in the grid $\mathcal{G}$, is defined by its 3D position. We project it to the pixel coordinate system of all camera views using the camera parameters, and get the corresponding joint confidence/response from $\mathcal{F}$. We compute the average confidence/response over all camera views as the unary potential for the hypothesis.

Limb Length Potential For each pair of joints (J_m, J_n) in the edge set $\mathcal{E}_{limb}$, we compute the average distance $\tilde{l}_{m,n}$ on the training set as limb length prior. During inference, the limb length pairwise potential is defined as:

$$\psi^{\text{limb}}(J_m, J_n) = \begin{cases} 1, & \text{if} \quad |l_{m,n} - \tilde{l}_{m,n}| \leq \epsilon, \\ 0, & \text{otherwise} \end{cases}, \quad (5)$$

where $l_{m,n}$ is the distance between J_m and J_n. The pairwise term favors 3D poses having reasonable limb lengths. In our experiments, ϵ is set to be 150mm.

Limb Orientation Potential We compute the dot product between the limb orientations of the estimated pose and the IMU orientations as the limb orientation potential

$$\psi^{\text{IMU}}(J_m, J_n) = \frac{J_m - J_n}{\|J_m - J_n\|_2} \cdot o_{m,n}, \quad (6)$$

where $o_{m,n}$ is the orientation (represented as a directional vector) of the limb measured by the IMU. This term favors poses whose limb orientations are consistent with the IMUs. We also experimented with the hard orientation constraint similar to what we did for limb length, but this soft limb orientation constraint gets better performance. A 3D pose estimator without/with *orientation potential* will be termed as **PSM** and **ORPSM**, respectively.

Inference We maximize the posterior probability, *i.e.* Eq. (4), over the discrete state space by the dynamic programming algorithm. In general, the complexity grows quadratically. In order to improve the speed, we adopt a recursive
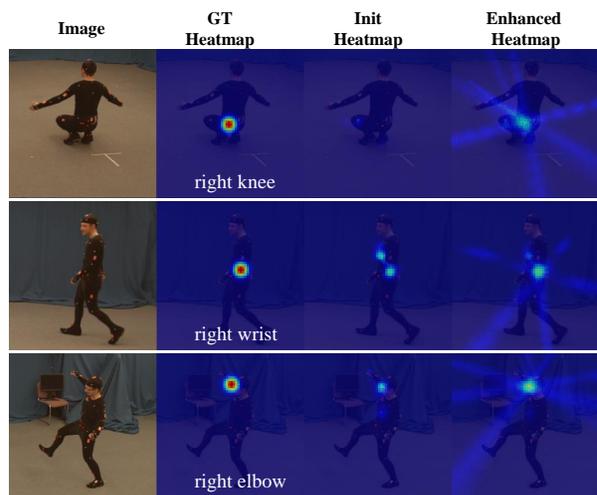

Figure 4. Three sample heatmaps estimated by *ORN*. The initially estimated heatmaps without fusion are inaccurate. After fusing multi-view features, the "Enhanced Heatmap" localizes the correct joints. Note there are blurred lines in the "Enhanced Heatmap" with each corresponding to the confidence contributed from one camera view. The lines intersect at the correct location.

variant of PSM [19] which iteratively refines the 3D poses. In practice, it takes about 0.15 seconds to estimate one 3D pose on a single Titan Xp GPU.

