Human Pose Estimation from Video and IMUs

2021.05.20.

고낙헌

저자 | Timo von Marcard, Gerard Pons-Moll, Bodo Rosenhahn

저널 | IEEE Transactions on Pattern Analysis and Machine Intelligence

목 차

1. 개요 및 요약
2. 관성 측정 장치(Inertial Measurement Unit)의 원리
3. Video-based Tracker
4. Hybrid Tracker with IMUs
5. Tracking Error Analysis
6. 결론
7. 향후 논문리뷰 및 연구 계획

1

개요 및 요약

연구요약

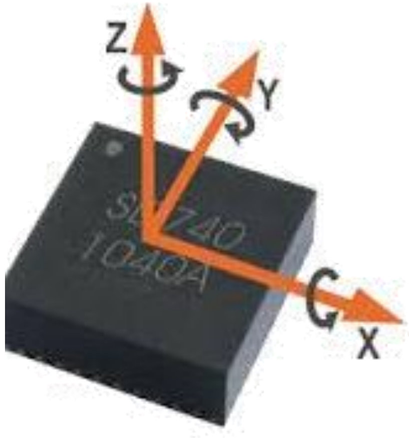
- 비디오 데이터와 방향 데이터를 융합하여 모션 캡쳐의 정확도를 향상시키는 연구

2

관성 측정 장치(IMU)의 원리

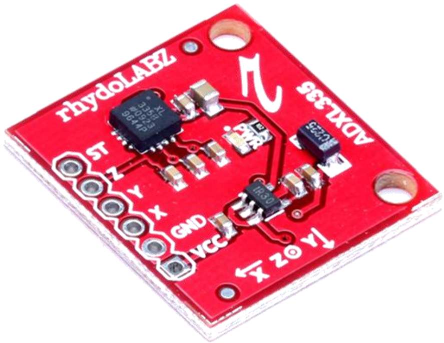
IMU란?

- IMU: 관성 측정 장치
- 자이로스코프, 가속도계, 지자기센서로 구성된 센서 집합체 → 센서의 방향벡터를 구함.



<3축 자이로스코프>

축 별로 시간당
몇 rad 회전했는지 측정



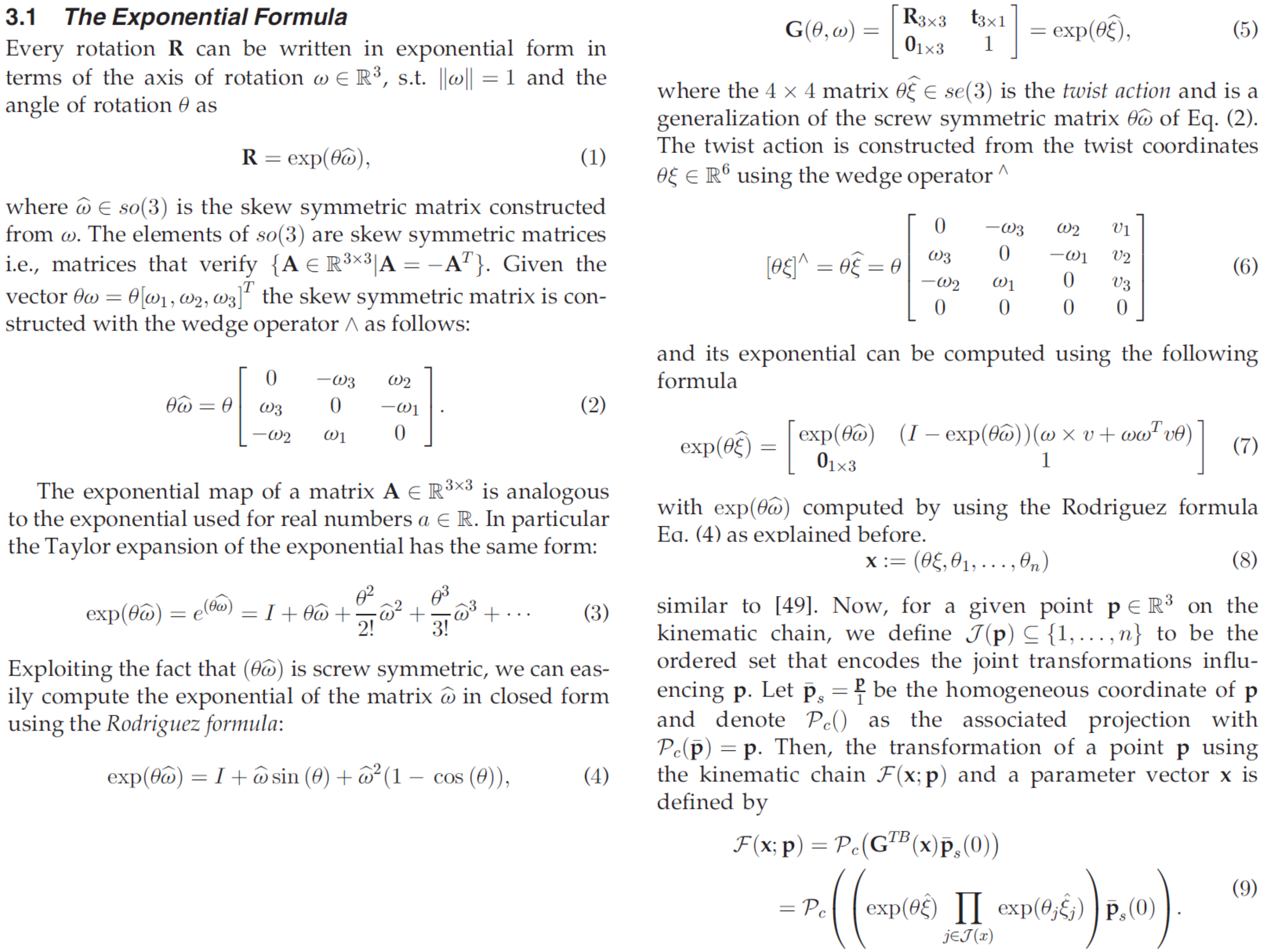
<가속도계>

중력 가속도를 분해하여
얼마나 기울어졌는지 측정

<지자기센서>

자북을 기준으로
자기선속의 세기를 측정하여
얼마나 틀어졌는지 측정

IMU의 원리



3.1 *The Exponential Formula*

Every rotation $\mathbf{R}$ can be written in exponential form in terms of the axis of rotation $\omega \in \mathbb{R}^3$, s.t. $\|\omega\| = 1$ and the angle of rotation θ as

$$\mathbf{R} = \exp(\theta\widehat{\omega}), \tag{1}$$

where $\widehat{\omega} \in so(3)$ is the skew symmetric matrix constructed from ω. The elements of $so(3)$ are skew symmetric matrices i.e., matrices that verify $\{\mathbf{A} \in \mathbb{R}^{3\times3} | \mathbf{A} = -\mathbf{A}^T\}$. Given the vector $\theta\omega = \theta[\omega_1, \omega_2, \omega_3]^T$ the skew symmetric matrix is constructed with the wedge operator $\wedge$ as follows:

$$\theta\widehat{\omega} = \theta \begin{bmatrix} 0 & -\omega_3 & \omega_2 \\ \omega_3 & 0 & -\omega_1 \\ -\omega_2 & \omega_1 & 0 \end{bmatrix}. \tag{2}$$

The exponential map of a matrix $\mathbf{A} \in \mathbb{R}^{3\times3}$ is analogous to the exponential used for real numbers $a \in \mathbb{R}$. In particular the Taylor expansion of the exponential has the same form:

$$\exp(\theta\widehat{\omega}) = e^{(\theta\widehat{\omega})} = I + \theta\widehat{\omega} + \frac{\theta^2}{2!}\widehat{\omega}^2 + \frac{\theta^3}{3!}\widehat{\omega}^3 + \cdots \tag{3}$$

