Title:
Deep Learning–based Pose and Shape Estimation of Human Body at Disaster Site Utilizing Synthetic Disaster Scene Generation

Authors:
Ken Nishioka, nishioka@ist.hokudai.ac.jp, Hokkaido University
Zechen Zhu, Zhu-Zechen@outlook.com, Hokkaido University
Satoshi Kanai, kanai@ssi.ist.hokudai.ac.jp, Hokkaido University
Hiroaki Date, hdate@ssi.ist.hokudai.ac.jp, Hokkaido University
Atsushi Konno, konno@ssi.ist.hokudai.ac.jp, Hokkaido University
Soichi Murakami, so-ichi@umin.ac.jp, Hokkaido University Hospital
Toshiaki Shichinohe, shichino@med.hokudai.ac.jp, Hokkaido University Hospital

Keywords:
Human pose and shape estimation, Deep Learning, Synthetic dataset, Disaster Medicine.

DOI: 10.14733/cadconfP.2025.166-171

Introduction:
During large-scale disasters, people are often left behind under the debris of collapsed buildings. In such situations, rescue and emergency medical services must infer the position and pose of the survivor and determine which body parts are caught in the debris. On the other hand, methods for estimating 3D human body pose and shape (HPS) from images using deep learning have been actively researched [6].

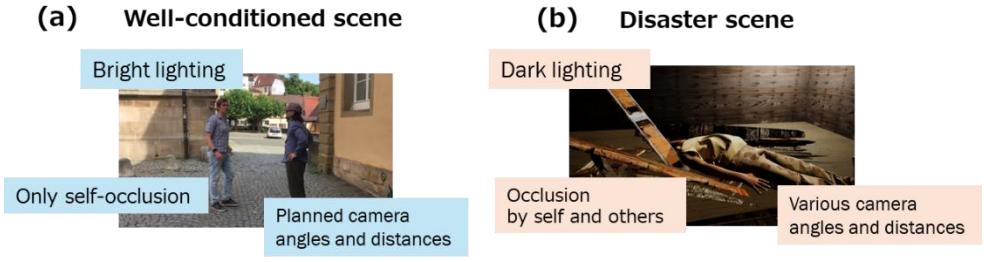

Fig. 1: Differences between well-conditioned and disaster scenes: (a) an image from the 3DPW dataset [5] and (b) an image from the "Disaster" dataset constructed herein.

However, most deep learning–based methods for HPS estimation from images are trained only on everyday and well-conditioned scenes (Fig. 1a). This makes them less effective when human bodies are partially occluded by debris and under irregular lighting conditions (Fig. 1b). Another critical issue is that it is very challenging to collect large amounts of training data from actual disaster sites and annotate human poses and shapes; the associated ethical considerations further complicate such data collection. Therefore, a 3D HPS estimation method that can stably estimate HPS from disaster-site images using deep learning must be developed, and a method that would systematically and efficiently generate training datasets for this purpose must be devised.

The majority of current deep learning–based HPS estimation methods [6] utilize the Skinned Multi-Person Linear Model (SMPL) [3], which is a parametric model of the human body, to construct learning models that estimate the HPS parameters. This requires collecting ground-truth HPS parameter values in real-world environments, which is a labor-intensive and error-prone process. To address this issue, Black et al. [1] have recently proposed a large-scale synthetic dataset, BEDLAM, for 3D HPS estimation. BEDLAM enables the efficient construction of large synthetic datasets of scenes containing human bodies in various poses and the corresponding SMPL pose and shape parameter values. Furthermore, it has been shown that deep learning models for HPS estimation trained only on BEDLAM can achieve the same or better estimation accuracy than models trained on datasets consisting only of authentic images. However, this dataset is designed for HPS estimation only in general indoor and outdoor scenes: it does not contain scenes with collapsed structures and scattered debris appearing in typical disaster sites.