5. Datasets and Metrics

Total Capture [27] To the best of our knowledge, this is the only benchmark providing images, IMUs and ground truth 3D poses. It places 8 cameras in the capture room to record the human motion. We use four of them (1, 3, 5 and 7) in our experiments for efficiency reasons. The performers wear 13 IMUs. We use eight of them as shown in Figure 3 (c). There are five subjects performing four actions including Roaming(**R**), Walking(**W**), Acting(**A**) and Freestyle(**FS**) with each repeating 3 times. Following the previous work [27], we use Roaming 1,2,3, Walking 1,3,
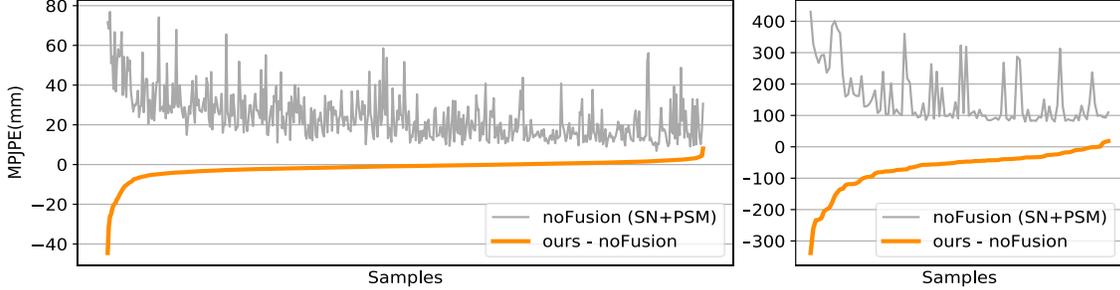
Figure 5. The grey line shows the 3D MPJPE error of the *noFusion* approach. The orange line shows the error difference between *our method (ORN+ORPSM)* and *noFusion*. If the orange line is below zero, it means *our method* has smaller errors. We split the testing samples into two groups according to the error scale of *noFusion*. The first group includes the samples whose errors are smaller than 80mm (shown in the left figure). The second group includes the rest of the samples (shown in the right figure). The samples are sorted by the orange line for the sake of readability.

Freestyle 1,2 and Acting 1,2 of *Subjects 1,2,3* for training our 2D pose estimator. We test on Walking 2, Freestyle 3 and Acting 3 of all subjects.

H36M [9]   To validate the general applicability of our approach, we also conduct experiments on the H36M dataset. Since this dataset does not provide IMUs, we create virtual IMUs (limb orientations) using the ground truth 3D poses for both training and testing, and only show proof-of-concept results. Following the dataset conventions, we use subjects $1, 5, 6, 7, 8$ for training and subjects $9, 11$ for testing. We train a single model for all actions.

Metrics   The Percentage of Correct Keypoints (PCK) metric is used for 2D pose evaluation. Specifically, PCKh@t measures the percentage of the estimated joints whose distance from the ground-truth joints is smaller than t times of the head length. Previous works report results when t is $\frac{1}{2}$. In our experiments, we provide results when t is set to be $\frac{1}{2}$, $\frac{1}{6}$ and $\frac{1}{12}$, respectively, in order to understand our approach more comprehensively. We use the Mean Per Joint Position Error (MPJPE) for 3D pose evaluation [9]. It computes the distance between estimated poses and the ground truth poses. We report the average error over all joints and all instances.

6. Experimental Results

6.1. 2D Pose Estimation

Table 1 shows the 2D pose estimation results of *ORN* and the baseline *SimpleNet (SN)*. We keep *SN* the same as *ORN* except it does not perform fusion. We can see from the table that when the threshold t is set to be $\frac{1}{2}$ as in most previous works, *ORN* outperforms *SN* by a large margin. The improvement for wrist, elbow and knee joints is most significant because they are frequently occluded by human body in the dataset. Figure 4 shows some examples explaining

how our approach improves localization over the baseline. For example, in the first example, initially, the right knee joint is not correctly detected because it is occluded by human body. Fusing the features from hip in multiple camera views helps localize it correctly.

We notice that the improvement (brought by visual-inertial fusion) on the hip joints is small. There are two possible reasons for the phenomenon. First, the hip joints are visible for most images in the dataset. So IMUs provide barely no additional information. Second, the hip joint detection rate for the baseline method is already very high so it leaves little room for improvement. The estimation results for the joints without IMUs, which are represented as *"Others"* in the table, are similar for the baseline and our approach, which is expected.

When we use a more rigorous threshold, for example when $t = \frac{1}{12}$, the detection rate for hip drops from 87.6% to 85.3% (SN vs. ORN). There are two reasons for this phenomenon: (1) the detection rate for hip is already very high for *SN*, leaving little space for improvement; (2) IMUs often have small noises which may affect fusion precision. This conclusion is supported by the subsequent experimental results on the H36M dataset: when we use GT IMUs, the detection rate also improves for hip. Actually, even on the Total Capture, the impact also becomes small when we use a larger threshold. For example, when the threshold is set to be $\frac{1}{6}$, the accuracy of *ORN* is slightly better than *SN* (97.7% vs. 97.5%).