Exploiting the fact that $(\theta\widehat{\omega})$ is screw symmetric, we can easily compute the exponential of the matrix $\widehat{\omega}$ in closed form using the *Rodriguez formula*:

$$\exp(\theta\widehat{\omega}) = I + \widehat{\omega}\sin(\theta) + \widehat{\omega}^2(1 - \cos(\theta)), \tag{4}$$

$$\mathbf{G}(\theta, \omega) = \begin{bmatrix} \mathbf{R}_{3\times3} & \mathbf{t}_{3\times1} \\ \mathbf{0}_{1\times3} & 1 \end{bmatrix} = \exp(\theta\widehat{\xi}), \tag{5}$$

where the 4×4 matrix $\theta\widehat{\xi} \in se(3)$ is the *twist action* and is a generalization of the screw symmetric matrix $\theta\widehat{\omega}$ of Eq. (2). The twist action is constructed from the twist coordinates $\theta\xi \in \mathbb{R}^6$ using the wedge operator $^\wedge$

$$[\theta\xi]^\wedge = \theta\widehat{\xi} = \theta \begin{bmatrix} 0 & -\omega_3 & \omega_2 & v_1 \\ \omega_3 & 0 & -\omega_1 & v_2 \\ -\omega_2 & \omega_1 & 0 & v_3 \\ 0 & 0 & 0 & 0 \end{bmatrix} \tag{6}$$

and its exponential can be computed using the following formula

$$\exp(\theta\widehat{\xi}) = \begin{bmatrix} \exp(\theta\widehat{\omega}) & (I - \exp(\theta\widehat{\omega}))(\omega \times v + \omega\omega^T v\theta) \\ \mathbf{0}_{1\times3} & 1 \end{bmatrix} \tag{7}$$

with $\exp(\theta\widehat{\omega})$ computed by using the Rodriguez formula Eq. (4) as explained before.

$$\mathbf{x} := (\theta\xi, \theta_1, \ldots, \theta_n) \tag{8}$$

similar to [49]. Now, for a given point $\mathbf{p} \in \mathbb{R}^3$ on the kinematic chain, we define $\mathcal{J}(\mathbf{p}) \subseteq \{1, \ldots, n\}$ to be the ordered set that encodes the joint transformations influencing $\mathbf{p}$. Let $\bar{\mathbf{p}}_s = \frac{\mathbf{p}}{1}$ be the homogeneous coordinate of $\mathbf{p}$ and denote $\mathcal{P}_c()$ as the associated projection with $\mathcal{P}_c(\bar{\mathbf{p}}) = \mathbf{p}$. Then, the transformation of a point $\mathbf{p}$ using the kinematic chain $\mathcal{F}(\mathbf{x}; \mathbf{p})$ and a parameter vector $\mathbf{x}$ is defined by

$$\begin{aligned} \mathcal{F}(\mathbf{x}; \mathbf{p}) &= \mathcal{P}_c(\mathbf{G}^{TB}(\mathbf{x})\bar{\mathbf{p}}_s(0)) \\ &= \mathcal{P}_c\left(\left(\exp(\theta\hat{\xi}) \prod_{j \in \mathcal{J}(x)} \exp(\theta_j\hat{\xi}_j)\right)\bar{\mathbf{p}}_s(0)\right). \end{aligned} \tag{9}$$

IMU의 한계

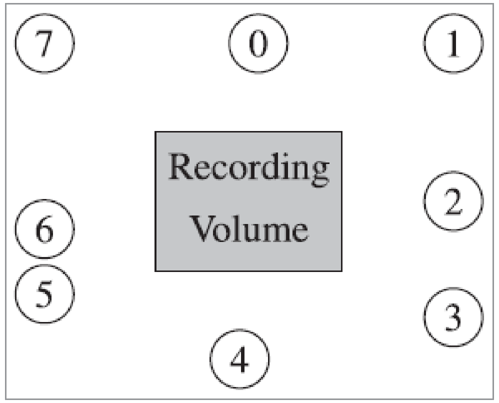
- 자기장 센서가 포함되어 있으므로, 여러 기기가 함께 사용되면 상호작용으로 인해 센서가 원래 위치에서 벗어난 것처럼 인식되는 ‘drift’ 현상 발생

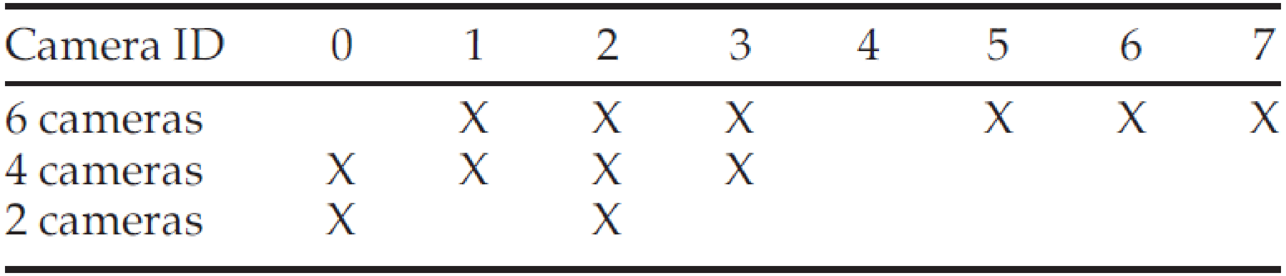
3

Video-based Tracker

Video만으로 모션 캡처를 구현했을 때

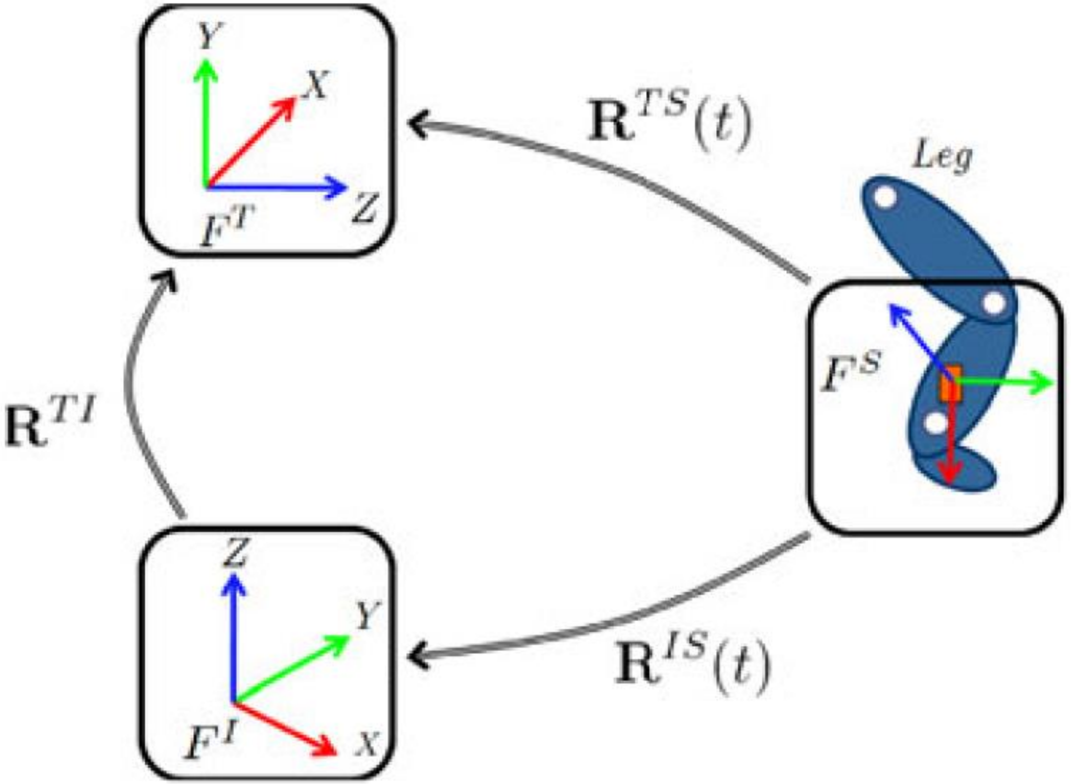
- 2개 이상의 multi-camera를 활용하여 2D 영상 복수 개를 3D 모션으로 복원함.
- 각 카메라가 바라보는 축의 방향을 변환하는 과정 필요