The purpose of the present study was to develop a prototype system that can automatically generate images of human bodies in disaster scenes and the corresponding HPS annotations. To this end, we employed a parametric 3D model of the human body and a game engine for simulating an indoor disaster scene caused by an earthquake, using the BEDLAM approach as a reference. We then investigated the effectiveness of these synthetic images as a training dataset for improving the stability and accuracy of HPS estimation.

Simulation of the disaster scene and generation of synthetic data:

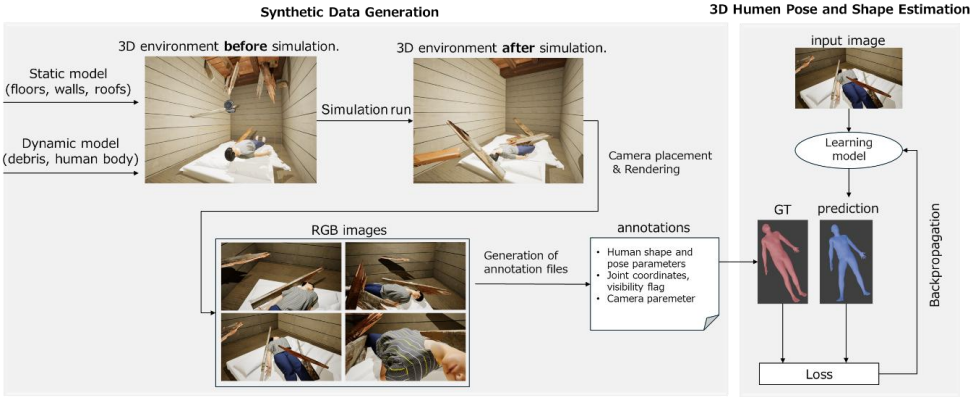

Fig. 2: Generation of synthetic training data and 3D HPS estimation for disaster scenes.

Figure 2 shows the overview of our simulation system for the construction of synthetic disaster scenes using a game engine (Unreal Engine, UE [4]) and the synthesis of artificial training data for an HPS-estimation deep learning model. The simulation flow consists of the following steps.

Preparation of the 3D model data

- **Environmental model:** Three types of floors, two types of *beds*, one type of *roof,* and six types of *walls* were constructed. The beds and roofs were assembled from 3D meshes and textures, while the other models (floors, walls) were assembled from textures only.
- **Human body models:** One male and one female template model with SMPL shape parameters were constructed based on BEDLAM models. Then, these template models were deformed by specifying the values of SMPL pose parameters that define the angles of the 23 joints in the human body and the orientation of the body. The models were initially converted from the Python binary format (.npz) to the FBX format and imported into UE as skeletal meshes (3D models with skeletons).
- **Clothing and hair models:** To ensure a variety of appearances, a 3D clothing model with 10 clothing texture patterns and 12 skin texture patterns was prepared for each gender. The 3D hair model was separately attached to the human body model with manual adjustment.

- **Debris (destroyed wood, concrete blocks) models:** Assuming the collapse of an ordinary house under a severe earthquake, 115 models of destroyed members, such as wooden pillars, wooden boards, and concrete blocks of various shapes and sizes, were prepared.

Simulation of the affected scene

(1) **Arrangement of the static models:** An eight-square-meter room is initially constructed by appropriately arranging the floor, walls, and ceiling, and a bed is placed inside the room. The floor aspect ratio of the room can be changed under the floor area constraint. The objects are treated as *static* objects that are stationary during the simulation.
(2) **Initial placement of dynamic models:** A human body model (gender, skin, and clothing textures randomly selected) is placed at a random position in the room. In addition, 0–100 debris models are randomly placed in different positions and orientations in the room. Each piece of debris is subjected to gravity proportional to its volume.
(3) **Running the simulation:** The physics simulation in UE is run to reproduce the human body's fall and the collisions among debris. The simulation stops after 10 s, and the scene data is acquired after all the objects are at a standstill.
(4) **Camera placement and rendering:** Ten cameras are randomly placed around the constructed scene, and RGB images are rendered. Figure 3 shows examples of the generated images.
(5) **Generation of an annotation dataset for deep learning–based HPS estimation:** The SMPL parameter values are archived into the annotation dataset. In addition, the visibility flags for each joint from a camera and the camera parameters are logged into the dataset.