We also evaluate the impact of cross-view fusion in *ORN*. As can be seen in Table 1, the multi-view fusion outperforms the same-view fusion consistently which validates its effectiveness. In addition, we find that the improvement is larger when we use a more rigorous threshold t. The results suggest that multi-view feature fusion helps localize the joints more precisely.
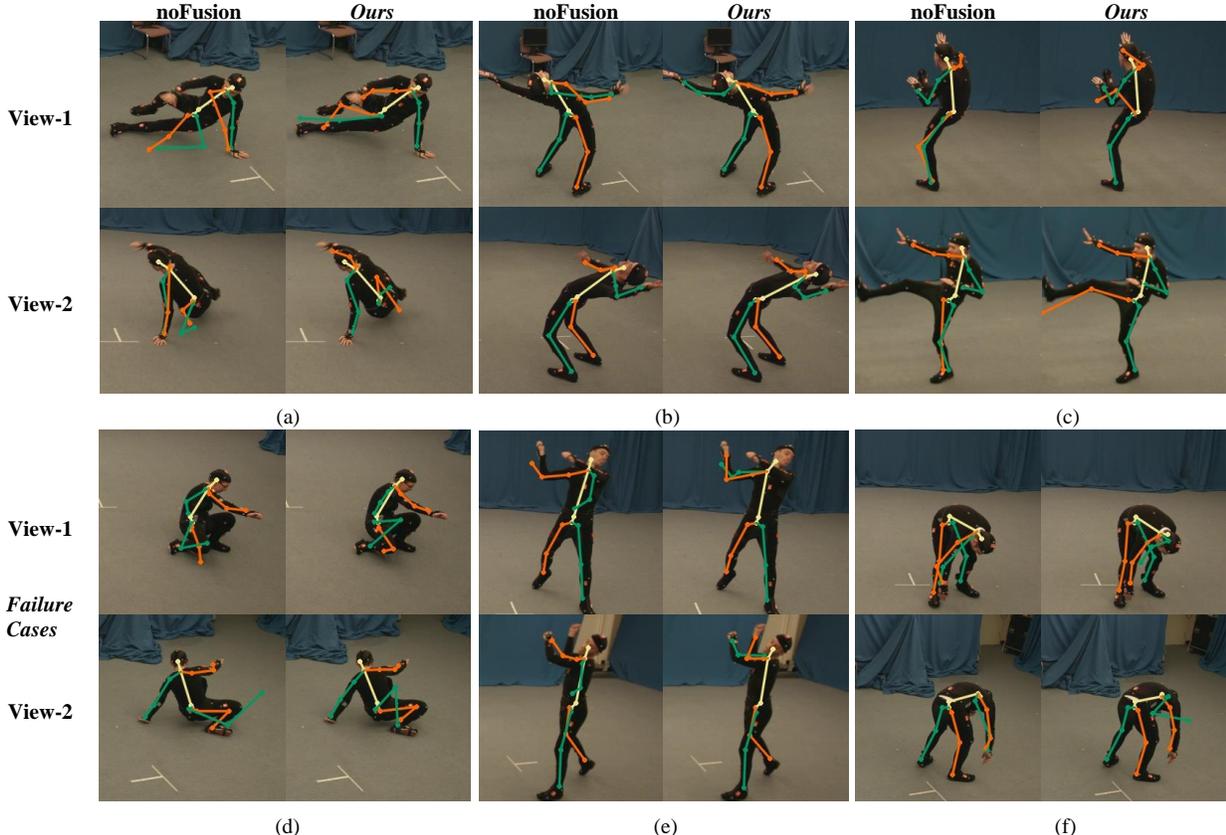
6

Figure 6. Sample 3D poses estimated by our approach and *noFusion*. We project the estimated 3D poses to the images and draw the skeletons. Left and right limbs are drawn in green and orange colors, respectively. (a-c) show examples when our method improves over *noFusion*. (d-f) show three failure cases. These rare cases mainly happen when both joints of a limb have large errors.

Table 2. 3D pose estimation errors (mm) of different variants of our approach on the Total Capture dataset. "Mean (six)" is the average error over the six joint types. "Others" is the average error over the rest of the joints. "Mean (All)" is the average error over all joints.