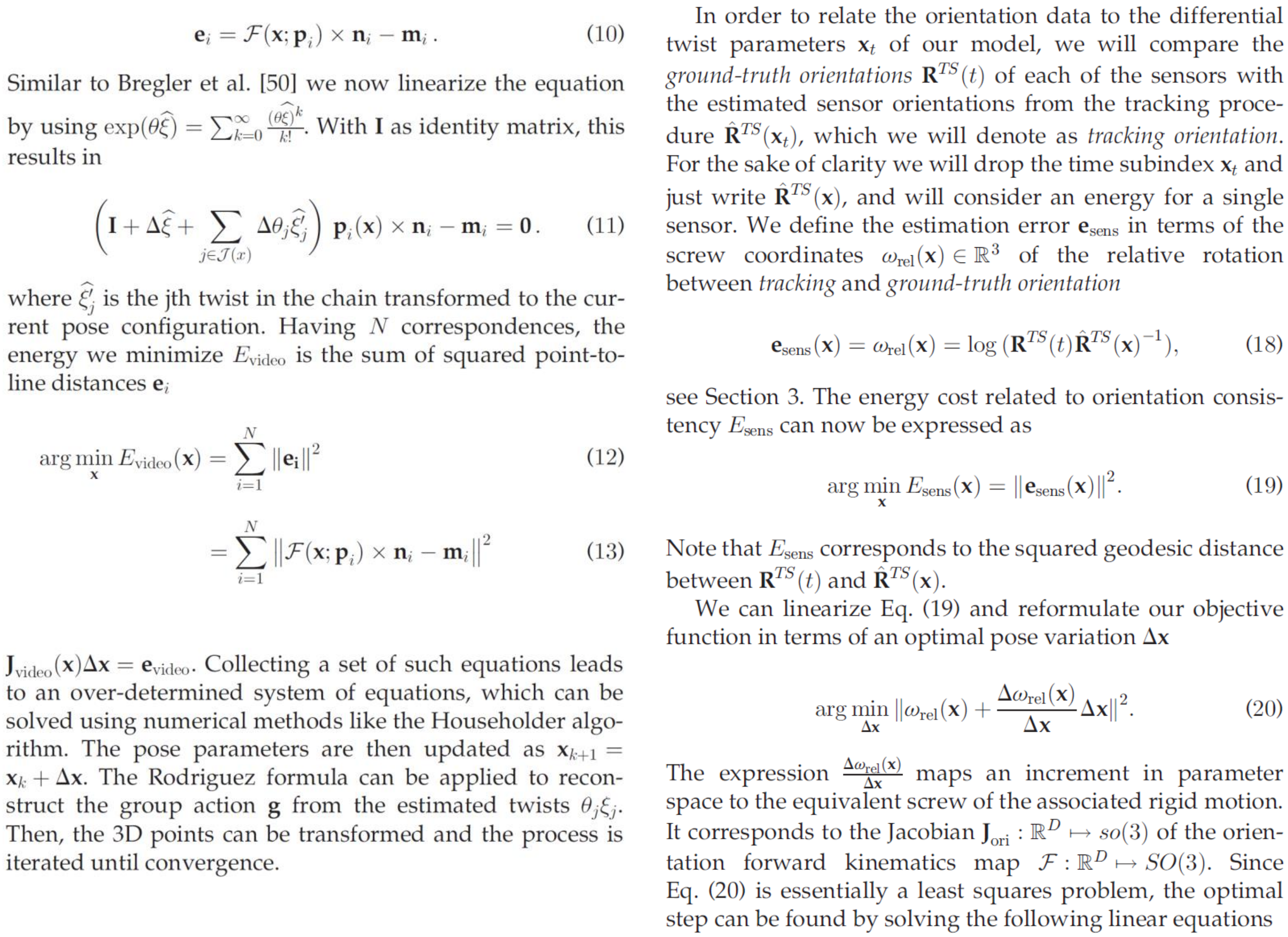


Camera ID	0	1	2	3	4	5	6	7
6 cameras		X	X	X		X	X	X
4 cameras	X	X	X	X				
2 cameras	X		X					

축 방향 변환 과정



Video-based Tracker의 원리



$$\mathbf{e}_i = \mathcal{F}(\mathbf{x};\mathbf{p}_i) \times \mathbf{n}_i - \mathbf{m}_i\,. \tag{10}$$

Similar to Bregler et al. [50] we now linearize the equation by using $\exp(\theta\widehat{\xi}) = \sum_{k=0}^{\infty} \frac{(\widehat{\theta\xi})^k}{k!}$. With **I** as identity matrix, this results in

$$\left(\mathbf{I} + \Delta\widehat{\xi} + \sum_{j \in \mathcal{J}(x)} \Delta\theta_j \widehat{\xi}'_j\right) \mathbf{p}_i(\mathbf{x}) \times \mathbf{n}_i - \mathbf{m}_i = \mathbf{0}\,. \tag{11}$$

where $\widehat{\xi}'_j$ is the jth twist in the chain transformed to the current pose configuration. Having N correspondences, the energy we minimize E_{video} is the sum of squared point-to-line distances $\mathbf{e}_i$

$$\arg\min_{\mathbf{x}} E_{\text{video}}(\mathbf{x}) = \sum_{i=1}^{N} \|\mathbf{e_i}\|^2 \tag{12}$$

$$= \sum_{i=1}^{N} \|\mathcal{F}(\mathbf{x};\mathbf{p}_i) \times \mathbf{n}_i - \mathbf{m}_i\|^2 \tag{13}$$

$\mathbf{J}_{\text{video}}(\mathbf{x})\Delta\mathbf{x} = \mathbf{e}_{\text{video}}$. Collecting a set of such equations leads to an over-determined system of equations, which can be solved using numerical methods like the Householder algorithm. The pose parameters are then updated as $\mathbf{x}_{k+1} = \mathbf{x}_k + \Delta\mathbf{x}$. The Rodriguez formula can be applied to reconstruct the group action **g** from the estimated twists $\theta_j\xi_j$. Then, the 3D points can be transformed and the process is iterated until convergence.

In order to relate the orientation data to the differential twist parameters $\mathbf{x}_t$ of our model, we will compare the *ground-truth orientations* $\mathbf{R}^{TS}(t)$ of each of the sensors with the estimated sensor orientations from the tracking procedure $\hat{\mathbf{R}}^{TS}(\mathbf{x}_t)$, which we will denote as *tracking orientation*. For the sake of clarity we will drop the time subindex $\mathbf{x}_t$ and just write $\hat{\mathbf{R}}^{TS}(\mathbf{x})$, and will consider an energy for a single sensor. We define the estimation error $\mathbf{e}_{\text{sens}}$ in terms of the screw coordinates $\omega_{\text{rel}}(\mathbf{x}) \in \mathbb{R}^3$ of the relative rotation between *tracking* and *ground-truth orientation*

$$\mathbf{e}_{\text{sens}}(\mathbf{x}) = \omega_{\text{rel}}(\mathbf{x}) = \log\left(\mathbf{R}^{TS}(t)\hat{\mathbf{R}}^{TS}(\mathbf{x})^{-1}\right), \tag{18}$$

see Section 3. The energy cost related to orientation consistency E_{sens} can now be expressed as

$$\arg\min_{\mathbf{x}} E_{\text{sens}}(\mathbf{x}) = \|\mathbf{e}_{\text{sens}}(\mathbf{x})\|^2. \tag{19}$$

Note that E_{sens} corresponds to the squared geodesic distance between $\mathbf{R}^{TS}(t)$ and $\hat{\mathbf{R}}^{TS}(\mathbf{x})$.

We can linearize Eq. (19) and reformulate our objective function in terms of an optimal pose variation $\Delta\mathbf{x}$

$$\arg\min_{\Delta\mathbf{x}} \|\omega_{\text{rel}}(\mathbf{x}) + \frac{\Delta\omega_{\text{rel}}(\mathbf{x})}{\Delta\mathbf{x}}\Delta\mathbf{x}\|^2. \tag{20}$$

