2022 IEEE International Conference on Industrial Technology (ICIT) | 978-1-7281-1948-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICIT48603.2022.10002823

A Skeleton-based View-Invariant Framework for Human Fall Detection in an Elevator

Rashid Ali, Ivan Surya Hutomo, Lan-Da Van, Yu-Chee Tseng, *IEEE Fellow*
Department of Computer Science,
National Yang Ming Chiao Tung University,
Hsinchu, Taiwan
rashid.c@nycu.edu.tw, hutomoivan.eic08g@nctu.edu.tw, ldvan@cs.nctu.edu.tw, yctseng@cs.nycu.edu.tw

Abstract—**This paper considers the emergency behavior detection problem inside an elevator. As elevators come in different shapes and emergency behavior data are scarce, we propose a skeleton-based view-invariant framework to tackle the camera view angle variation issue and the data collection issue. The proposed emergency fall detection model only needs to be trained for a target camera, which is deployed in an elevator at a manufacture's lab, from which a large amount of training data can be collected. The deployment of a source camera, which is in a customer-side elevator, hence can be customized and almost no training effort is needed. Our framework works in four stages. First, a 2D RGB input image is taken from the source camera and a 2D human skeleton is obtained by 2D pose estimation (AlphaPose). Second, the 2D skeleton is converted to a 3D human skeleton by 3D pose estimation (3D pose baseline). Third, a pre-trained rotation-translation (RT) transform (Procrustes analysis (PA)) aligns the 3D pose representations to the target camera view. Finally, a dual 3D pose baseline deep neural networks (D3PBDNN) model for human fall detection is proposed to perform the recognition task. We gather a human fall detection dataset inside different elevators from various view angles and validate our proposal. Experimental results successfully attain almost equivalent accuracy to that of a source camera-trained model.**

Keywords—view-invariant, fall detection, 2D/3D pose estimation, deep neural network, Procrustes analysis, skeleton

I. Introduction

Elevators have become a common mode of transportation because they provide convenience and speed in our daily lives. Since an elevator is a closed and unnoticed environment, any emergency, such as falling, may have dire consequences. An intuitive way to recognize emergency behaviors inside an elevator is by video human pose estimation. However, even in the same elevator, accuracy may vary a lot because in a limited space, camera viewing angle may not be able to capture whole human body skeleton. On the other hand, it is difficult to derive one generalized recognition model to fit all types of elevators because elevators come in different shapes and emergency datasets are scarce. Intuitively, to attain high accuracy, collecting huge training data from all viewing angles for all types of elevators seems to be inevitable.

In reality, camera deployment can be highly customized by end elevator users. View angles of cameras are very likely to differ from one to the others. Training data for each specific environment is always scarce, especially for emergency behaviors. As a result, we are motivated to develop a general framework to attain single-view angle learning that can be transformed to different view angles. In this work, we propose a skeleton-based view-invariant framework that relies on 2D pose estimation, 3D pose estimation, rotation and translation (RT) transform, and fall detection model.

In the literature, camera calibration [1-4] is a widely used to approach for our purpose. The works in [1, 2, 4] propose multi-camera calibration methods. However, the drawbacks include the need of calculating a world coordinate of these cameras, which is a tricky issue, and the difficulty in calibrating the surveillance cameras with a checkerboard inside an elevator, which is an uninterruptable environment. The work [3] exploits Procrustes analysis (PA) to calibrate the position of a camera in a pedestrian area based on the 3D positions of pedestrians' heads and feet. These approaches require cameras and objects be close. Recently, several works [5-10] try to solve the view invariant problem via different approaches. It is noted that the pose estimation techniques [11-14] play an important role to the above solutions. In our work, we also apply PA on 3D skeletons, but the two cameras to be calibrated do not need to stay in the same elevator, and the elevators are not necessarily of the same type. We only train fall detection for a target camera and all other source cameras only need to transform their recognized skeletons to the former camera. In this way, we only require collecting data and train our model from the target camera's viewing angle, thus greatly accelerating industrial applications in elevator safety. The contributions of this work are as follows: 1) propose a skeleton-based view-invariant framework, 2) propose a dual 3D pose baseline deep neural networks (D3PBDNN) for human fall detection, and 3) target the elevator safety application.