2D	3D	Hip	Knee	Ankle	Shoulder	Elbow	Wrist	*Mean (Six)*	Others	Mean (All)
SN	PSM	**17.2**	35.7	41.2	50.5	54.8	56.8	37.1	20.3	28.3
ORN	PSM	17.4	29.9	35.2	49.6	44.2	45.1	32.8	20.4	25.4
SN	*ORPSM*	18.3	25.8	34.0	44.8	44.2	49.8	32.1	19.9	25.5
ORN	*ORPSM*	18.5	**24.2**	**30.1**	**44.8**	**40.7**	**43.4**	**30.2**	19.8	24.6

6.2. 3D Pose Estimation

We first evaluate our 3D pose estimator through a number of ablation studies. Then we compare our approach to the state-of-the-arts. Finally, we present results on the H36M dataset validating the generalization capability of the proposed approach.

Ablation Study   We denote the baseline which uses *SN* and *PSM* to estimate 2D and 3D pose as ***noFusion*** baseline. The main results are shown in Table 2. First, using *ORN* consistently decreases the 3D error no matter what 3D pose estimators we use. In particular, the improvement on the elbow and wrist joint is as large as 10mm when we use

PSM as the 3D estimator. This significant error reduction is attributed to the improved 2D poses. Figure 6 (a-c) visualize three typical examples where *ORN* gets better results: we project the estimated 3D poses to the images and draw the skeletons. It is guaranteed that if the 2D locations are correct for more than one view, then the 3D joint location is at the correct position. We also plot the 3D error of every testing sample in Figure 5. Our approach improves the accuracy for most cases because the orange line is mostly below zero. See the caption of the figure for the meanings of the lines. In addition, we can see that the improvement is larger when the *noFusion* baseline has large errors. There are a small number of cases where fusion does not improve joint detection results as shown in Figure 6 (d-f).

Table 3. 3D pose estimation errors MPJPE (mm) of different methods on the Total Capture dataset. "Aligned" means whether we align the estimated 3D poses to the ground truth poses by Procrustes.

Approach	IMUs	Temporal	Aligned	Subjects(S1,2,3)			Subjects(S4,5)			Mean
				W2	A3	FS3	W2	A3	FS3	
PVH [27]				48.3	94.3	122.3	84.3	154.5	168.5	107.3
Malleson *et al.* [15]	✓	✓		-	-	65.3	-	64.0	67.0	-
VIP [28]	✓	✓	✓	-	-	-	-	-	-	26.0
LSTM-AE [26]		✓		**13.0**	23.0	47.0	**21.8**	40.9	68.5	34.1
IMUPVH [6]	✓	✓		19.2	42.3	48.8	24.7	58.8	61.8	42.6
Qiu *et al.* [19]				19.0	21.0	28.0	32.0	33.0	54.0	29.0
SN + PSM				14.3	18.7	31.5	25.5	30.5	64.5	28.3
SN + PSM		✓		12.7	16.5	28.9	21.7	26.0	59.5	25.3
ORN + ORPSM	✓			14.3	**17.5**	**25.9**	23.9	**27.8**	**49.3**	**24.6**
ORN + ORPSM	✓		✓	12.4	14.6	22.0	19.6	22.4	41.6	20.6

Table 4. 3D pose estimation error (mm) on the H36M dataset. We use virtual IMUs in this experiment. We show results for the six joints which are affected by IMUs. "Mean (six)" is the average error over the six joint types. "Others" is the average error over the rest of the joints. "Mean (All)" is the average error over all joints.

Methods	Hip	Knee	Ankle	Shoulder	Elbow	Wrist	*Mean (Six)*	Others	Mean (All)
noFusion (SN + PSM)	23.2	28.7	49.4	29.1	28.4	32.3	31.9	18.3	27.9
ours (ORN + ORPSM)	**20.6**	**18.6**	**28.2**	**25.1**	**21.8**	**24.2**	**23.1**	18.3	21.7

Second, from the second and third rows of Table 2, we can see that using ORPSM alone achieves a similar 3D error as ORN alone. This means 3D fusion is related to 2D fusion in some way— although 3D fusion does not directly improve the 2D heatmap quality, it uses 3D priors to select better joint locations having both large responses as well as small discrepancy with respect to the prior structures. But in some cases, for example, when the responses at the correct locations are too small, using the 3D prior is not sufficient. This is verified by the experimental results in the fourth row — if we enforce 2D and 3D fusion simultaneously, the error further decreases to 30.2mm. It suggests the two components are actually complementary.