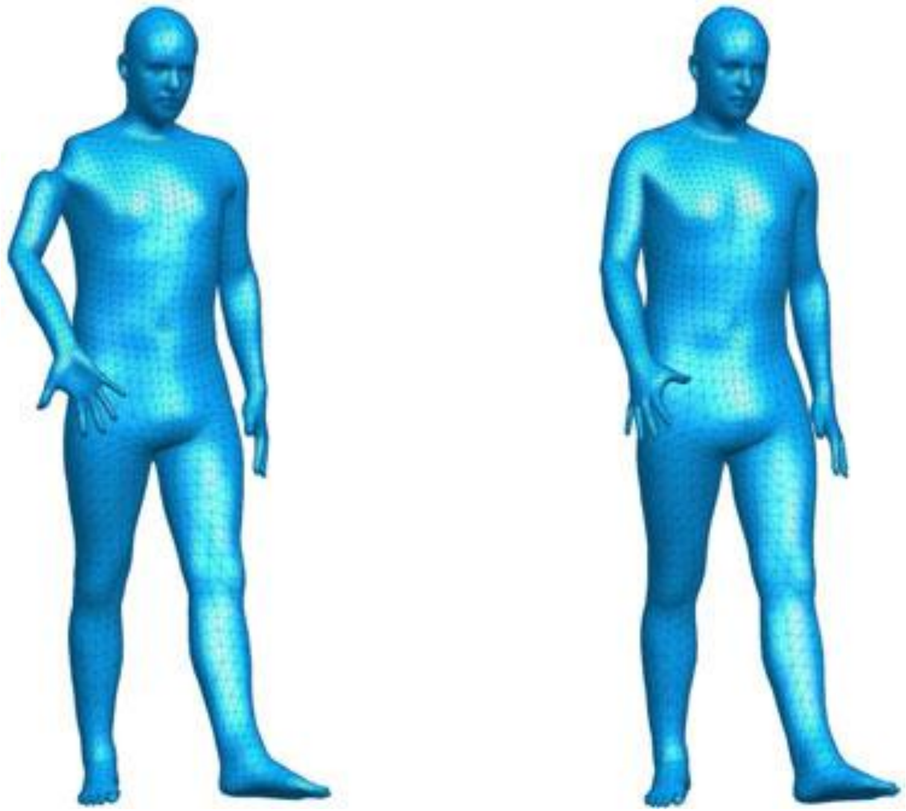
The expression $\frac{\Delta\omega_{\text{rel}}(\mathbf{x})}{\Delta\mathbf{x}}$ maps an increment in parameter space to the equivalent screw of the associated rigid motion. It corresponds to the Jacobian $\mathbf{J}_{\text{ori}} : \mathbb{R}^D \mapsto so(3)$ of the orientation forward kinematics map $\mathcal{F} : \mathbb{R}^D \mapsto SO(3)$. Since Eq. (20) is essentially a least squares problem, the optimal step can be found by solving the following linear equations

4

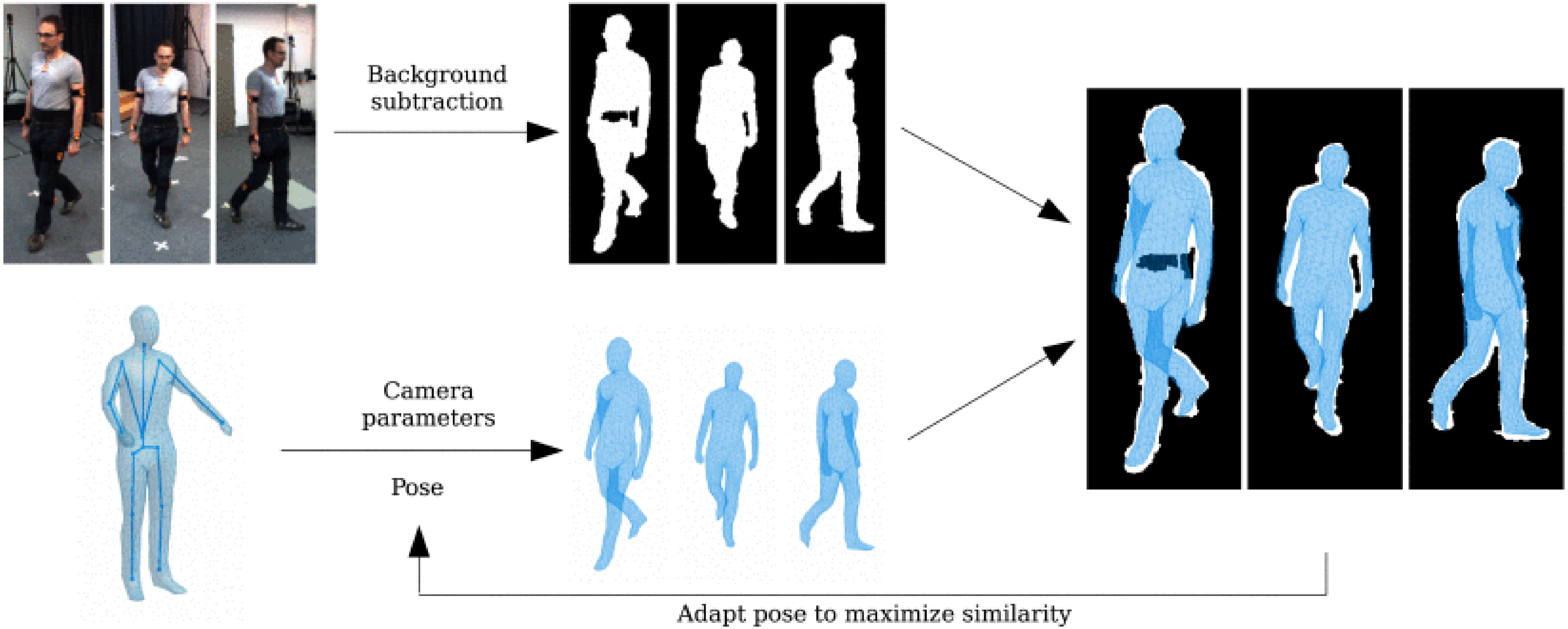
Hybrid Tracker with IMUs

Video 데이터와 IMU 방향데이터의 장점 결합

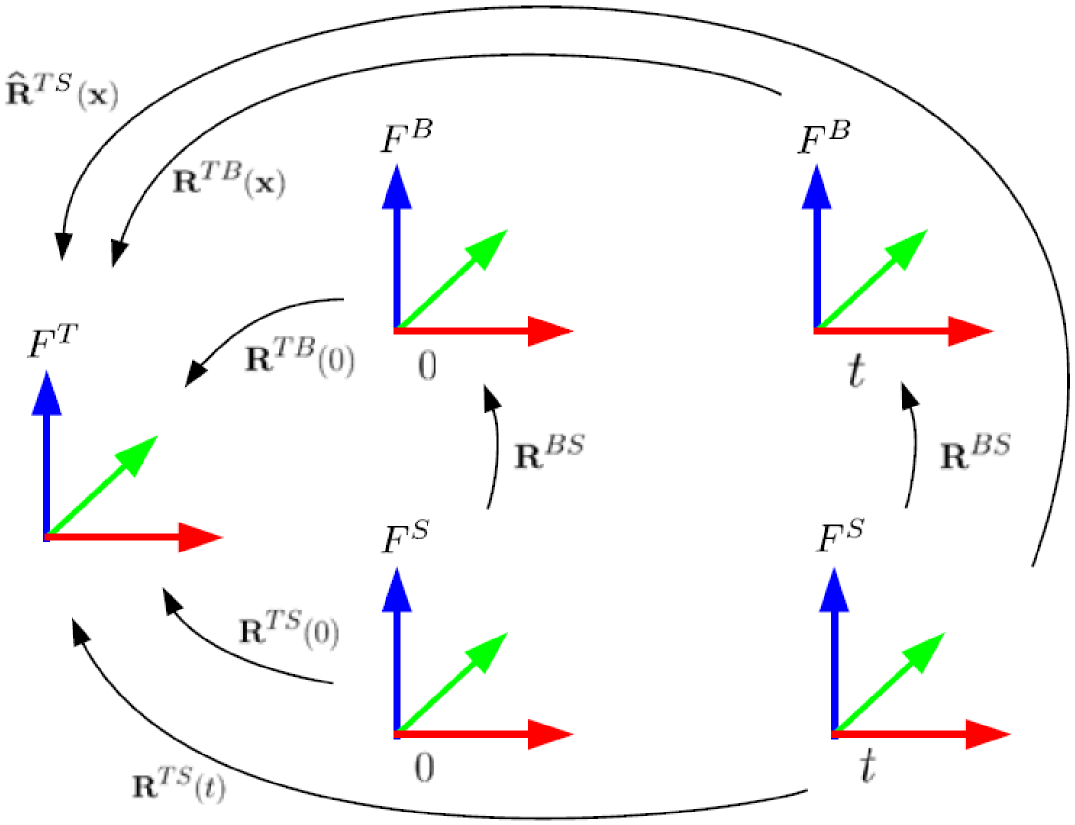
- 좌우 아래 팔, 좌우 정강이, 허리 총 다섯 군데에만 IMU 부착
- 고전 모션 캡쳐 연구에서는 IMU를 최소 20군데 이상 부착
- Video 데이터를 사용하여 모션 모형을 복원하면 IMU의 정확한 위치를 알고 있으므로 drift-free한 정확도를 얻을 수 있음.
- 사지가 뻗어나가는 부위와 몸통 부위에만 IMU를 부착하여 인체 모형의 생김새를 확정할 수 있음.