II. Proposed Framework

We consider two cameras A and B such that A is the target camera and B is the source camera. A and B are deployed in target and source elevators, respectively (one may imagine common RGB camera, while B is another 2D camera. The whole proposed framework is shown in Fig. 1. that they are deployed in manufacturer and customer sides, respectively). All data collection and model training are done in the target elevator. Therefore, we assume that A is a common RGB camera, while B is another 2D camera. The whole proposed framework is shown in Fig. 1.

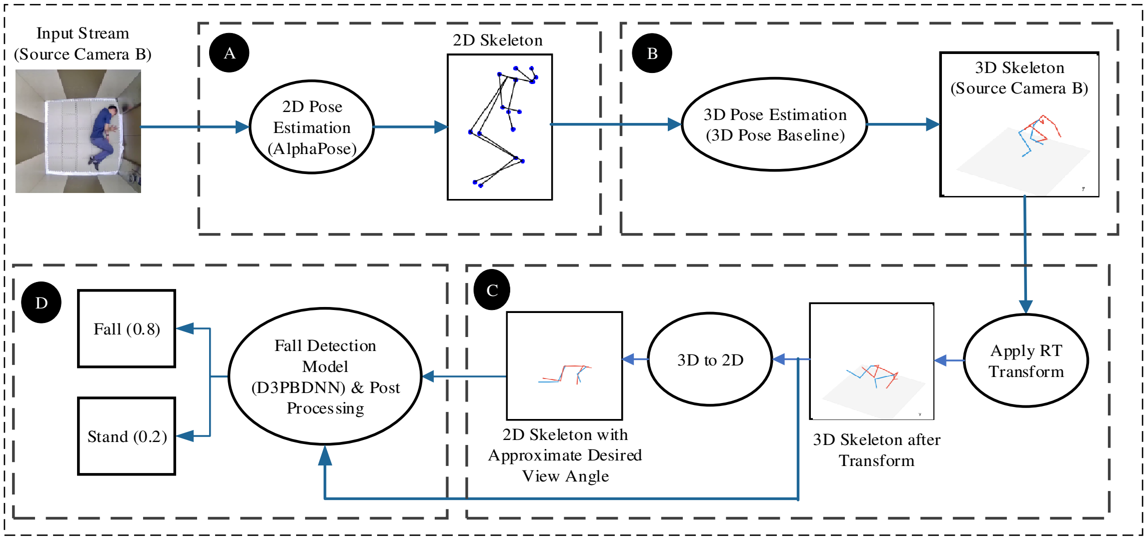

Input Stream
(Source Camera B)
A
2D Pose Estimation (AlphaPose)
2D Skeleton
B
3D Pose Estimation (3D Pose Baseline)
3D Skeleton (Source Camera B)
Apply RT Transform
3D Skeleton after Transform
C
3D to 2D
2D Skeleton with Approximate Desired View Angle
D
Fall Detection Model (D3PBDNN) & Post Processing
Fall (0.8)
Stand (0.2)

Fig. 1. The block diagram of the proposed skeleton-based view-invariant framework.

A. 2D Pose Estimation

Block A converts each RGB image taken by the source camera B to a 2D skeleton by AlphaPose [12]. We adopt AlphaPose because it is a top-down multi-person approach based on YOLO [15]. AlphaPose uses the COCO format with 17 joints.

B. 3D Pose Estiamtion

Block B converts 2D skeleton from the COCO format to the Human3.6M format [11]. The Human3.6M format has 17 joints locations per skeleton with a different indexing from the COCO format. Through use of index mapping, the Human3.6M model converts 2D skeleton to 3D skeleton by the multilayer neural network (called 3D pose baseline neural network [13]). The neural network is trained using the Human3.6M dataset that consists of both 2D and 3D skeleton ground truth in the Human3.6M skeleton format. We search for the best hyperparameters using the Optuna [16] and the results are shown in Table 1.