State-of-the-arts   Finally we compare our approach to the state-of-the-arts on the Total Capture dataset. The results are shown in Table 3. First, we can see that *IMUPVH* [6] which uses IMUs even gets worse results than *LSTM-AE* [26] which does not use IMUs. The results suggest that getting better visual features is actually more effective than performing late fusion of the (possibly inaccurate) 3D poses obtained from images and IMUs, respectively. Our approach, which uses IMUs to improve the visual features, also outperforms [6] by a large margin.

The error of the state-of-the-art is about 29mm [19] which is larger than 24.6mm of ours. This validates the effectiveness of our IMU-assisted early visual feature fusion. Note that the error of VIP [28] is obtained when the 3D pose estimations are aligned to ground truth which should be compared to 20.6mm of our approach.

We notice that the error of our approach is slightly larger than [26] for the "W2 (walking)" action. We tend to think it is because LSTM can get significant benefits when it is applied to *periodic* actions such as "walking". This is also observed independently in another work [6]. Besides, the error 14.3mm for "W2" of Subject 1,2,3 is not further reduced after fusion since *noFusion* method has already achieved extraordinarily high accuracy.

Generalization   To validate the wide applicability of our approach, we conduct experiments on the H36M dataset [9]. The results of different methods are shown in Table 4. We can see that our approach (ORN+ORPSM) consistently outperforms the baseline *noFusion* which validates its general applicability. In particular, the improvement is significant for the Ankle joint which is often occluded. Since we use the ground truth IMU orientations in this experiment, the results are not directly comparable to other works.

7. Summary and Future Work

We present an approach for fusing visual features through IMUs for 3D pose estimation. The main difference from the previous efforts is that we use IMUs in a very early stage. We evaluate the approach through a number of ablation studies, and observe consistent improvement resulted from the fusion. As the readings from the IMUs usually have noises, our future work will focus on learning a reliability indicator, for example based on temporal filtering, for each sensor to guide the fusion process.

References

[1] Sikandar Amin, Mykhaylo Andriluka, Marcus Rohrbach, and Bernt Schiele. Multi-view pictorial structures for 3D human pose estimation. In *BMVC*, 2013.

[2] Vasileios Belagiannis, Sikandar Amin, Mykhaylo Andriluka, Bernt Schiele, Nassir Navab, and Slobodan Ilic. 3d pictorial structures for multiple human pose estimation. In *CVPR*, pages 1669–1676, 2014.

[3] Magnus Burenius, Josephine Sullivan, and Stefan Carlsson. 3D pictorial structures for multiple view articulated pose estimation. In *CVPR*, pages 3618–3625, 2013.

[4] Junting Dong, Wen Jiang, Qixing Huang, Hujun Bao, and Xiaowei Zhou. Fast and robust multi-person 3d pose estimation from multiple views. *arXiv preprint arXiv:1901.04111*, 2019.

[5] Juergen Gall, Bodo Rosenhahn, Thomas Brox, and Hans-Peter Seidel. Optimization and filtering for human motion capture. *IJCV*, 87(1-2):75, 2010.

[6] Andrew Gilbert, Matthew Trumble, Charles Malleson, Adrian Hilton, and John Collomosse. Fusing visual and inertial sensors with semantics for 3d human pose estimation. *IJCV*, 127(4):381–397, 2019.

[7] Andrew Gilbert, Marco Volino, John Collomosse, and Adrian Hilton. Volumetric performance capture from minimal camera viewpoints. In *ECCV*, 2018.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[9] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3D human sensing in natural environments. *T-PAMI*, pages 1325–1339, 2014.

[10] Yasamin Jafarian, Yuan Yao, and Hyun Soo Park. Monet: Multiview semi-supervised keypoint via epipolar divergence. *arXiv preprint arXiv:1806.00104*, 2018.

[11] Hanbyul Joo, Tomas Simon, Xulong Li, Hao Liu, Lei Tan, Lin Gui, Sean Banerjee, Timothy Godisart, Bart Nabbe, Iain Matthews, et al. Panoptic studio: A massively multiview system for social interaction capture. *T-PAMI*, 41(1):190–204, 2019.