By repeating the above process, many training samples (RGB images, SMPL parameter values, visibility flags, and camera parameters) can be automatically generated and stored in the annotation dataset.

Fig. 3: Examples of generated images of disaster scenes.

Deep learning–based HPS estimation using synthetic disaster scenes:

This section describes the experiments showing how the estimation performance of the existing 3D HPS estimation network can be improved by retraining it on the synthetic image datasets of disaster scenes generated using the developed system.

Herein, experiments were conducted using BEDLAM [1] as a reference. BEDLAM uses a deep learning model called CLIFF [2] for HPS estimation; thus, herein, CLIFF was also used as the HPS estimation model.

CLIFF is a framework that estimates the SMPL parameters [3] defining 3D HPS from the input images (Fig. 4). In this framework, an object detection process first detects the position of the human in the image. Next, the image of the human part is input into the CLIFF, and the 3D HPS fitted to the image is estimated. By inputting these parameters into SMPL, the 3D coordinates of the epidermal mesh vertices and joint positions can be estimated. When rendering the human body shape as an image, the mesh vertices and joints are projected onto the image using the camera parameters.

Dataset

A summary of the datasets used herein is provided in

Tab. **1**. The publicly available 3DPW dataset [5] for HPS estimation and our synthetic “Disaster” dataset were used together. 3DPW is a dataset consisting of approximately 51,000 single-camera images of people in different poses and corresponding 3D annotations obtained from inertial measurement units attached to limbs. In our experiments, 3DPW was used for training and validation, but not for testing.

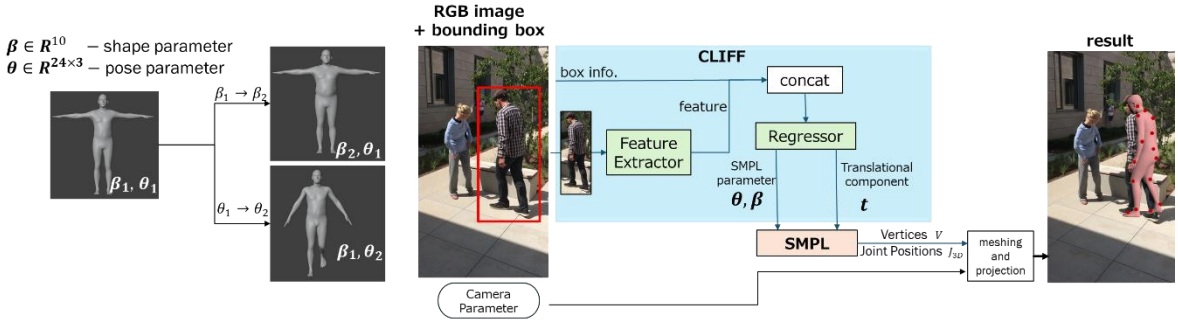

Fig. 4: (Left) SMPL parameter description. (Right) Process flow of 3D HPS estimation with CLIFF (figure created based on [2]).

Disaster-Sim is a dataset constructed as described above; it was used for training and validation. Additionally, the *Disaster-Manual* dataset was created through the manual placement of 3D models and used for testing because manual placement allows to create more natural scenes than those in *Disaster-Sim*.

In addition, real-world image datasets, *Real-DebrisField* and *Real-Lab*, were used as test data. *Real-DebrisField* contains images taken at a simulated debris field in *the Hirosaki University of Health and Welfare Junior College USAR facility*. These images were taken under the assumption of a disaster situation in which a survivor's body is sandwiched between concrete blocks, with the entire body not necessarily visible because of the blocks. *Real-Lab* contains images of two subjects recorded at different occlusion rates (100%, 90%, 80%, 60%, 50%, 40%, 20%, and 0%). The images of the subjects were taken from the same position, and only the occlusion board was moved. After shooting, the 2D joint positions of the subjects and bounding boxes were manually determined.