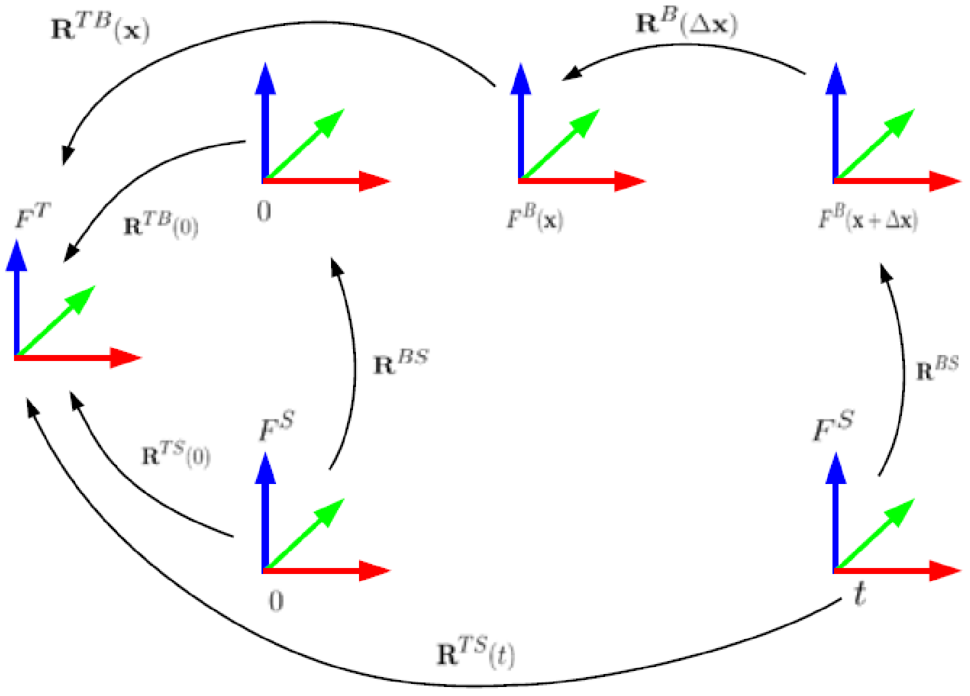


데이터 결합 원리

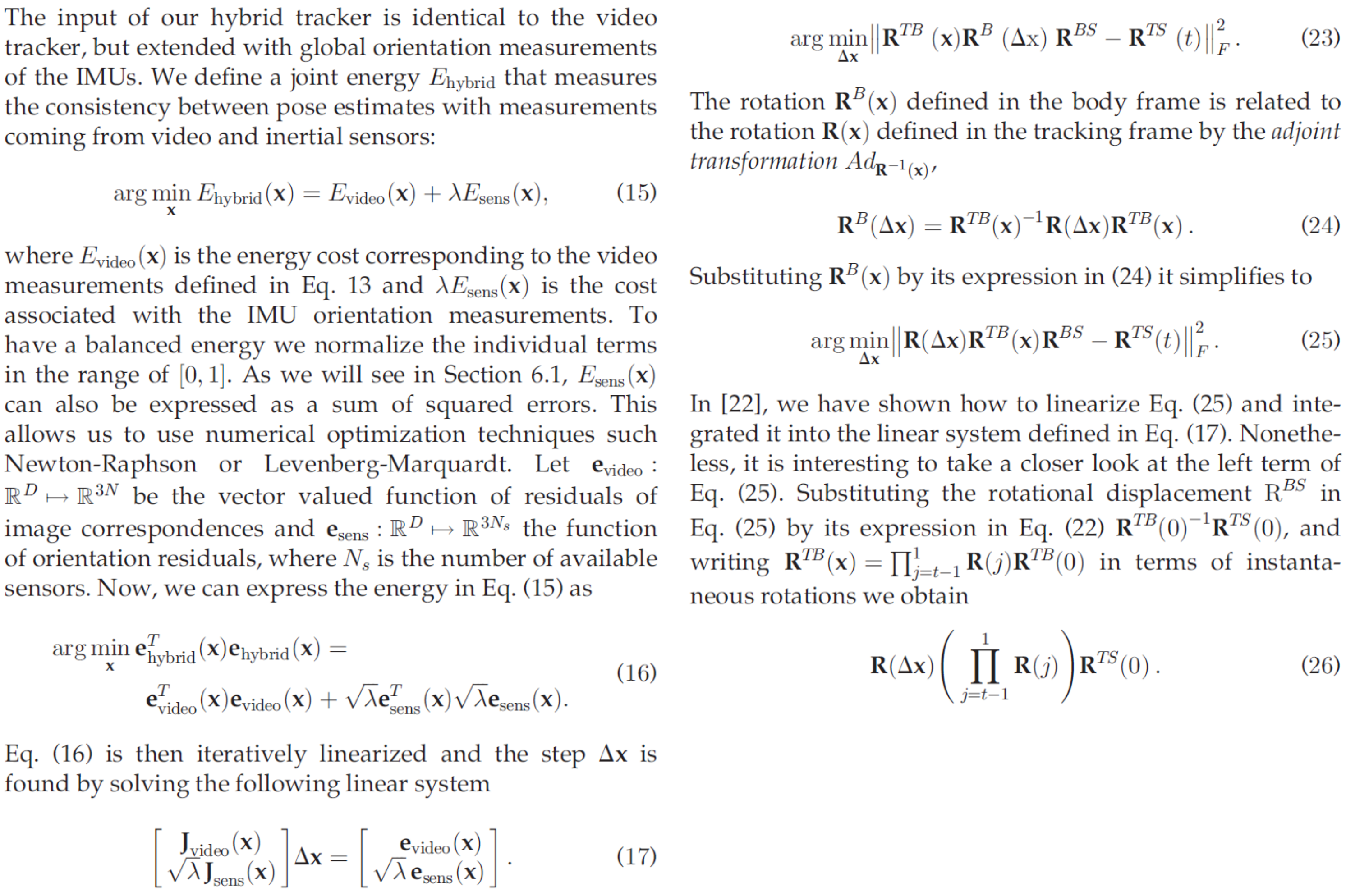

Background subtraction
Camera parameters
Pose
Adapt pose to maximize similarity

데이터 결합 원리


$\hat{\mathbf{R}}^{TS}(\mathbf{x})$
$\mathbf{R}^{TB}(\mathbf{x})$
F^B
F^B
F^T
$\mathbf{R}^{TB}(0)$
0
t
$\mathbf{R}^{BS}$
$\mathbf{R}^{BS}$
F^S
F^S
$\mathbf{R}^{TS}(0)$
0
t
$\mathbf{R}^{TS}(t)$


$\mathbf{R}^{TB}(\mathbf{x})$
$\mathbf{R}^{B}(\Delta\mathbf{x})$
F^T
$\mathbf{R}^{TB}(0)$
0
$F^B(\mathbf{x})$
$F^B(\mathbf{x}+\Delta\mathbf{x})$
$\mathbf{R}^{BS}$
$\mathbf{R}^{BS}$
F^S
F^S
$\mathbf{R}^{TS}(0)$
0
t
$\mathbf{R}^{TS}(t)$