[12] Ilya Kostrikov and Juergen Gall. Depth sweep regression forests for estimating 3D human pose from images. In *BMVC*, page 5, 2014.

[13] Yebin Liu, Carsten Stoll, Juergen Gall, Hans-Peter Seidel, and Christian Theobalt. Markerless motion capture of interacting characters using multi-view image segmentation. In *CVPR*, pages 1249–1256. IEEE, 2011.

[14] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *TOG*, 34(6):248, 2015.

[15] Charles Malleson, Andrew Gilbert, Matthew Trumble, John Collomosse, Adrian Hilton, and Marco Volino. Real-time full-body motion capture from video and imus. In *3DV*, pages 449–457. IEEE, 2017.

[16] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3D human pose estimation. In *ICCV*, page 5, 2017.

[17] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G. Derpanis, and Kostas Daniilidis. Harvesting multiple views for marker-less 3D human pose annotations. In *CVPR*, pages 1253–1262, 2017.

[18] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *CVPR*, pages 7753–7762, 2019.

[19] Haibo Qiu, Chunyu Wang, Jingdong Wang, Naiyan Wang, and Wenjun Zeng. Cross view fusion for 3d human pose estimation. In *ICCV*, pages 4342–4351, 2019.

[20] Helge Rhodin, Mathieu Salzmann, and Pascal Fua. Unsupervised geometry-aware representation for 3d human pose estimation. In *ECCV*, September 2018.

[21] Helge Rhodin, Jörg Spörri, Isinsu Katircioglu, Victor Constantin, Frédéric Meyer, Erich Müller, Mathieu Salzmann, and Pascal Fua. Learning monocular 3d human pose estimation from multi-view images. In *CVPR*, pages 8437–8446, 2018.

[22] Daniel Roetenberg, Henk Luinge, and Per Slycke. Xsens mvn: full 6dof human motion tracking using miniature inertial sensors. *Xsens Motion Technologies BV, Tech. Rep*, 1, 2009.

[23] Ronit Slyper and Jessica K Hodgins. Action capture with accelerometers. In *ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, pages 193–199. Eurographics Association, 2008.

[24] Jochen Tautges, Arno Zinke, Björn Krüger, Jan Baumann, Andreas Weber, Thomas Helten, Meinard Müller, Hans-Peter Seidel, and Bernd Eberhardt. Motion reconstruction using sparse accelerometer data. *TOG*, 30(3):18, 2011.

[25] Denis Tome, Matteo Toso, Lourdes Agapito, and Chris Russell. Rethinking pose in 3D: Multi-stage refinement and recovery for markerless motion capture. In *3DV*, pages 474–483, 2018.

[26] Matthew Trumble, Andrew Gilbert, Adrian Hilton, and John Collomosse. Deep autoencoder for combined human pose estimation and body model upscaling. In *ECCV*, pages 784–800, 2018.

[27] Matthew Trumble, Andrew Gilbert, Charles Malleson, Adrian Hilton, and John Collomosse. Total capture: 3D human pose estimation fusing video and inertial sensors. In *BMVC*, pages 1–13, 2017.

[28] Timo von Marcard, Roberto Henschel, Michael J Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *ECCV*, pages 601–617, 2018.

[29] Timo Von Marcard, Gerard Pons-Moll, and Bodo Rosenhahn. Human pose estimation from video and imus. *T-PAMI*, 38(8):1533–1547, 2016.

[30] Timo von Marcard, Bodo Rosenhahn, Michael J Black, and Gerard Pons-Moll. Sparse inertial poser: Automatic 3d human pose estimation from sparse imus. In *Computer Graphics Forum*. Wiley Online Library, 2017.

[31] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *ECCV*, pages 466–481, 2018.

[32] Wei Yang, Wanli Ouyang, Xiaolong Wang, Jimmy Ren, Hongsheng Li, and Xiaogang Wang. 3d human pose estimation in the wild by adversarial learning. In *CVPR*, pages 5255–5264, 2018.

[33] Xingyi Zhou, Qixing Huang, Xiao Sun, Xiangyang Xue, and Yichen Wei. Towards 3D human pose estimation in the wild: a weakly-supervised approach. In *ICCV*, pages 398–407, 2017.