Dataset	#Samples	Used for	Annotation
3DPW [29]	22,735	Training, Validation	Fully
Disaster-Sim	7,010	Training, Validation	Fully
Disaster-Manual	802	Testing	Fully
Real-DebrisField	17	Testing	None
Real-Lab	69	Testing	2D joint position only

Tab. 1: Details on the datasets used in the experiment.

Learning model

CLIFF [2] was used as the learning model. CLIFF consists of a feature extractor and regressor (Fig. 4, right). The *f*eature extractor is a deep learning–based 2D pose estimation model, and the regressor consists of multiple fully connected layers and dropout layers. In this experiment, the feature extractor and regressor were fine-tuned simultaneously.

Evaluation metrics

Mean per joint position error (MPJPE), Procrustes-aligned mean per joint position error (PA-MPJPE), and mean per vertex position error (MPVPE) have been previously used to assess the performance of 3D HPS estimation [6]. Thus, herein we also adopted these evaluation metrics for the *Disaster-Manual* dataset. At the same time, the 2D joint error L_{2D} (mean value of absolute error of each joint position on the image) was used as the evaluation metric on the *Real-Lab* dataset.

Results

The model trained on *Disaster-Sim + 3DPW* shows the most accurate results by almost all metrics, with an MPJPE of 172.04 mm. In contrast, the HPS model trained on only 3DPW exhibits notably worse performance, with an MPJPE of 449.26 mm. In addition, models trained on *Disaster-Sim* or *Disaster-Sim + 3DPW* estimated human body shapes and poses better than those trained on 3DPW alone. This result indicates that existing HPS datasets, such as 3DPW, are insufficient for accurate pose and shape estimation from images of human bodies buried in debris, as observed in *Disaster-Manual*. Therefore, the proposed synthetic dataset is more effective for training the models to estimate the HPS of bodies buried in debris.

datasets	*Disaster-Manual* Evaluations [mm]			*Real-Lab* L_{2D} [pixel]			
	MPJPE↓	*PA-MPJPE*↓	*MPVPE*↓	Face, left↓	Back, left↓	Face, Front↓	Back, Front↓
3DPW	449.26	150.17	496.38	115	181	156	158
Disaster-Sim	180.06	113.62	211.67	**63**	89	65	59
3DPW +Disaster-Sim	**172.04**	**107.61**	**202.43**	71	**87**	**62**	**56**

Tab. 2: Evaluation metrics for 3D HPS estimation on different datasets.

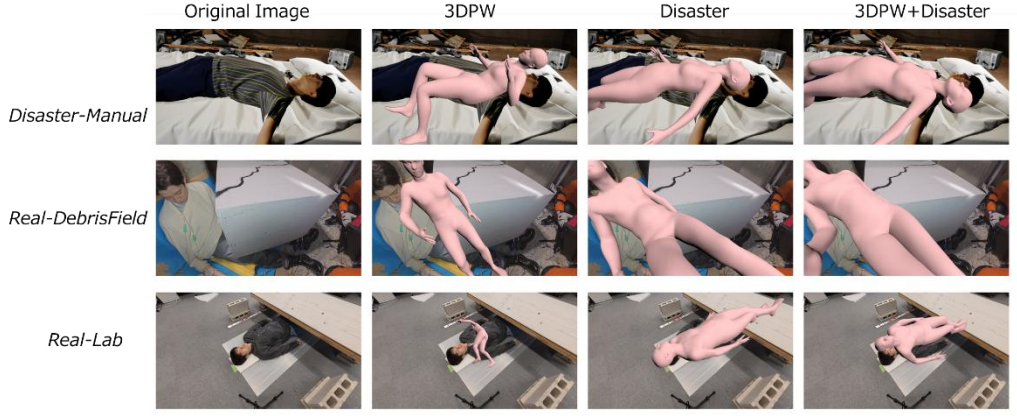

Fig. 5: Example estimation results produced by models trained on each of the three datasets.