데이터 결합 원리



The input of our hybrid tracker is identical to the video tracker, but extended with global orientation measurements of the IMUs. We define a joint energy E_{hybrid} that measures the consistency between pose estimates with measurements coming from video and inertial sensors:

$$\arg\min_{\mathbf{x}} E_{\text{hybrid}}(\mathbf{x}) = E_{\text{video}}(\mathbf{x}) + \lambda E_{\text{sens}}(\mathbf{x}), \tag{15}$$

where $E_{\text{video}}(\mathbf{x})$ is the energy cost corresponding to the video measurements defined in Eq. 13 and $\lambda E_{\text{sens}}(\mathbf{x})$ is the cost associated with the IMU orientation measurements. To have a balanced energy we normalize the individual terms in the range of $[0, 1]$. As we will see in Section 6.1, $E_{\text{sens}}(\mathbf{x})$ can also be expressed as a sum of squared errors. This allows us to use numerical optimization techniques such Newton-Raphson or Levenberg-Marquardt. Let $\mathbf{e}_{\text{video}} : \mathbb{R}^D \mapsto \mathbb{R}^{3N}$ be the vector valued function of residuals of image correspondences and $\mathbf{e}_{\text{sens}} : \mathbb{R}^D \mapsto \mathbb{R}^{3N_s}$ the function of orientation residuals, where N_s is the number of available sensors. Now, we can express the energy in Eq. (15) as

$$\begin{aligned} \arg\min_{\mathbf{x}} \mathbf{e}^T_{\text{hybrid}}(\mathbf{x})\mathbf{e}_{\text{hybrid}}(\mathbf{x}) &= \\ \mathbf{e}^T_{\text{video}}(\mathbf{x})\mathbf{e}_{\text{video}}(\mathbf{x}) &+ \sqrt{\lambda}\mathbf{e}^T_{\text{sens}}(\mathbf{x})\sqrt{\lambda}\mathbf{e}_{\text{sens}}(\mathbf{x}). \end{aligned} \tag{16}$$

Eq. (16) is then iteratively linearized and the step $\Delta\mathbf{x}$ is found by solving the following linear system

$$\begin{bmatrix} \mathbf{J}_{\text{video}}(\mathbf{x}) \\ \sqrt{\lambda}\mathbf{J}_{\text{sens}}(\mathbf{x}) \end{bmatrix} \Delta\mathbf{x} = \begin{bmatrix} \mathbf{e}_{\text{video}}(\mathbf{x}) \\ \sqrt{\lambda}\,\mathbf{e}_{\text{sens}}(\mathbf{x}) \end{bmatrix}. \tag{17}$$

$$\arg\min_{\Delta\mathbf{x}} \left\| \mathbf{R}^{TB}(\mathbf{x})\mathbf{R}^{B}(\Delta\text{x})\,\mathbf{R}^{BS} - \mathbf{R}^{TS}(t) \right\|^2_F. \tag{23}$$

The rotation $\mathbf{R}^B(\mathbf{x})$ defined in the body frame is related to the rotation $\mathbf{R}(\mathbf{x})$ defined in the tracking frame by the *adjoint transformation* $Ad_{\mathbf{R}^{-1}(\mathbf{x})}$,

$$\mathbf{R}^B(\Delta\mathbf{x}) = \mathbf{R}^{TB}(\mathbf{x})^{-1}\mathbf{R}(\Delta\mathbf{x})\mathbf{R}^{TB}(\mathbf{x}). \tag{24}$$

Substituting $\mathbf{R}^B(\mathbf{x})$ by its expression in (24) it simplifies to

$$\arg\min_{\Delta\mathbf{x}} \left\| \mathbf{R}(\Delta\mathbf{x})\mathbf{R}^{TB}(\mathbf{x})\mathbf{R}^{BS} - \mathbf{R}^{TS}(t) \right\|^2_F. \tag{25}$$

In [22], we have shown how to linearize Eq. (25) and integrated it into the linear system defined in Eq. (17). Nonetheless, it is interesting to take a closer look at the left term of Eq. (25). Substituting the rotational displacement R^{BS} in Eq. (25) by its expression in Eq. (22) $\mathbf{R}^{TB}(0)^{-1}\mathbf{R}^{TS}(0)$, and writing $\mathbf{R}^{TB}(\mathbf{x}) = \prod_{j=t-1}^{1} \mathbf{R}(j)\mathbf{R}^{TB}(0)$ in terms of instantaneous rotations we obtain

$$\mathbf{R}(\Delta\mathbf{x})\left(\prod_{j=t-1}^{1} \mathbf{R}(j)\right)\mathbf{R}^{TS}(0). \tag{26}$$

5

Tracking Error Analysis

센서 Error 결과

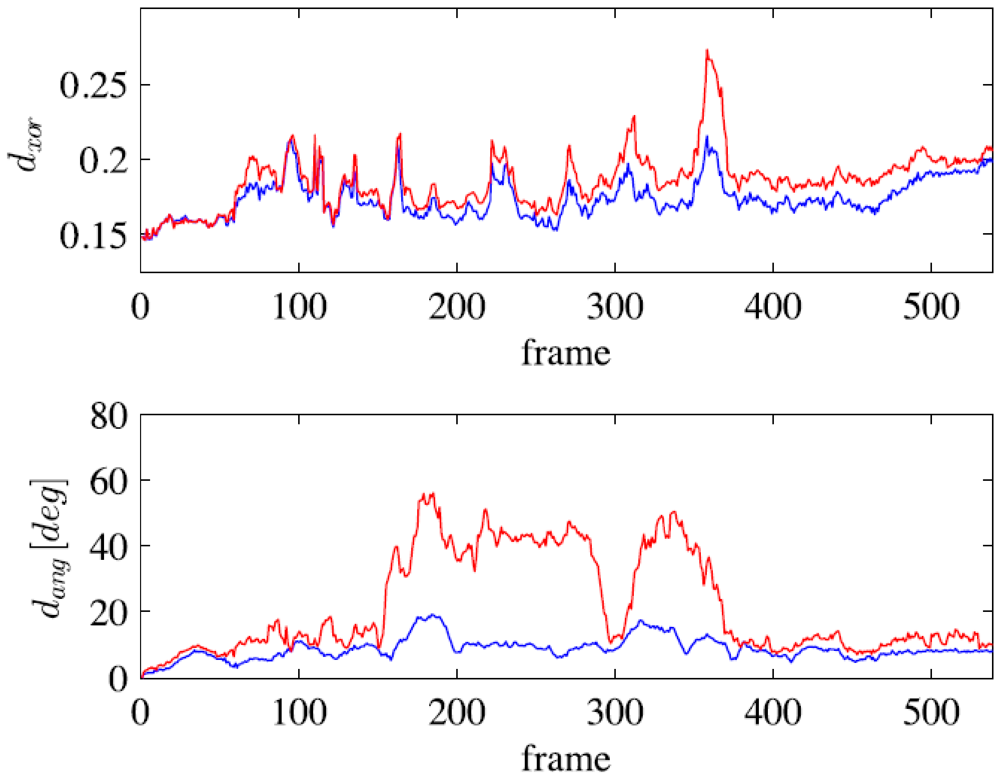


Fig. 8. Frame-wise XOR and orientation error for a walking sequence. The hybrid tracker (blue) performs well for the entire sequence. The video tracker (red) shows some large orientation errors between frames 160 and 370. Interestingly, this is almost invisible in the XOR error curve.

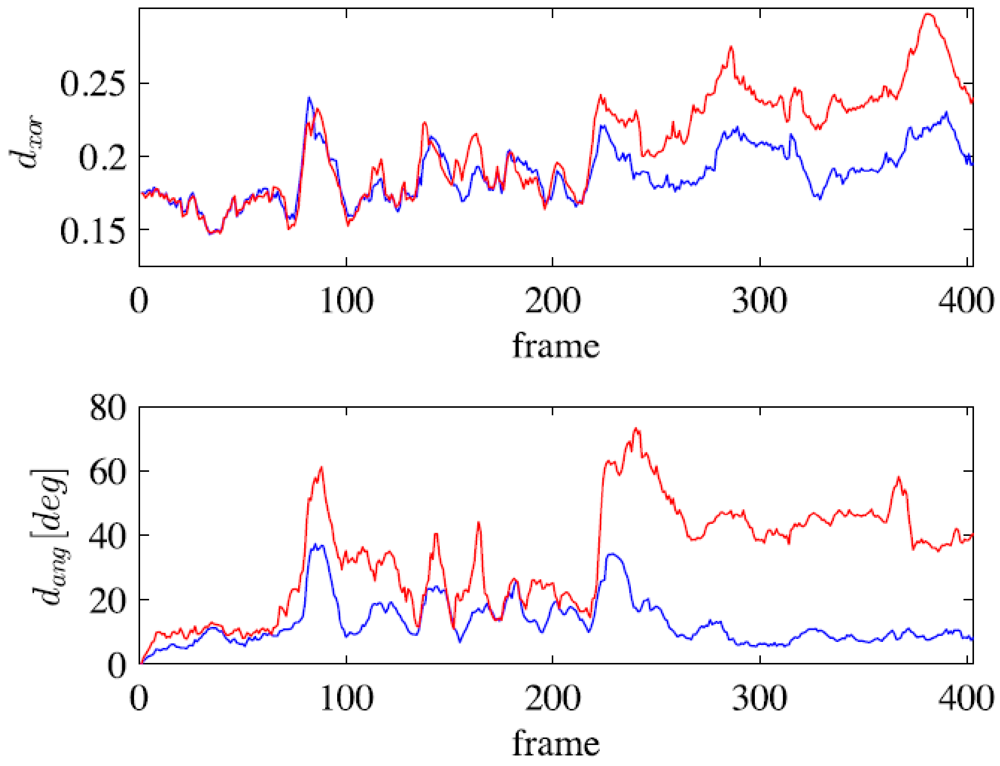


Fig. 9. Frame-wise XOR and orientation error for a dynamic punching sequence. The video tracker (red) struggles to track the complex motion and cannot recover from frame 210 on. The hybrid tracker (blue) performs better with respect to both error metrics.