C. Rotation and Translation (RT) Transform

Fig. 2 elaborates how RT transform works to get the optimized mapping from source camera view to target camera view using PA. First, it takes two standing postures from target camera A and source camera B at once and then approximately matches the views of both cameras using PA. Next, the inference pipeline acquires rotation R and translation T values once the RT transform is learned. Finally, these values are applied to source camera view skeletons to match

them with target camera view skeletons. The PA [17-19] that has been widely used in shape comparing and matching in [7, 14, 20] uses singular value decomposition (SVD) to compare and match the shapes between the target and source objects by rotation and translation values. The details are elaborated as follows. First, we have two 3D skeletons from the target angle A and source angle B, and we find out the centers (centroidA and centroidB) of these two skeletons. Then, we re-center our skeletons and calculate covariance matrix H between them by (1), where P_a and P_b are skeleton points.

$$H= \sum_{i=1}^{n} (P_A^i \text{- centroid}_A)(P_B^i \text{- centroid}_B) \qquad (1)$$

Next, we decompose this covariance matrix H using SVD [21] to obtain the decomposition matrices U, S, and V in (2).

$$[U,S,V]=\mathrm{SVD}(H) \qquad (2)$$

After obtaining U, S, and V, we can find the rotation matrix, R, and the translation matrix, T, by (3) and (4), respectively.

$$R=V^*U^\wedge T \qquad (3)$$

$$T= -R^* \text{ centroid}_A + \text{centroid}_B \qquad (4)$$

T*ABLE* 1. HYPERPARAMETER SEARCH SCOPE AND RESULTS FOR 3D POSES BASELINE NEURAL NETWORK USING OPTUNA

Param Name	**Search Scope**			**Best Parameter**
	Low	***High***	***Step***	
LR	1.0e − 5	1.0e − 1	Continuous	0.0006495
LR Decay	50000	150000	50000	150000
Gamma	0.8	1	0.05	0.9
Linear Size	256	2048	256	1024
Dropout	0.3	0.8	0.1	0.3
Num Stage	1	6	1	5
Optimizer	Adam, RMSprop, SGD			Adam

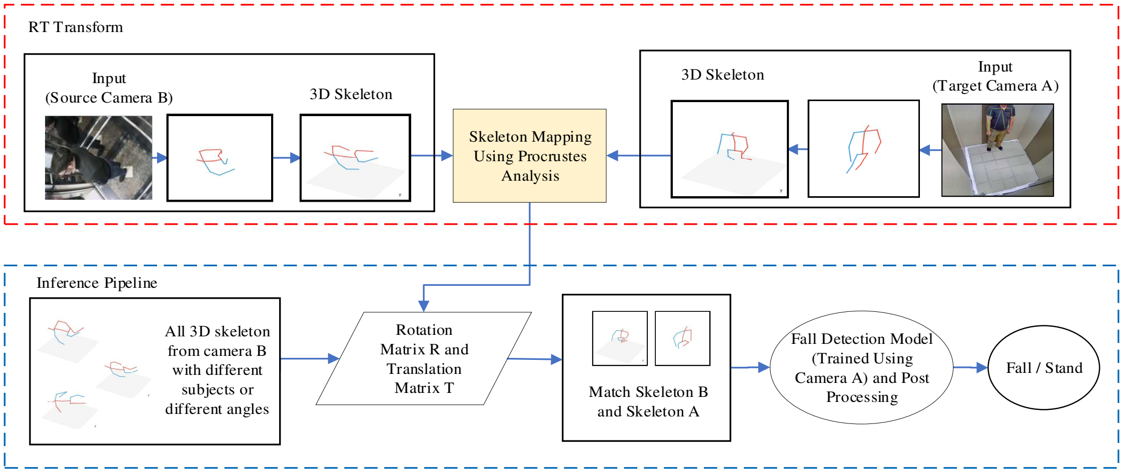


Fig. 2. The presented RT transform.