The results of pose and shape estimation on actual images (*Real-DebrisField* and *Real-Lab*) are shown in Fig. 5. The model trained on 3DPW yielded inadequate results for the supine pose and human shapes on *Real-DebrisField*. In contrast, the model trained on *Disaster-Sim* yielded a more accurate supine pose. However, the errors of joint angle estimation on the actual images remain significant for both models, indicating the need for further improvement.

The model trained on *Disaster-Sim* achieved better accuracy on *Real-Lab* than the one trained on 3DPW. In addition, the position of the subject's head estimated by the HPS model trained on 3DPW is notably misaligned (Fig. 5). This misalignment was attributed to a small number of images where the subject is lying down in the 3DPW dataset. At the same time, most of the subjects in the *Disaster-Sim* images are in lying poses. Thus, the model trained on this dataset yielded more accurate results than models trained on data that included subjects only in everyday standing poses. Therefore, the proposed synthetic dataset is effective for training models to determine HPS from real-world images involving subjects in lying poses.

Conclusions:

Herein, we developed a synthetic data generation system to simulate disaster scenes. The system employs a game engine to automatically synthesize large datasets of diverse images containing models of human bodies under collapsed rooms and debris, which would be difficult to collect in a real environment. Furthermore, using a parametric human model, the proposed system generates the ground-truth HPS data. Importantly, a CLIFF deep learning model trained on the simulated dataset showed higher HPS estimation accuracy than those trained on real-world datasets.

Future studies should further improve the accuracy of HPS estimation from disaster site images. In addition, manual adjustments should be automated to ensure variations in clothing and hairstyles and reproduce the natural poses of the human body. Furthermore, as the errors in joint angles estimated from real disaster-site images remain considerable even for the model trained on the proposed dataset, future studies should augment the proposed dataset through the addition of shielding conditions closer to real environments.

Ken Nishioka, https://orcid.org/0009-0003-2185-0290
Zechen Zhu, https://orcid.org/0009-0006-2907-8428
Satoshi Kanai, https://orcid.org/0000-0003-3570-1782
Hiroaki Date, https://orcid.org/0000-0002-6189-2044
Konno Atsushi, https://orcid.org/0000-0003-3288-8844
Soichi Murakami, https://orcid.org/0000-0003-2227-9367
Toshiaki Shichinohe, https://orcid.org/0000-0001-6614-462X

References:

[1] Black, M.J; Patel, P.; Tesch, J.; Yang, J.: BEDLAM: A synthetic dataset of bodies exhibiting detailed lifelike animated motion, 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, 8726-8737. https://doi.org/10.1109/CVPR52729.2023.00843

[2] Li, Z.; Liu, J.; Zhang, Z.; Xu, S.; Yan, Y.: CLIFF: Carrying location information in full frames into human pose and shape estimation, Proceedings of the European conference on computer vision (ECCV), 13665, 2022, 590-606. https://doi.org/10.1007/978-3-031-20065-6_34

[3] Loper, M.; Mahmood, N.; Romero, J.; Pons-Moll, G.; Black, M.J.: SMPL: A skinned multi-person linear model, Seminal Graphics Papers: Pushing the Boundaries, 2, 2023, 851-866. https://doi.org/10.1145/3596711.3596800

[4] Unreal Engine, https://www.unrealengine.com/, Epic Games.

[5] Von Marcard, T.; Henschel, R.; Black, M.J.; Rosenhahn, B.; Pons-Moll, G.: Recovering accurate 3d human pose in the wild using imus and a moving camera, Proceedings of the European Conference on Computer Vision (ECCV), 2018, 614–631. https://doi.org/10.1007/978-3-030-01249-6_37

[6] Zheng, C.; Wu, W.; Chen, C.; Yang, T.; Zhu, S.; Shen, J.; Kehtarnavaz, N.; Shah, M.: Deep learning-based human pose estimation: A survey, ACM Computing Surveys, 56(1), 2023, 1–37. https://doi.org/10.1145/3603618