연구 결과

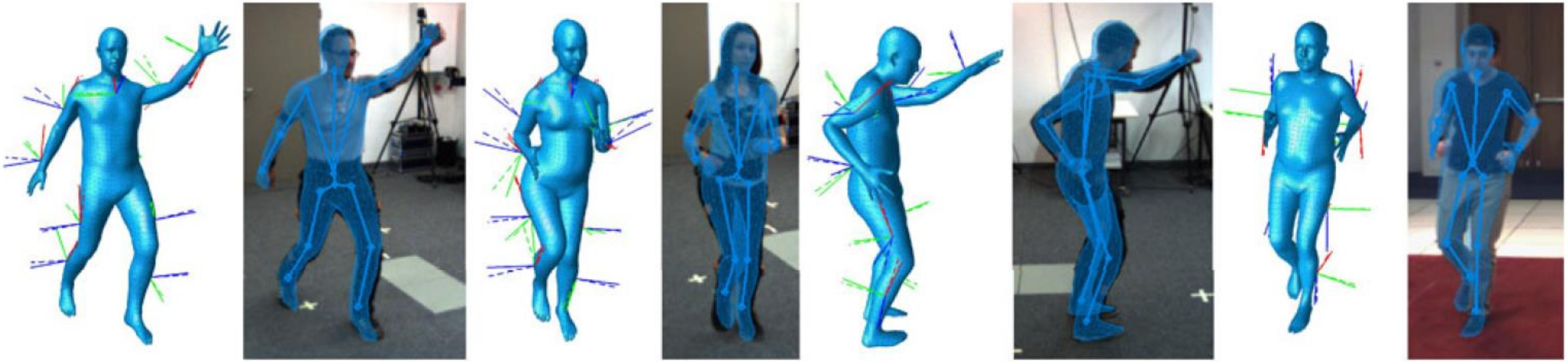
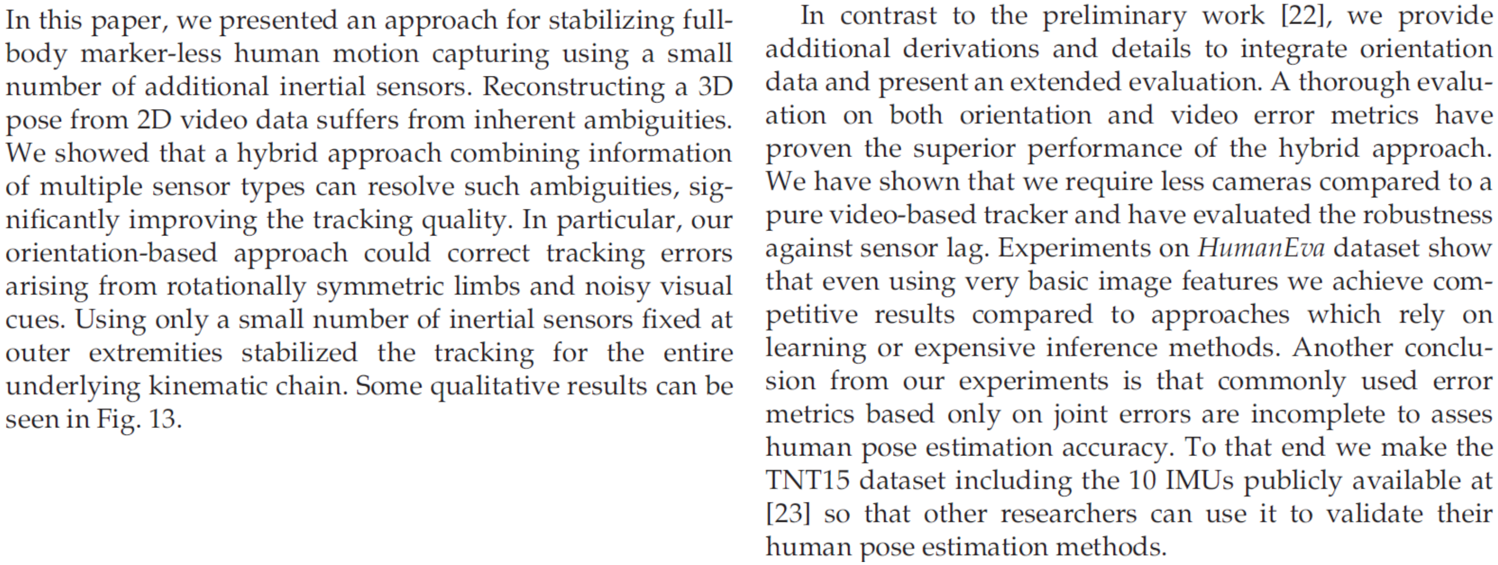


Fig. 13. We show three exemplary frames of the *TNT15* dataset and one of *HumanEva* (right most example). Each frame is is illustrated by two images. The images on the left depict the estimated model pose; ground-truth orientations are shown in solid lines and estimated orientations in dashed lines. The right images show the mesh projected to the respective RGB-image. For the *TNT15* sequences we have used a lower resolution mesh for tracking, which is actually visible in the respective RGB-images.

6

결론

결론



In this paper, we presented an approach for stabilizing full-body marker-less human motion capturing using a small number of additional inertial sensors. Reconstructing a 3D pose from 2D video data suffers from inherent ambiguities. We showed that a hybrid approach combining information of multiple sensor types can resolve such ambiguities, significantly improving the tracking quality. In particular, our orientation-based approach could correct tracking errors arising from rotationally symmetric limbs and noisy visual cues. Using only a small number of inertial sensors fixed at outer extremities stabilized the tracking for the entire underlying kinematic chain. Some qualitative results can be seen in Fig. 13.

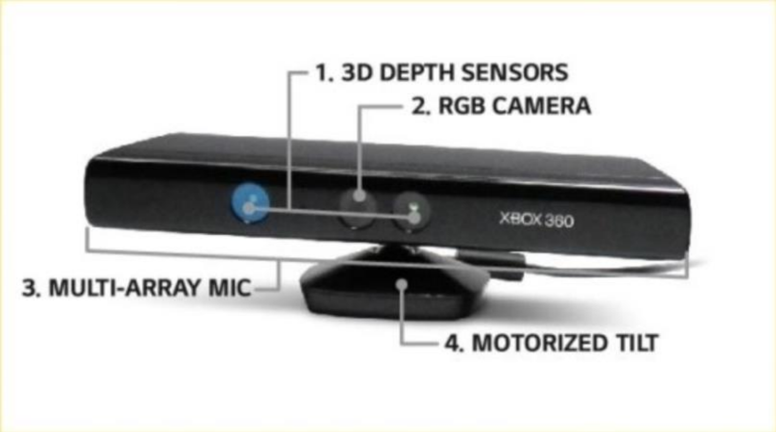
In contrast to the preliminary work [22], we provide additional derivations and details to integrate orientation data and present an extended evaluation. A thorough evaluation on both orientation and video error metrics have proven the superior performance of the hybrid approach. We have shown that we require less cameras compared to a pure video-based tracker and have evaluated the robustness against sensor lag. Experiments on *HumanEva* dataset show that even using very basic image features we achieve competitive results compared to approaches which rely on learning or expensive inference methods. Another conclusion from our experiments is that commonly used error metrics based only on joint errors are incomplete to asses human pose estimation accuracy. To that end we make the TNT15 dataset including the 10 IMUs publicly available at [23] so that other researchers can use it to validate their human pose estimation methods.