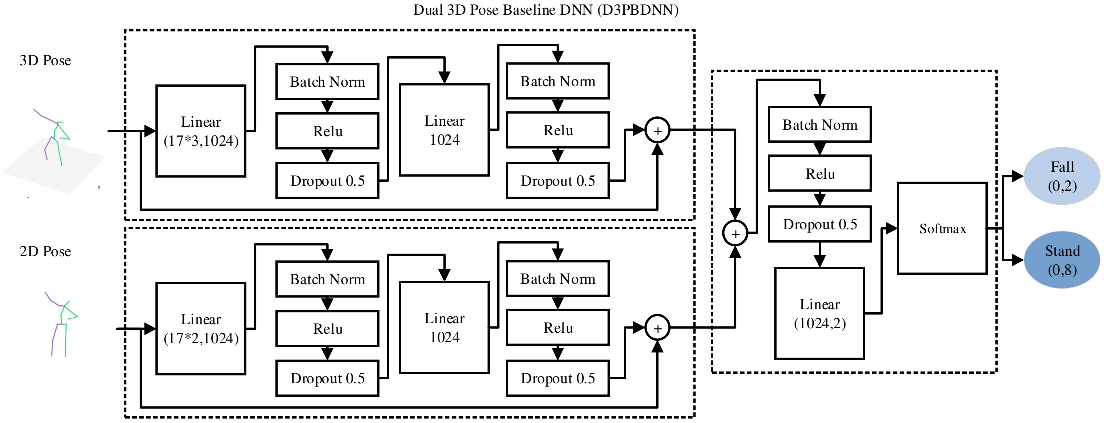


Fig. 3. The proposed fall detection network architecture.

Finally, once these transformation matrices are learned, Block C transforms the source camera skeleton B to target camera skeleton A as shown in Fig.1.

D. *Fall Detection Model and Post Processing*

Block D contains our proposed fall detection neural network and post processing for emergency behavior detection. The proposed D3PBDNN based on [13] consists of two modality-specific subnetworks in Fig. 3. It can process 2D and 3D pose representations simultaneously. We exploit 2D and 3D pose representations to add more information and extract rich features. Each subnetwork [13] consists of two stacks of the linear layer with batch normalization (BN), ReLU, and dropout. For each subnetwork, we add the residual connection to avoid information loss. Then, we concatenate the 1024-dimensional 2D and 3D features and feed them to a linear and SoftMax layer for predicting class scores. To further improve our result, we adopt the majority voting for the post processing. In our work, we use 5-voting queues for majority voting. If the number of falling detection is more than or equal to 3, we determine the current frame as fall. Similarly, standing uses this concept.

III. Results

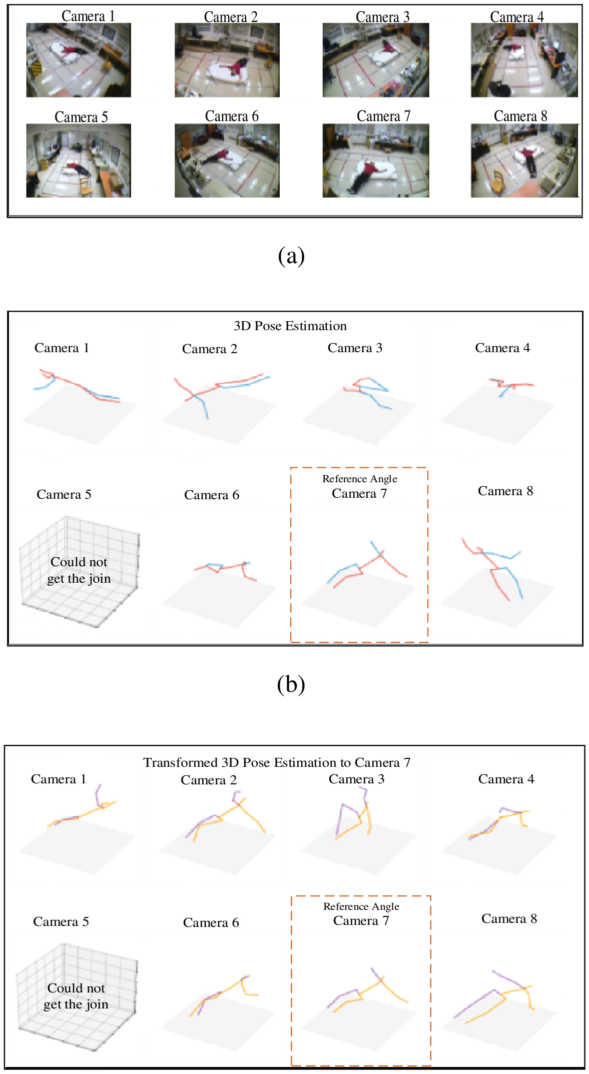
In this section, we evaluate our proposed framework employing quantitative analysis and qualitative comparison.

A. *3D Pose Estimation Evaluation*

During the training, we use a standard protocol of the Human3.6M dataset. Subjects 1, 5, 6, 7, and 8 are used for training, and on the other hand, Subjects 9 and 11 are used for evaluation [11]. We use the average error in millimeters between ground truth and prediction after central hip alignment. Through Table 1, we can reduce the millimeter error of 3D prediction from 45.5 mm to 40.3 mm using our optimized hyperparameters.

B. *Rotaton-Translation Transform Evaluation*

The RT transform is evaluated by multi cameras falling dataset [22]. This dataset has 24 scenarios recorder with 8 different camera views as shown in Fig. 4(a). The same action is recorded by 8 cameras from 8 different view angles. Hence, we can use this dataset to qualitatively evaluate the RT transform as shown in Fig. 4(b), where camera 7 is presumed



(a)

(b)

(c)

Fig. 4. The qualitative evaluation of the proposed RT transform.

TABLE 2. MEASUREMENT EVALUATION RESULTS OF FIG .4

Evaluation Metrics	**RT Transform**	**Average (All Cams)**	
Cosine Similarity (Larger is better)	Before	0.28	30.43%
	After	0.92	100%
Euclidian Distance (Smaller is better)	Before	1826.75	100%
	After	628.20	34.39%
MJPE Error (Smaller is better)	Before	386.21	100%
	After	123.99	32.10%

as a target. In Fig. 4(c), we apply RT transform to match the angle of target camera 7 with other source cameras. The quantitative evaluation is measured by the average of cosine similarity, Euclidean distance, and mean per joint position error (MJPE) before and after RT transform as shown in Table 2. The results show that the presented RT transform improves cosine similarity, Euclidean distance, and MJPE by 69.57%, 65.61%, and 67.9%, respectively. That means the presented approach can successfully approximate the target view angle.

C. *Fall Detection Evaluation*

We train a fall detection model using the untransformed 2D and 3D pose representations from the target camera angle A. The model takes both 2D and 3D pose representations as input. Then, we evaluate the model with 2D and 3D pose representations from different camera view angles B, C, D and E. We took angles A, B, and C from our custom dataset inside elevators. For angle D, we use footage from the Le2I dataset [23]. We use multiple cameras fall dataset [22] for angle E. For a fair comparison, we evaluate the fall detection model with untransformed and transformed pose representations in terms of accuracy, precision, recall, and F1-Score per image frame as shown in Table 3. The experimental result shows that, during the inference phase, through the RT transform, the D3PBDNN model can attain an accuracy by 0.83 at least in Table 3.