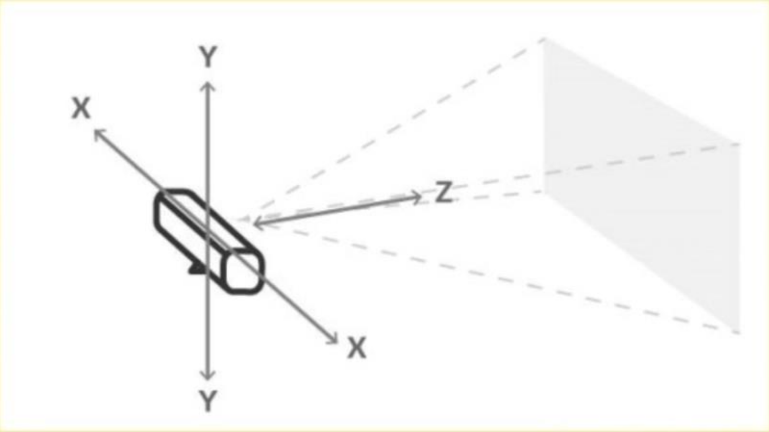
7

향후 논문리뷰 및 연구 계획

Pose 인식

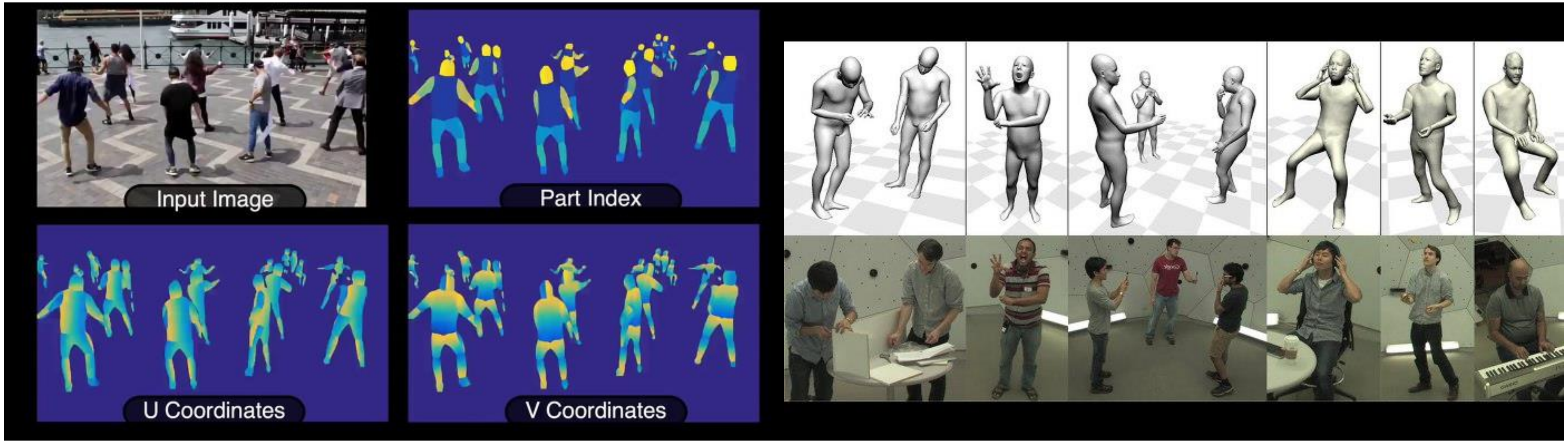
- single-camera를 사용할 수 있도록 기존 PoseNet 모델 활용
- RGB 카메라 + 적외선 카메라를 활용한 Kinect 방식 활용





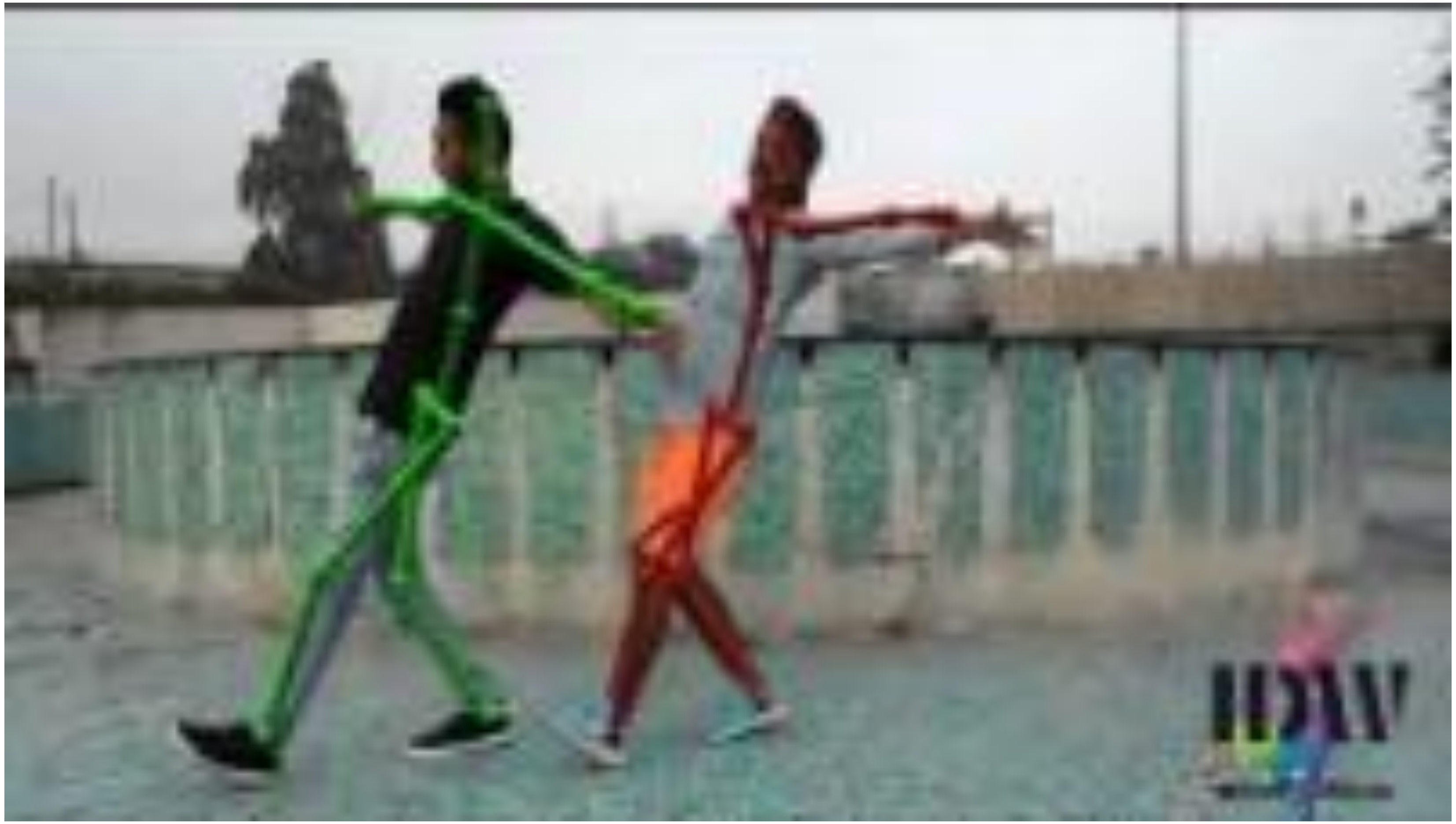
Pose 인식

- 2D input을 3D 표면에 매핑하는 DensePose 모델 활용



Pose 인식

- Xnect: Real-time Multi-person 3D Motion Capture with a Single RGB Camera



IMU 적용

- IMU에서 방향벡터를 뽑아내는 방법에 관한 연구
- 무릎이나 발목 중 하나의 관절 좌표만 인식돼도 정강이의 방향벡터가 있다면, 한 점으로부터 해당 방향으로의 일직선 내에 있는 다른 관절의 좌표를 유추할 수 있음.
- 국내 문헌 중 아두이노, 라즈베리파이 등에서 IMU를 사용하는 예제가 존재함.

감사합니다

2021.05.20.

고낙헌