D. *Comparison*

Our qualitative comparison with the selected skeleton-based view-invariant action recognition frameworks is shown in Table 4. Since each work uses a different approach and dataset and our work focuses on fall detection inside the elevator, it will not be fair if the proposed framework is compared by head-to-head. Thus, we qualitatively compare our proposed framework with the previous frameworks. The work [5] proposed VNect and alignment for view invariant and expansion plus LSTM for action recognition. The work [6] proposed a novel end-to-end view adaptive framework that automatically alters the camera angle at each frame to obtain the consistent skeleton representation under the new view. The work [7] proposed learning view-invariant probabilistic embedding for 2D joint keypoints and applied nearest neighbor search for action recognition. The work [8] used motion trajectories of the skeleton to accomplish view invariance across different viewpoints. The work [9] proposed viewpoint-aware action recognition using viewpoint categorization network, skeleton-based features from multi-view image data and random forest. However, the above skeleton-based view-invariant action recognition works focus on human action rather than emergency human fall detection. In contrast to the above-mentioned approaches except [9], our framework trains the model with 2D and 3D pose representations from the target camera angle and uses RT transform to align pose representations of the source camera angle to the target camera angle. In other words, it does not require training data from different angles to achieve view invariance.

IV. CONCLUSION

In this paper, an effective skeleton-based view-invariant framework for human fall detection is proposed and evaluated using the datasets from different camera view angles. Our evaluation results show that the proposed framework with the modified D3PBDNN for source camera view angles can achieve equivalent accuracy of that of a trained target camera model. Through the experiments, the developed technology can largely save many times of training models and overcome the camera view variation issues inside elevators. In the near future, we can integrate this technology with the elevator scheduling [24, 25].

V. ACKNOWLEDGEMENT

The research is co-sponsored by Hitachi Yungtay Elevator Co., Ltd., Pervasive Artificial Intelligence Research (PAIR) Labs, and Ministry of Science and Technology (MOST).

TABLE 3. FALL DETECTION RESULTS OF OUR PROPOSED FRAMEWORK

Training				**Testing**					
Models	***Angle***	***F1-Score (Train)***	***F1-Score (Valid)***	***Angle***	***RT Transform***	***Accuracy***	***Precision***	***Recall***	***F1-Score***
Fig. 1	A	0.98	0.98	A	×	0.8	0.58	0.82	0.68
				B	×	0.81	0.94	0.76	0.84
					✓	0.9	0.98	0.86	0.91
				C	×	0.97	0.98	0.97	0.98
					✓	0.98	1.0	0.97	0.99
				D	×	0.62	0.4	0.83	0.54
					✓	0.87	0.86	0.56	0.68
				E	×	0.81	0.35	0.55	0.43
					✓	0.83	0.37	0.38	0.37

TABLE 4. COMPARISON WITH THE SELECTED SKELETON-BASED VIEW-INVARIANT FRAMEWORKS

References	**Baptista [5]**	**Zhang [6]**	**Sun [7]**	**Rawat [8]**	**Kim [9]**	**This Work**
Target Application	***Human Action Recognition***	***Human Action Recognition***	***Human Action Recognition***	***Human Action Recognition***	***Human Action Recognition***	***Human Fall Detection in an Elevator***
View Invariant Approach	VNect + Alignment	View Adaptive (VA) Neural Network via VA-RNN, VA-CNN, VA-Fusion	Probabilistic Embedding	Motion Trajectory Matching	Viewpoint Categorization Network	2D Pose Estimation (AlphaPose) + 3D Pose Estimation (3D Pose Baseline) + RT Transform (PA)
Action Recognition Approach	Expansion + LSTM	Classification Network	Nearest Neighbor Search		2D to 3D Skeleton Lifting + Random Forest	D3PBDNN + Post Processing
Action Recognition Input Data Type	3D Skeleton	3D Skeleton	2D Skeleton	3D Skeleton	2D and 3D Skeleton	2D and 3D Skeleton

REFERENCES

[1] R. Y. Tsai, "A versatile camera calibration technique for high-accuracy 3D machine vision metrology using off-the-shelf TV cameras and lenses," *IEEE J. Robot. Autom.*, vol. 3, no. 4, pp. 323–344, Aug. 1987.

[2] O. D. Faugueras and G. Toscani, "The calibration problem for stereoscopic vision," *Sensor Devices and Systems for Robotics*, Springer Berlin Heidelberg, pp. 195–213, 1989.

[3] J. Guan et al., "Extrinsic calibration of camera networks based on pedestrians," *Sensors*, vol. 16, issue 5, May 2016.

[4] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 11, pp. 1330–1334, Nov. 2000.

[5] R. Baptista, E. Ghorbel, K. Papadopoulos, G. G. Demisse, D. Aouada, and B. Ottersten, "View-invariant action recognition from RGB data via 3D pose estimation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2019, pp. 2542–2546.

[6] P. Zhang *et al.*, "View adaptive neural networks for high performance skeleton-based human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 1963–1978, Aug. 2019.

[7] J. J. Sun, J. Zhao, L. C. Chen, F. Schroff, H. Adam, and T. Liu, "View-invariant probabilistic embedding for human pose," in *Proc. European Conference on Computer Vision*, Aug. 2020, pp. 53-70, Springer, Cham.

[8] Y. S. Rawat and S. Vyas, "View-invariant action recognition," *Comput. Vis.*, pp. 1–10, Sep. 2020.

[9] S. H. Kim and D. Cho, "Viewpoint-aware action recognition using skeleton-based features from still images," *Electron.*, vol. 10, no. 9, pp. 1–12, 2021.

[10] G. Wei, C. Lan, W. Zeng and Z. Chen, "View invariant 3D human pose estimation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 12, pp. 4601-4610, Dec. 2020.

[11] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 7, pp. 1325–1339, Jul. 2014.

[12] H. S. Fang, S. Xie, Y. W. Tai, and C. Lu, "RMPE: Regional multi-person pose estimation," in *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 2017-October, pp. 2353–2362, Nov. 2016.

[13] J. Martinez, R. Hossain, J. Romero, and J. J. Little, "A simple yet effective baseline for 3d human pose estimation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2640–2649.

[14] Y. Li, K. Li, S. Jiang, Z. Zhang, C. Huang, and R. Y. D. Xu, "Geometry-driven self-supervised method for 3D human pose

estimation," in *Proc. AAAI Conference on Artificial Intelligence*, Apr. 2020, vol. 34, no. 7, pp. 11442-11449.

[15] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," arXiv preprint arXiv:1804.02767, 2018.

[16] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Jul. 2019, pp. 2623–2631.

[17] J. C. Gower, "Generalized Procrustes analysis," *Psychometrika*, vol. 40, issue 1, pp. 33-51, Mar. 1975.

[18] C. Goodall, "Procrustes methods in the statistical analysis of shape," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 53, issue 2, pp. 285-321, 1991.

[19] C. Wang and S. Mahadevan, "Manifold alignment using procrustes analysis," in *Proc. 25th International Conference on Machine Learning*, 2008, pp. 1120–1127.

[20] H. Temiz, B. Gökberk, and L. Akarun, "Multi-view reconstruction of 3D human pose with procrustes analysis," in *Proc. 2019 Ninth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, Nov. 2019, pp. 1-5.

[21] K. S. Arun, T. S. Huang, and S. D. Blostein, "Least-squares fitting of two 3-D point sets," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 9, no. 5, pp. 698–700, Sep. 1987.

[22] E. Auvinet, C. Rougier, J. Meunier, A. St-Arnaud, and J. Rousseau, "Multiple cameras fall data set," Tech. Rep 1350, University of Montreal, Jul. 2010.

[23] I. Charfi, J. Miteran, J. Dubois, M. Atri, and R. Tourki, "Optimized spatio-temporal descriptors for real-time fall detection: comparison of support vector machine and Adaboost-based classification," *J. Electron. Imaging*, vol. 22, no. 4, p. 041106, Jul. 2013.

[24] L. D. Van, Y. B. Lin, T. H. Wu, and T. H. Chao, "Green elevator scheduling based on IoT communications," *IEEE Access*, vol. 8, pp. 38404-38415, Mar. 2020.

[25] L. D. Van, Y. B. Lin, T. H. Wu, and Y. C. Lin, "An intelligent elevator development and management system," *IEEE Systems Journal*, vol. 14, no. 2, pp. 3015-3026, Jun. 2020.